Étude Bibliographique de Master 2

Spécialité : **AIGLE**

Personnalisation de page web : application à l'amélioration de l'accessibilité du web

par **Franck PETITDEMANGE**

Mars 2014

Sous la direction de **Marianne HUCHARD, Michel MEYNARD, Yoann BONAVERO**

Contents

1 Introduction et motivation

Le world wild web (www) est un reseau de ressource. La publication de ces ressources repose sur un langage universellement compréhensible et accepté par tous les ordinateurs : HTML. Historiquement conçut pour faciliter l'échange d'article dans la communauté dans la scientifique. La démocratisation du web à fait radicalement évoluer le contenu d'une page web, sans pour autant que le langage ne suivent ces évolutions. Ainsi les auteurs de page web ont détourné les pratiques de conception d'une page de manière anarchique. Ce manque d'homogénétité complique la compréhension du contenu publié sur le www par une machine. Faisant perdre la propriété universelle du web voulu par son créateur Tim Berners Lee :

> "La puissance du Web réside dans son universalité. L'accès à tous, quel que soit son handicap est un aspect essentiel"

Ceci introduit la motivation de ce stage et les problématiques qui en découlent.

Le sujet du stage est la personnalisation d'une page web. L'objectif est de fournir des méthodes et des outils afin d'adapter une page suivant les souhaits d'un lecteur. On s'interesse à une application pour l'amélioration de l'accessibilité des pages web pour les personnes en situation de handicap visuel.

L'adaptation d'une page web implique notre première problématique : la restructuration d'une page web. On souhaite expérimenter une approche basée sur les méta-modèles. L'idée étant d'extraire la structure d'une page et d'en construire une représentation plus abstraite. Cela doit nous permettre de s'affranchir de la diversité de conception de ces dernières. A partir de cette représentation on veut lui appliquer des transformations, puis générer une nouvelle page conforme aux transformations.

La conception des pages web s'articule autour : d'un langage pour décrire la structure du document (HTML) et d'un langage pour décrire la mise en forme du document (CSS). Les pages sont constituées d'éléments hétérogènes : une page est constituée d'un contenu principal, d'un menu de navigation, de publicité, etc... Chacun de ces éléments représentent une sous-structure de la page. Lorsqu'on regarde une page web depuis un navigateur, on constate que ces éléments sont structurés de façon sémantique, ils sont organisés selon leur sens. La difficulté dans la tâche de d'extraction de la structure d'une page est dû au manque d'expressivité de HTML. En effet, la norme actuelle de HTML (HTML 4), ne fournit pas de moyen de délimiter les éléments du document en fonction de leur sémantique. Par exemple, on ne peut pas délimiter de manière explicite la structure d'un menu dans une page avec ce langage. Le constat est que l'information de la structure d'une page apparaît principalement dans la mise en page. La structure d'une page est explicitée à travers l'utilisation de police, de couleur ou plus généralement d'élément visuel pour caractériser les contenus qui ont la même signification.

La seconde problématique est la définition d'un protocole d'acquisition et d'apprentissage automatique des souhaits de transformation d'une page. (...)

2 Modele de page web

2.1 Introduction

Modèle

Definition. Un modèle est une représentation simplifié d'une partie d'un système. C'est une abstraction du système étudié suivant un point de vue. Par exemple une carte routière est une abstraction d'un réseau routier, il existe plusieurs type de carte suivant ce que l'on veut étudier (chemin pédestre, chemin routier etc). L'intérêt d'un modèle est de mieux comprendre un système.

> "Pour un observateur A, M est un modèle de l'objet O, si M aide A à répondre aux questions qu'il pose sur O" (Minsky)

Métamodèle

Definition. Pour exprimer un modèle, nous avons besoin de pouvoir exprimer ces concepts. Un métamodèle, c'est un modèle qui fournit un langage pour exprimer un modèle. Littéralement, c'est un modèle de modèle.

