Académie de Montpellier
Université Montpellier II
Sciences et Techniques du Languedoc

Étude Bibliographique de Master 2

effectuée au Laboratoire d'Informatique de Robotique
et de Micro-électronique de Montpellier

Spécialité : **AIGLE**

Personnalisation de page web

par **Franck PETITDEMANGE**

Mars 2014

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

1 Introduction

2 Modèle de page web

2.1 Introduction

Un modèle est une abstraction du système étudier 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)
Dans le cadre de notre sujet, on souhaite à terme l'adaptation d'une page web. Ici le système étudié est une page web.

Nous nous intéressons à un méta-modèle qui puisse décrire la structure, ainsi que le comportement et les propriétés des différents objets constitutifs d'une page web afin d'en exprimer un modèle. Le méta-modèle doit fournir un langage assez riche pour permettre d'identifier les objets que souhaite modifier un utilisateur dans notre page web.

Nous avons étudié 3 support à la réalisation d'un méta-modèles pouvant décrire les objets constitutif d'une page web : HTML 4, HTML 5 et la norme ARIA.

2.2 HTML 4

Généralité HTML 4 est un langage permettant de représenter les données sur le web. C'est le langage standard actuelle des pages web. Les données décrite peuvent être de type texte, ou plus généralement de type multimédia. Il permet de structurer le contenu et d'inclure une mise en page.
HTML 4 définit des types d' éléments qui représentent des structures ou des comportements voulut dans une page web. Il permet de décrire trois composant principaux :

- Le contenu textuel : paragraphe, titre (beaucoup de balise de présentation)
- Les liens hypertextes
- Les images, objets etc..

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 est unique dans le document. Il possède plusieurs rôles dans HTML :

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

L'attribut class, au contraire, assigne un ou plusieurs noms de classe à un élément. 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 » (aussi connu comme sous le nom « niveau texte »). 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 ¡div¿ et ¡span¿ 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

2.4 ARIA

2.5 Discussion

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 n'est pas un support assez riche pour être le support d'un méta-modèle exprimant le modèle d'une page web.

3 Tree pattern matching

Dans le cadre de notre sujet on explore différentes pistes pour la détection d'objet type définit à priori. Ses problèmes sont étudiés dans le domaine de recherche d'information (RI). Une approche de ce domaine est la comparaison d'arbre. Cette approche est apparenté aux problématiques de correspondance de motif dans les structures arborescente (tree pattern matching). De manière analogue aux problèmes de correspondance de motif dans une chaîne (string pattern matching) qui prend entrée un motif et un texte et qui produit en sortie la localisation d'une sous-chaîne de caractères du texte en entrée correspondant aux motifs recherchés. Dans le tree pattern matching, le motif et le texte sont des structures arborescentes. Le problème de recherche de motif (pattern matching) consiste à trouver tous les sous-arbres du texte isomorphe avec le pattern en entrée. Dans la littérature ont trouve plusieurs mesures de distance entre deux arbres utiles à la comparaison d'arbre.