

## 1 Détection d'objets HTML

**Mapping de structures** [6] propose une méthode de mapping entre les pages web pour en isoler des sous-structures communes que l'on pourra identifier comme des objets. L'inconvénient est que l'on ne peut découvrir qu'une partie des objets dans les pages web avec cette approche.

**Segmentation par Pattern** [4] propose un partitionnement d'une page web après génération de celle-ci par un moteur graphique sur la base d'un modèle de présentation défini a priori. Le modèle de présentation correspond à un pattern de mise en forme "standard" des pages web : header, footer, menu lat. gauche, menu lat. droit.

**Segmentation par densitométrie** [3] propose un partitionnement d'une page d'après la variation de densité textuelle de chaque segment textuelle.

**Segmentation par indice visuelle** [1] propose un partitionnement d'une page en fonction des propriétés de mise en forme associées à chaque noeud du DOM. L'approche propose un découpage de la page depuis la racine jusqu'aux feuilles. Le découpage du DOM est réalisé sur la base d'heuristiques et d'un algorithme de fusion des noeuds du DOM.

## 2 Annotation d'objets HTML

**annotation basée sur les fonctionnalités** [2] proposent un modèle de représentation des objets HTML à partir duquel on peut construire des fonctions pouvant détecter la sémantique des objets dans une page. La découverte des objets repose sur un prétraitement (découpage) similaire à [1].

**annotation basé sur la localisation** Les auteurs proposent d'annoter les objets comme étant informatifs ou non-informatif. Cette décision est prise d'après la taille et la position de chaque objet dans la page [7].

**annotation basé sur la répétition** Les auteurs proposent un partitionnement des objets comme étant informatif ou non en calculant une valeur d'entropie pour chaque bloc [5] grâce à une mesure : ITF-DF.

## 3 Adaptation de documents numériques

La consultation de fonds documents de type légataires présente une grande quantité d'information. Leurs consultations impliquent une surcharge cognitive pour les lecteurs. Les auteurs de l'article [8] propose une approche pour réduire cette surcharge en appliquant des adaptations de structurelle et de présentation du contenu suivant les préférences des utilisateurs. Ces préférences prennent en compte les informations qui intéressent le lecteur. Cette acquisition se fait par l'analyse de son comportement au travers l'interface de navigation des documents.

## Références

- [1] Deng Cai, Shipeng Yu, Ji-Rong Wen, and Wei-Ying Ma. Extracting content structure for web pages based on visual representation. In *Web Technologies and Applications*, pages 406–417. Springer, 2003.
- [2] Jinlin Chen, Baoyao Zhou, Jin Shi, Hongjiang Zhang, and Qiu Fengwu. Function-based object model towards website adaptation. In *Proceedings of the 10th international conference on World Wide Web*, pages 587–596. ACM, 2001.
- [3] Christian Kohlschütter and Wolfgang Nejdl. A densitometric approach to web page segmentation. In *Proceedings of the 17th ACM conference on Information and knowledge management*, pages 1173–1182. ACM, 2008.
- [4] Milos Kovacevic, Michelangelo Diligenti, Marco Gori, and Veljko Milutinovic. Recognition of common areas in a web page using visual information : a possible application in a page classification. In *Data Mining, 2002. ICDM 2003. Proceedings. 2002 IEEE International Conference on*, pages 250–257. IEEE, 2002.
- [5] Shian-Hua Lin and Jan-Ming Ho. Discovering informative content blocks from web documents. In *Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 588–593. ACM, 2002.
- [6] Karane Vieira, Altigran S da Silva, Nick Pinto, Edleno S de Moura, Joao Cavalcanti, and Juliana Freire. A fast and robust method for web page template detection and removal. In *Proceedings of the 15th ACM international conference on Information and knowledge management*, pages 258–267. ACM, 2006.
- [7] Chaw Su Win and Mie Mie Su Thwin. Informative content extraction by using eifce [effective informative content extractor].
- [8] Corinne Amel Zayani, Ikram Amous, and André Péninou. Adaptation visuelle de documents légataires. *Document numérique*, 12(1) :11–29, 2010.