Dans le cadre de notre sujet, on souhaite à terme l'adaptation d'une page web. Ici le système étudié est une page web. On s'interesse à la perception qu'un utilisateur peut avoir d'une page à travers un navigateur web. Il est important d'en cerner les concepts intrinsèques. L'objectif étant de concevoir un métamodele suffisament riche à l'expression d'un modèle conforme à la représentation qu'un utilisateur à d'une page, afin d'en faciliter l'expression de transformation.

Une page web possède le rôle d'affichage d'un contenu structuré et mise en forme. Le contenu peut être de type texte, image, vidéo, etc. On peut voir une page comme une composition d'élément graphique agencé dans l'espace de la page et apportant une information. Par exemple, une page est composé d'un menu de navigation, d'une entête, d'un pied de page, de formulaire et de widget (on entend par widget des éléments de contrôle, comme des barres de progression, des selecteurs de couleurs, etc). Chacun de ces éléments possède un type, un état et des propriétés.

Le type correspond à la nature de celui-ci. Il représente ce qu'un utilisateur attend de ce dernier. Il est attendu d'un menu de navigation qu'il nous permette de naviguer dans les page du site web.

Certain objet fournissent une interaction avec l'utilisateur qui est susceptible de modifier l'état d'un objet. Par exemple certain menu cache les différents liens de navigations, au moment où l'utilisateur survole le menu celui-ci affiche les différents liens.

Les objets possèdent également un ensemble de propriétés, quelque fois intrinsèque à leur type. Par exemple un formulaire possède des champs de saisie. D'autre propriétés peuvent être la couleur et la police du texte, la hauteur, la largeur, et la position d'une objets dans le page.

On s'interesse à un modèle capable de modéliser une page dans le but d'en abstraire la structure, les differents elements la composant, les relations entre

ces éléments, leurs natures et leurs comportements. On propose d'étudier et de comparer le langage standart du publication de document sur le web, HTML, dans la norme HTML 4 et 5 mais aussi une toxonomie pour la description d'interface graphique ARIA.

2.2 HTML 4

HTML 4 [6] est un langage permettant de structurer document sur le web. C'est le langage standard actuelle des pages web. Les données décritent peuvent être de type texte, ou plus généralement de type multimédia. Il permet de structurer le contenu et d'inclure des élements de mise en page. HTML 4 utilise la notion de section et sous-section d'un document pour décrire sa structure.

Structuration générique HTML 4 propose un mécanisme générique pour la composition du contenu formant la structure des pages web. Ce mécanisme gravite autour des éléments de type en bloc, en ligne et leurs identifiants respective : id et classe.

id et class Chaque élément peut se voir attribuer un identifiant ou une classe d'appartenance.

id assigne un nom à un élément. Ce nom doit être unique dans le document. Il possède plusieurs rôle dans HTML :

- sélecteur dans une feuille de style
- ancre cible dans lien hypertexte
- etc...

class, au contraire, assigne un ou plusieurs noms de classe à un élément ; on peut dire de l'élément qu'il appartient à ces classes. Un nom de classe peut être partagé par plusieurs instances d'éléments. L'attribut class a plusieurs rôles dans HTML :

- comme sélecteur dans une feuille de style (quand l'auteur souhaite assigner une information de style à un ensemble d'éléments) .
- pour un traitement universel par les agents utilisateurs.

En bloc et en ligne Certains éléments, qui peuvent apparaître dans l'élément BODY, sont dits être de niveau « bloc » tandis que d'autres sont dits de niveau « en-ligne ». La distinction se fondent sur plusieurs notions :

- Le modèle de contenu : les éléments de bloc peuvent contenir des éléments en-ligne et d'autres éléments de bloc. Les éléments en-ligne ne peuvent contenir que des données et d'autres éléments en-ligne. L'idée inhérente à cette distinction structurelle, c'est que les éléments de bloc créent des structures « plus grandes » que les éléments en-ligne.

- Le formatage : les éléments de bloc sont formatés différemment des éléments en-ligne. En général, les éléments de bloc commencent sur une nouvelle ligne, et non les éléments en-ligne.

