

Inférence de la structure d'une page web en vue d'améliorer son accessibilité

Encadré par : Y. Bonavero, M. Huchard et M. Meynard

Franck PETITDEMANGE



26 juin 2014

Sommaire

- 1 Introduction
- 2 État de l'art
- 3 Réalisation
- 4 Conclusion

Sommaire

1 Introduction

2 État de l'art

3 Réalisation

4 Conclusion

Accessibilité du web

Un enjeu sociétal important

Definition

Accessibilité : capacité d'accéder aux informations contenues dans une page et d'interagir avec.

Problèmes d'accessibilité (spécifique aux basses visions)

- Surcharge visuelle
- Police de caractère
- Contraste de couleur

Accessibilité du web

Besoin de comprendre la structuration d'une page

Problèmes des outils d'accessibilité

- Pas de traitement des couleurs locales
- Pas de prise en compte des profils utilisateurs

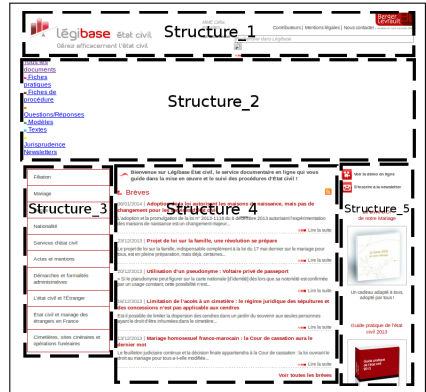
Besoin

- Comprendre les informations structurant une page web

Page web

Page web

- Technologies :
HTML/CSS/Javascript
- Contenu hétérogène décrit par
différentes structures logiques



Comment inférer les différentes structures logiques dans une page web ?

Difficultés

- Manque d'expressivité de HTML 4
- Pas de construction standard des structures logiques
- Écart entre la structure DOM et l'affichage dans un navigateur

Approche

- Étude des langages de publication de page web
- Étude des techniques d'extraction de structure d'une page

Sommaire

1 Introduction

2 État de l'art

- Étude des Langages de publication
- Étude de méthodes d'extraction de structure

3 Réalisation

4 Conclusion

Évolution de la sémantique (1/2)

```
<ul class='menu'>
  <li><a href=".">l1</li>
  <li><a href=".">l2</li>
</ul>
<p class='menu'>
  <a href=".">l1</a>
  <a href=".">l2</a>
</p>
```



HTML 4

- Peu de sémantique
- Diversité de représentation
- Structure logique implicite

Évolution de la sémantique (2/2)

HTML 5

- Structure logique explicitée
- Sémantique pour décrire l'interface de la page est limitée

ARIA

- Ontologie d'une interface graphique
- Trop élaborée pour nos besoins mais est plus expressif



Évolution de la sémantique (2/2)

HTML 5

- Structure logique explicitée
- Sémantique pour décrire l'interface de la page est limitée

ARIA

- Ontologie d'une interface graphique
- Trop élaborée pour nos besoins mais est plus expressif



CSS

Un langage de mise en forme

Propriétés de mise en forme :

- avant-plan/arrière-plan
- police
- ...

Mécanisme de positionnement

- relatif
- absolu
- flottant

Synthèse

HTML 4 langage actuellement le plus exploité. Les inconvénients sont :

- la diversité de représentation d'une même structure logique
- la faible expressivité au regard des concepts décrits dans les pages web
- l'absence de structuration explicite

Notre approche

- Proposer un Méta-modèle concrétisant mieux les concepts des pages et permettant de s'abstraire de la diversité de représentation des structures

Mapping

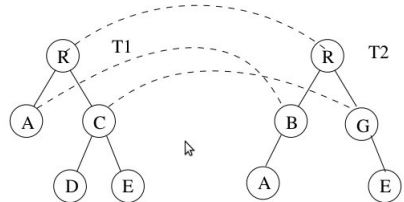
(Vieira et al., *A fast and robust method for web page template detection and removal*)

Mapping descendant restrictif

Permet de faire correspondre les plus grandes sous-structures communes entre deux arbres.

