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 sujet du stage s'inscrit dans le contexte de personnalisation d'un page web. L'objectif visé est de fournir des méthodes et des outils afin d'adapter une page suivant les souhaits d'un utilisateur.

La première problématique est 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 du monde. 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 méta-modele 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. Le WC3 décrit une page web comme une ressource. Chaque ressource étant une unité d'information. Cependant dans la conception contemporaine des pages, celle-ci peut encore être découpé en unité d'information. Ainsi on peut voir une page comme une composition d'élément graphique apportant chacun 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. 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.

2.2 HTML 4

HTML 4 [4] est un langage permettant de structurer les données 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 définit des types d'éléments qui représente des structures ou des comportement voulu dans une page web. HTML 4 permet de décrire trois composant principaux :

1. Le contenu textuel : paragraphe, titre (beaucoup de balise de présentation)
2. Les liens hypertextes
3. Les images, objets

Structuration générique HTML propose un mécanisme générique pour la structuration des documents 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.

Le 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.

2.3 HTML 5

HTML 5 étend HTML 4 en amenant de nouveaux éléments. Ces éléments ajoutent une valeur sémantique aux structures de HTML 4. HTML 5 spécifie sont modèle de structure générique par une catégorisation des éléments d'une page. (...)

2.4 ARIA

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.*

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.*

ARIA

3 Extraction structure

3.1 Introduction

3.2 Approche segmentation visuelle

3.3 Approche stochastique

3.4 Discution

4 Detection d'objet

4.1 Introduction

Dans le cadre de notre sujet on explore différentes pistes pour la détection d'objet type définit à priori. La contrainte imposée par le web est la diversité des structures ayant la même fonction. On émet l'hypothèse que les structures proches ont la même fonction[1].

4.2 Tree pattern matching

Ces problèmes sont étudiés dans le domaine de la recherche d'information (RI). Ce domaine étudie la manière de retrouver des informations dans un corpus.

Une approche est la comparaison d'arbre. Cette approche dans la RI consiste à trouver tous les sous-arbres isomorphe avec le pattern en entrée. Pour faire la comparaison entre deux structures, on utilise le concept de distance entre deux arbres.

Definition. La distance entre deux arbres est calculée par 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 [3]. 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.

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

4.3 Arbre de décision

4.4 Discution

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] Kaizhong Zhang Bruc A.Shapiro. Comparing multiple rna secondary structures using tree comparisons.

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

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

[4] W3C. Html 4.01 specification.