Regroupement des éléments Les éléments et utilisé conjointement avec les attributs id et classe sont au cœur du mécanisme générique de structuration d'un document. Ces éléments définissent le contenu comme étant en-ligne (SPAN) ou de bloc (DIV) mais n'imposent aucune autre expression de présentation sur le contenu. La sémantique de ces balises est neutre, elle ne fournit aucune informations, il n' y a aucune information sur l'usage que son auteur en fait.

Figure 1: Architecture page web HTML 4

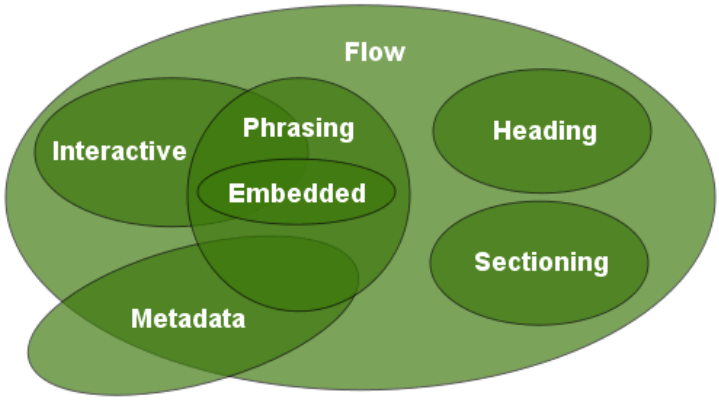
2.3 HTML 5

HTML 5 [7] étend HTML 4 en amenant de nouveaux éléments. La norme HTML 4 est toujours rétroactive, cependant quelques éléments ont été dépréciés, particulièrement les balises de mise en forme.

Cette nouvelle spécification de HTML n'est pas encore aux rangs de standard. HTML 5 tend à combler le fosser dans la description du contenu des pages web. HTML 5 remplace la structure générique de HTML 4 par un nouveau modele. Ce modèle apporte de nouveaux élements, catégorise les elements et apporte une sémantique à la composition des éléments.

Structuration HTML 5 propose toujours le même mécanisme générique de HTML 4 (éléments de type en bloc, en ligne et leurs identifiants respective : id et classe). HTML 5 propose une catégorisation des éléments spécifiant chacun un modèle de contenu, c'est à dire les règles définissant le type de contenu qu'il peut avoir.

Figure 2: Catégories des éléments HTML 5



- Metadata content (contenu de meta-données) : catégorie des elements qui modifient le comportement ou la présentation du document, insèrent des liens vers d'autres documents ou comportent des informations sur la structure même des données.
- Flow content (contenu de flux) : catégorie principale, regroupe la quasi-totalité des éléments disponibles en HTML5
- Sectioning content (contenu sectionnant) : catégorie des éléments qui créent une section dans le plan d'un document, définit la portée contextuelle des éléments
- Heading content (contenu de titre) : définit le titre d'un contenu
- Phrasing content (contenu phrasé) : catégorie qui regroupe les elements définissant du texte

- Embedded content (contenu integré) : contient tous les éléments qui appel des ressources externe au document
- Interactive content (contenu interactif) : regroupe les éléments conçuent pour une interaction avec l'utilisateur

Regroupement des éléments HTML 5 apporte de nouveaux élements permettant de mieux définir la structure sémantique d'un document. Une sémantique accompagne la composition des elements. Par exemple, quand un élément <article> est imbriqué dans un autre élement <article>, cette dernière représente un article relatif à l'élément contenant, comme un commentaire. Un élement <adress> dans un élement <article> est compris comme l'adresse de l'auteur de l'article.