Idée

Identifier les structures logiques des pages web par correspondance



Segmentation

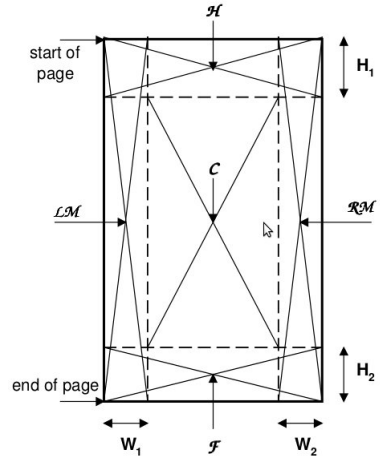
par pattern de présentation (*Milos Kovacevic et al., Recognition of Common Areas in a Web Page Using Visual Information : a possible application in a page classification*)

Observation

Les concepteurs de page web suivent approximativement les mêmes schémas de présentation

Idée

Regrouper les nœuds du DOM de la page suivant leurs coordonnées après la mise en page par le navigateur



Segmentation

par densitométrie textuelle (*Kohlschütter et al, A densitometric approach to web page segmentation*)

Étape 1 : identification de segments de petites tailles

La page est vue comme une séquence de caractères entrelacés identifiés par des balises HTML. Les segments sont calculés d'après les variations dans le rythme des séquences pour lesquels on calcule une densité textuelle.

Exemple :

Ici deux segments seront calculés

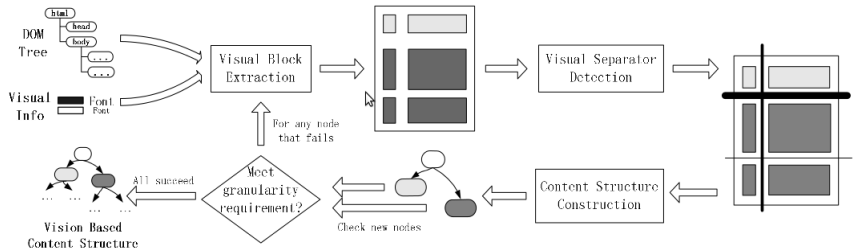
```
<a>lienA</a>lienB<a>lienC</a><p>un paragraphe</p>
```

Étape 2 : grossissement successif des segments par fusion

Les segments contigus dont la densité textuelle est proche sont fusionnées successivement.

Segmentation

par indice visuel (Cai et al., *Extracting content structure for web pages based on visual representation*)



Idée

Regroupement par indice visuel

Synthèse

Les inconvénients

- Le *Mapping* ne permet pas d'extraire la structure globale de la page
- La segmentation par pattern est trop dépendante de la présentation de la page
- La segmentation par densitométrie ne prend pas en compte les écarts possibles entre le DOM et le rendu final
- Le calcul des séparateurs dans l'approche par indice visuel est une opération coûteuse $O(n^2)$

Notre approche : approche par segmentation visuelle

Propose un **découpage global** de la page, **indépendant des patterns de présentation** et permet un **découpage fin** dans la structure d'une page.

Sommaire

1 Introduction

2 État de l'art

3 **Réalisation**

- Méta-modèle
- Extraction structure
- Annotation structure

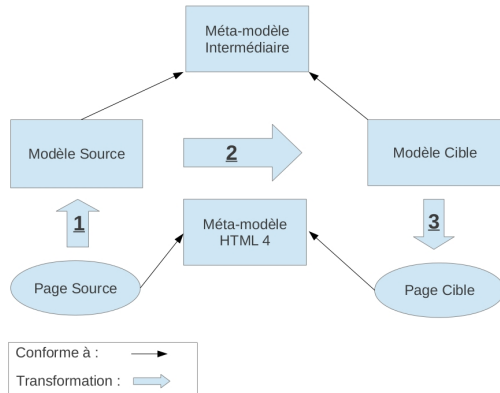
4 Conclusion

Approche générale

Une approche d'Ingénierie Dirigée par les Modèles

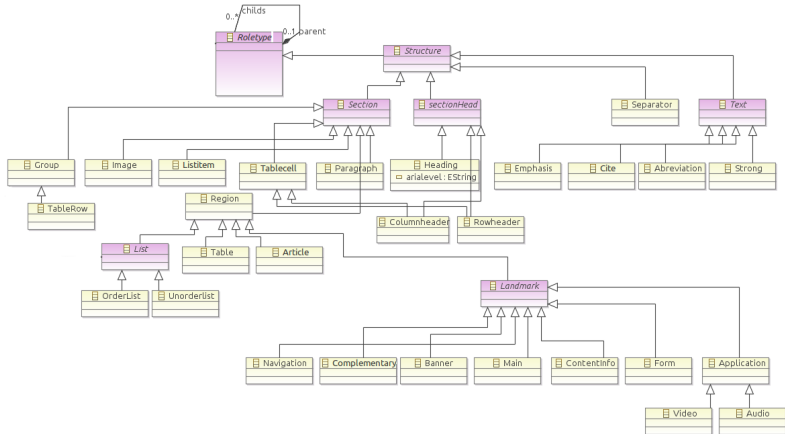
Avantages

- Meilleure expression des préférences
- Écriture d'adaptation indépendante des langages



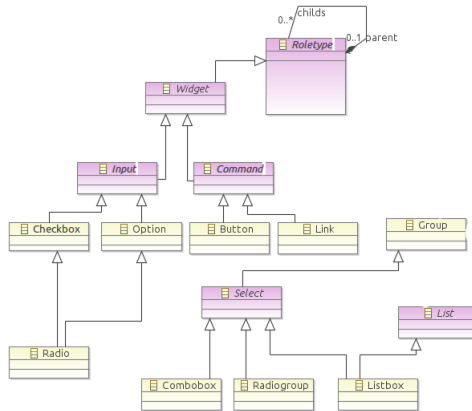
Méta-modèle intermédiaire

Éléments structurels



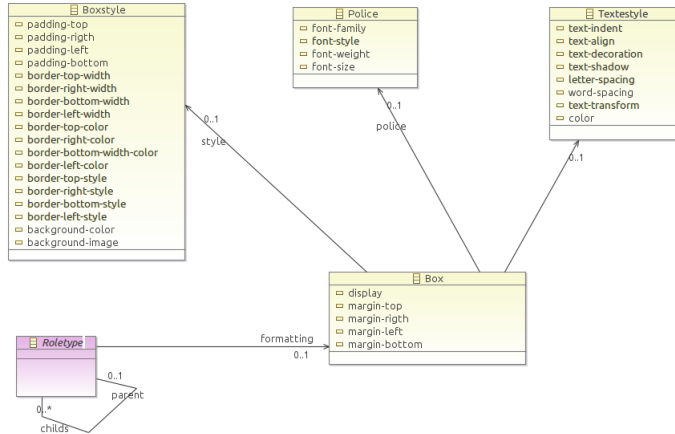
Méta-modèle intermédiaire

Éléments d'interaction



Méta-modèle intermédiaire

Éléments de mise en forme



Démarche générale

```
function CONSTSTRUCTLOG(noeudDom, noeudStructInter)  
    poolNoeuds  $\leftarrow$  0  
    DIVISIONDOM(noeudDom, poolNoeuds)  
    for i in poolNoeuds do  
        APPENDCHILD(arbreIntermediaire, poolNoeud[i])  
        CONSTSTRUCTLOG(poolNoeud[i], arbreIntermediaire)  
    end for  
end function
```

```
function DIVISIONDOM(noeudDom, poolNoeuds)  
    if DIVISIBLE(noeudDom) == TRUE then  
        for all Enfant de noeudDom do  
            DIVISIONDOM(noeudDom, poolNoeuds)  
        end for  
    else  
        poolNoeuds  $\leftarrow$  poolNoeuds + noeudDom  
    end if  
end function
```


Concepts et heuristiques

Concepts

Nœuds conteneurs : propriété CSS display égale à *bloc*

Nœuds de mise en forme : propriété CSS display à *inLine*

Taille d'un nœud : taille relatif à sa position dans le DOM et à la somme du poids des nœuds qu'il contient

Distance visuelle : signale les changements de propriété de mise en forme des nœuds

Règles

Règle 2 : si le nœud possède un seul enfant et que cet enfant n'est pas un nœud de données alors on parcourt ce nœud

Règle 4 : si l'un des enfants du nœud est un nœud conteneur et que sa taille est supérieure à un certain seuil alors on parcourt ce nœud

Règle 8 : si la fonction de distance visuelle est vérifiée pour l'un des nœuds enfants alors on parcourt ce nœud et on extrait les enfants qui vérifient la fonction

Processus d'extraction

Segmentation globale page Berger-Levrault



légibase état civil

Où lire effacement l'état civil

MINISTÈRE
DES AFFAIRES
ÉTRANGÈRES

Contributions | Mentions Régulées | News Contributor

Recherche dans le légibase

Tous les documents

- Fiches pratiques
- Fiches de procédure
- Questions/Réponses
- Modèles
- Textes
- Jurisprudence
- Newsletters

Mission
Service
Délicé
Nationalité
Services d'état civil
Actes et mentions
Démarches et formalités administratives
L'état civil et l'hérédité
État civil et mariage des étrangers en France
Consentements, actes civils et applications homologues

Demarche sur Légibase état civil, le service documentaire en ligne qui vous guide dans la mise en œuvre et le suivi des procédures d'état civil !