- Section : permet de définir les grandes sections d'un document
- Article : représente un contenu autonome dans une page
- nav : représente une section de liens vers d'autre page ou des fragments de cette page
- aside : représente une section de la page dont le contenu est indirectement lié à ce qui l'entoure et qui pourrait être séparé de cet environnement
- header : représente un groupe d'introduction ou une aide à la navigation. Il peut contenir des éléments de titre, mais aussi d'autres éléments tels qu'un logo, un formulaire de recherche, etc.
- footer : représente le pied de page, ou de la section, ou de la racine de sectionnement la plus proche

Figure 3: Architecture HTML 5

2.4 ARIA

ARIA (Acessible Rich Internet Application) [5] est la spécification d'une ontologie décrivant une interface graphique. Elle fournit des informations sur la structuration d'un document et décrit les élements qui composent l'interface au moyen d'un ensemble de rôles, d'états et de propriétés .

Rôle Les rôles permettent d'identifier la fonction de chaque elements d'une interface. Ils sont regroupé en trois catégories :

- Widget Roles : référence d'un ensemble de widget (alertdialog, button, slider, scrollbar, menu, etc)
- Document Structure Roles : décrit les structures qui organisent un document (article, definition, heading, ect)
- Landmark Roles : décrit les régions principales d'un document (main, navigation, search, etc)

Etats et propriétés ARIA permet d'associer des états et propriétés à des wigets.

Un état est une configuration unique d'un objet. Par exemple, on peut définir l'état d'un bouton par l'état *aria-checked* qui peut prendre trois propriétés suivant l'interaction avec l'utilisateur : *true - false - mixed*

On peut associé un ensemble de propriété par exemple la valeur minimal ou maximal que l'on doit remplir dans un champs de saisit *aria-valuemin, aria-valuemax.*

(http://www.w3.org/TR/wai-aria/states_and_properties#state_prop_att -> à compléter)

2.5 Discussion

HTML 4 *HTML 4 fournit quelque idiome pour structurer les éléments d'un document. Il fournit un langage riche pour décrire les éléments de type texte et les hyperliens, mais un langage peu adapter pour décrire la structure d'une page web, dans le contexte actuel de l'exploitation de ses dernières. En effet le contenu décrit dans les pages web actuelles est très hétérogène, HTML 4 net permet pas de delimiter de manière explicite la structure sémantique d'une page. HTML 4 n'est pas un support assez riche pour être le support d'un méta-modèle exprimant le modèle d'une page web suivant nos besoins.*

Id et class permette la strucuturation sémantique, à défault des balises de regroupement et

HTML 4 repose sur le notion de section et sous section. Le mécanisme reposant sur les division ne permet pas de structurer le document de façon hierarchique mais ne donne pas un sens aux elements. Pour le conception d'un modéle on peut connaitre la struture d'un document mais on ne connaît pas les rôles associés aux documents.

HTML 5 *Ces balises balises répondent à un besoin de structuration du document possedant des données hétérogènes. Contrairement à HTML 4, la sémantique de la structure et les relations des éléments est définit de manière explicite. Repose moins sur l'utilisation des id et class. On connait mieux le rôle des élément entre eux.*

ARIA

3 Extraction structure

3.1 Introduction

L'extraction est la première phase du processus de restructuration. On peut la décrire comme le processus de découverte des éléments qui composent un tout. Ici on veut récuperer la structure d'une page web. On veut extraire la structure d'une page web suivant le point de vu de notre méta-modele, c'est à dire découvrir les élements d'une page conforme aux élements de notre méta-modele. Le méta-modèle refletant le point de vu du lecteur de contenu, on peut également dire que l'on veut découvrir les éléments conforme au point de vu de l'utilisateur.

La difficulté de ce processus, comme soulevé dans l'introduction, provient du manque de sémantique des élements du langage de la norme actuelle de HTML (HTML 4), ainsi que le detournement des éléments syntaxiques dans la conception des pages par leur auteur.

HTML 4 utilise la notion de section et sous-section d'un document pour décrire sa structure. Une section est définit par un element de division <DIV> avec des elements d'entête (<h1>, <h2>, <h3>, ...). La relation de ces elements de division et d' entete conduisent à la structure du document.

- <DIV> définit une section mais à une valeur sémantique neutre. La valeur sémantique n'est connu que par son auteur au travers l'attribution de classe d'appartemence. Ces classes permettent d'associer une mise en forme aux élements. Il n'y a pas de contrainte sur l'imbrication des balises. On ne peut pas savoir si elles representent une section, une sous-section.