1. Brèves

09/09/2014 | Adoption de la loi autorisant les maires de naissance, mais pas de changement pour les services d'état civil

La loi n° 2013-1121 du 6 décembre 2013 autorisant l'implémentation des maires de naissance est entrée en vigueur le 1er janvier 2014. Elle ne concerne pas les services d'état civil.

23/10/2013 | Projet de loi sur la famille, une révolution se prépare

Le projet de loi sur la famille, l'indisposabilité est présenté la loi du 27 mai dernier sur le mariage pour tous, nous envoie plusieurs propositions, mais déjà, certaines.

29/10/2013 | Uniformité d'un pseudonyme : véhicule prêt de passeport

Si le pseudonyme peut figurer sur la carte nationale d'identité, il est très clair que la nationalité est confirmée par un usage constant, une possibilité n'est...

26/10/2013 | Limitation de l'acte à un conjoint : le régime juridique des sépultures et des concessions n'est pas applicable aux couples

Est-il possible de limiter le déplacement des cendres dans un jardin du coconner aux autres personnes après le décès d'une situation dans la situation...

18/10/2013 | Mariage homosexuel franco-marocain : la Cour de cassation aura le dernier mot

Le mariage polygamie continue et la décision l'acte appartient à la Cour de cassation. La loi ouvrant droit au mariage pour tous a été modifiée...

Sans inscription de l'Agence

 L'inscription de l'agence

Le Livre d'Or de notre Mariage




Un cadeau unique à tous, adapté par vous !

Cadre pratique de l'état civil 2013



Voir toutes les brèves



légibase état civil

Ouvrez efficacement l'état civil

Structure 1

Menu

Contribuer

Mentions légales

Nous contacter

Accueil

Documents

Fiches

Informations

Statistiques

Procédures

Questions/Réponses

Moyens

Textes

Jurisprudence

Prescriptions

Structure 2

Menu de données en ligne

Statistiques de la nomenclature

Filiation
Mariage
Nationalité
Services d'état civil
Actes et mentions
Démarches et formalités administratives
L'état civil et l'hérédité
État civil et mariage des étrangers en France
Célibats, état civils et registres familiaux

Structure 3

Présentation sur Légibase l'état civil, le service documentaire en ligne qui vous guide dans la mise en œuvre et la suite des procédures d'État civil

Brèves

CONFÉRENCES | Adoption d'un enfant, les procédures de nationalité, mais pas de mariage pour tous

CONFÉRENCES | **Projet de loi sur la famille, une révolution en préparation**

Le projet de loi sur la famille, indissolublement accompagné à la loi du 10 mai dernier sur le mariage pour tous, est en pleine préparation, mais déjà, certaines...

CONFÉRENCES | **Utilisation d'un passeportier : valuaire privé de passeport**

Si le passeportier peut signer sur la carte nationale d'identité des lors que sa compétence est confirmée par un usage constant, cette possibilité n'est...

CONFÉRENCES | **Limitation de l'accès à un civetiste : la régime juridique des signatures et des connotations n'est pas applicable aux centres**

C'est possible de limiter la disparation des centres dans un cadre de pouvoir aux seules personnes appartenant d'être volutaires dans la civetiste...

CONFÉRENCES | **Mariage homosexuel Franco-marocain : la Cour de cassation aura le dernier mot**

Le fractionnaire judiciaire concerné et la décision finale appartiendra à la Cour de cassation. Si le courant se fait au mariage pour tous à telle mesure...

[Voir toutes les brèves](#)

▼

▼

▼

▼

▼

▼

▼

▼

▼

▼

Structure 4

CONFÉRENCES | **Projet de loi sur la famille, une révolution en préparation**

Le projet de loi sur la famille, indissolublement accompagné à la loi du 10 mai dernier sur le mariage pour tous, est en pleine préparation, mais déjà, certaines...

CONFÉRENCES | **Utilisation d'un passeportier : valuaire privé de passeport**

Si le passeportier peut signer sur la carte nationale d'identité des lors que sa compétence est confirmée par un usage constant, cette possibilité n'est...