- Chaque section correspond à une partie d'un document. Ces parties ne décrivent pas systematiquement un contenu linéaire. Par exemple on peut trouver des blocs correspondant à des publicités, sans rapport avec la section parent.

- Chaque section représentent une partie du document, cependant certaine section contiennent pas d' informations en rapport avec le contenu du document mais sur le site web. C'est le cas des menus, des logos, etc...

En effet, le mécanisme de structuracturation des pages est trop générique, ce qui le rend ambiguie la sémantique des elements. Ce mécanisme repose en parti sur l'élement <DIV> qui permet de diviser le document, la sémantique de cet element est neutre, on ne connait pas la signification de cette division. De plus il n y'a pas de restiction dans les imbrications des élements <DIV> ce qui rend difficile de connaître les limites sémantiques.

L' ambiguité syntaxique Le processus d'extraction est rendu compliqué par les mécanisme trop générique et le détournement qui est fait de la syntaxe des élements du langage.

Le mécanisme de structuration générique proposé par HTML 4 est bien adapté la description d'une structure hierarhcique dans document mais peu à

un contenu heterogène. On ne pas identifier le rôle des différents élements dans la structure.

La mauvaise utilisation des balises. Par exemple, la balise (dans le spécification de HTML 4) est utilisé pour divisé la structure du document de manière hierarchique, or en pratique elle est également utilisé pour appliquer un style de mise en forme sur un ensemble d'élement. La balise <hr> est un autre exemple de dérive.

Il est en résulte que la structure sous jacente peut ne pas être représentatif de l'intention de l'auteur du document. En d'autre terme la représentation visuelle peut ne pas être coherente avec la structure sous-jacente. Quelques approches (citer quelques articles) Nous traitons ici d'une approche par segmention visuel, une segmentation basé sur l'interpretation qu'un humain à d'une page au travers d'un navigateur internet. Puis nous parlemenrons d'une approche basé sur un modèle probabilistique.

3.2 Approche segmentation visuelle

L'approche proposé par les auteurs [1] présente un algorithme de partitionnement basé sur les élements de mise en forme des documents web. L'algorithme pourcourt le DOM de la page web et partitionne l'arbre suivant des heuristiques prenants en compte le type de noeud rencontré, les noeuds frère et parent.

3.3 Approche stochastique

L'approche proposé par les auteurs [3] jjjj

3.4 Discution

4 Detection d'objet

4.1 Introduction

On s'interesse ici au sens des sous-structures recupérés dans la phase d'extraction. Plus précisement on veut savoir à quel element de notre méta-modèle correspond la sous-structure. On veut faire correspondre les sous-structures aux elements de notre méta-modèle.

L'environnement web impose certaine contraintes sur notamment l'évolution des structures de données, mais également la diversité des arcthitecutes d'un site à l'autre. Cependant on peut constater qu'avec l'utilisation massive de

La section précendente décrit un moyen d'extraire la structure d'une page. On obtient une structure semantique de la page. On s'interesse à reconnaitre les éléments de cette structure suivant des elements de notre modèles.

La contrainte imposée par le web est la diversité des structures ayant la même fonction.

4.2 Wrapper

4.3 Tree-to-tree distance

Une approche est la comparaison d'arbre. Cette approche consiste à trouver la plus petite ou la moins coûteuse séquence d'opération d'édition (substitution, suppression et insertion) qui permet la transformation d'un arbre vers un autre.

Notons Λ un noeud vide. Une opération d'édition est ecrit $b \to c$, ou b et c sont soit un noeud, soit Λ .

- $b \to c$ est une opération de substition si $b \neq \Lambda$ et $c \neq \Lambda$,
- une opération de suppression si $b \neq \Lambda \doteq c$,
- et une opération d'insertion si $b = \Lambda \neq c$

Pour exprimer la séquence d'opération élémentaire qui transforme l'arbre, on utilise le concept de mapping, introduit [4]. Un mapping établie une correspondance un-à-un entre les nœuds de deux arbres ordonnés et qui préservent l'ordre des nœuds.

Definition. Un Mapping M de l'arbre T1 vers l'arbre T2 est un ensemble de paire ordonnée d'entier (i, j), $1 \leqslant i \leqslant n1$, $1 \leq j \leq n2$, satisfaisant les conditions suivantes, pour tous $(i1, j1), (i2, j2) \in M$:

- i1=i2 si et seulement si, j1=j2 (one-to-one condition);
- t1[i1] est à droite de t1[i2], si est seulement si, t2[j1] est à gauche de t2[j2] (preservation de l'ordre des noeuds frères);
- t1[i1] est un ancêtre de t1[i2] si et seulement si, t2[j1] est un ancêtre de t2[j2] (preservation de l'ordre des ancêtres);

Definition. Soit M un mapping entre les arbres T1 et T2 décrivant des opérations de modification. S est l'ensemble de pair $(i, j) \in M$, D l'ensemble des nœuds T1[i] n'ayant pas de paire $(i, j) \in M$, et I l'ensemble des nœuds T2[j] n'ayant pas de paire $(i, j) \in M$. Le coût du mapping est donné par $|S|p + |I|q + |D|r|$, ou p est le coût des substitution non identique, q est le coût des insertions (1), r est le coût d'une suppression (1), le coût des substitution identique est 0.

Quatre mesures de distances Plusieurs mesures de distance existent, chacune d'elle mettant en avant certaine propriétés fonctionnelles d'une structure. Ces mesures de distance imposent des contraintes sur le mapping. Dans la littérature en recense principalement 4 mesures : Distance de modification, distance d'alignement, distance de sous-arbre isolé, distance descendante.

Distance de modification C'est le coût minimum du mapping d'un arbre T1 vers T2.

Distance d'alignement Coût minimum du mapping n'ayant que des opérations d'insertion de nœud pour que T1 et T2 soit isomorphe. Principalement utile pour connaître le coût de recouvrement entre deux arbres. Utile pour faire ressortir les régions homologue ou similaire d'une structure.

Distance de sous-arbre isolé La distance est obtenu en imposant sur le mapping la contrainte que deux sous-arbres disjoints de T1 doivent être mappés à deux sous-arbres disjoints de T2. S'avère très utile dans le comparaison d'arbre de classification[2]. Elle correspond au côut minimum du mapping de sous-arbres isolés de T1 vers T2.

Distance descendante *Utilisé pour trouver la plus grande sous-strucucture commune entre deux arbres (...)*

Algorithmes [4] propose un algorithme de programmation dynamique pour résoudre la question de distance d'arbre en temps séquentiel $O(|T1| \times |T2| \times min(depth(T1), leaves(T1)) \times min(depth(T2), leaves(T2)))$.

4.4 Arbre de décision

4.5 Discution

Tree-to-tree distance Dans le cadre de notre problématique, il est intéressant d'utiliser la distance de modification entre la structure d'un objet définit à priori et une structure extraite d'une page web. Par exemple, si nous souhaitons connaître le rôle d'un objet extrait d'une page, on peut le comparer à la collection d'objet définit à priori et déterminer de quel objet il est le plus proche. Cette distance à un rôle d'indice de similarité.

5 Conclusion

References

[1] Ji-Rong Wen Wei-Ying Ma Deng Cai, Shipeng Yu. Extracting content structure for web pages based on visual representation.

[2] K. Takana E. Takana. The tree-to-tree editing distance problem.

[3] Francis Maes Patrick Gallinari Guillaume Wisniewski, Ludovic Denoyer. Modele probabiliste pour l extraction de structures dans les documents semi-structures : Application aux documents.

[4] KUO-CHUNG TAI. The tree-to-tree correction problem.

[5] W3C. Accessible rich internet applications 1.0.

[6] W3C. Html 4.01 specification.

[7] W3C. Html 5 specification.