CONFÉRENCES | **Limitation de l'accès à un civetiste : la régime juridique des signatures et des connotations n'est pas applicable aux centres**


C'est possible de limiter la disparation des centres dans un cadre de pouvoir aux seules personnes appartenant d'être volutaires dans la civetiste...

CONFÉRENCES | **Mariage homosexuel Franco-marocain : la Cour de cassation aura le dernier mot**

Le fractionnaire judiciaire concerné et la décision finale appartiendra à la Cour de cassation. Si le courant se fait au mariage pour tous à telle mesure...

Structure 5

de notre Mariage



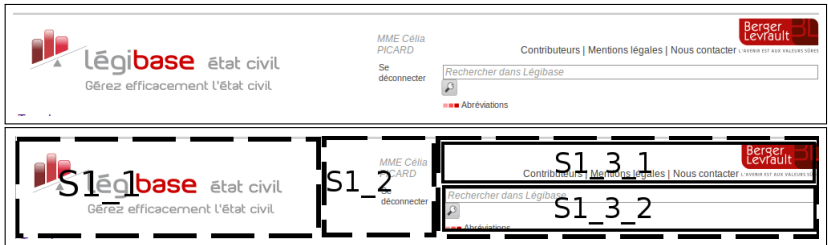
Un cadastre, adapté à tous, adapté par tout !

Guide pratique de l'état civil 2013

avec un guide de l'état civil

Processus d'extraction

Segmentation locale page Berger-Levrault



Approche par fonctionnalité

Construction de fonction d'annotation (*Chen et al., Function-based object model towards website adaptation*)

Classe de nœuds :

- Informatif
- Navigation
- Interaction
- Décoration

Chaque classe possède les propriétés :

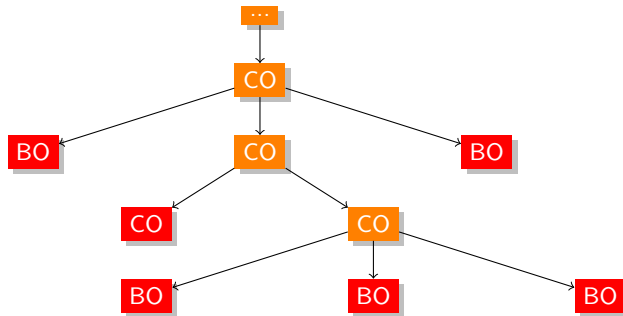
- Hyperlien
- Sémantème
- Décoration
- Alignement
- **Position**

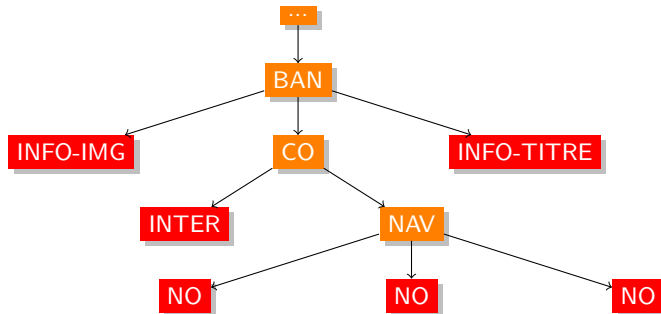
Approche par fonctionnalité

fonctions de détection proposées

Banner :

ContentInfo :





Sommaire

1 Introduction

2 État de l'art

3 Réalisation

4 Conclusion

Difficultés

Difficultés

- Domaine de recherche éloigné des connaissances de l'équipe de recherche
- Définir la problématique par rapport à la question de recherche (recherche de motifs d'intérêt dans un arbre DOM)
- Recherche d'articles traitant de la problématique

Résultats

Résultats

- Proposition d'une approche IDM
- État de l'art sur les langages de publication de page web
- État de l'art des techniques d'extraction de structure
- Proposition d'un méta-modèle
- Adaptation et implémentation d'une méthode pour extraire les structures d'une page
- Proposition de pistes pour annoter les structures extraites

Perspectives

Perspectives à court terme

- Évaluation de la méthode d'extraction de structure
- Implémentation, évaluation et élargissement du processus d'annotation

Perspectives à long terme

- Évaluation des modèles intermédiaires générés par le processus d'extraction et d'annotation sur un grand nombre de pages
- Acquisition des préférences utilisateurs
- Intégrer les préférences utilisateurs dans le processus de transformation