0.1 DÃ©tection d'objets HTML

Mapping de structures [1] propose une mÃ©thode de mapping entre les pages web pour en isoler des sous-structures communes que l'on pourra identifier comme des objets. L'inconvÃ©nient est que l'on ne peut dÃ©couvrir qu'une partie des objets dans les pages web avec cette approche.

Segmentation par Pattern [4] propose un partitionnement d'une page web aprÃ¨s gÃ©nÃ©ration de celle-ci par un moteur graphique sur la base d'un modÃ¨le de prÃ©sentation dÃ©fini a priori. Le modÃ¨le de prÃ©sentation correspond Ã un pattern de mise en forme "standard" des pages web : header, footer, menu lat. gauche, menu lat. droit.

Segmentation par densitomÃ©trie [5] propose un partitionnement d'une page d'aprÃ¨s la variation de densitÃ© textuelle de chaque segment textuelle.

Segmentation par indice visuelle [3] propose un partitionnement d'une page en fonction des propriÃ©tÃ©s de mise en forme associÃ©es Ã chaque noeud du DOM. L'approche propose un dÃ©coupage de la page de la racine jusqu'aux feuilles du DOM sur la base d'heuristiques et d'un algorithme de fusion des noeuds du DOM.

0.2 Annotation d'objets

annotation basÃ©e sur les fonctionnalitÃ©s [2] proposent un modÃ¨le de reprÃ©sentation des objets HTML Ã partir duquel on peut construire des fonctions pouvant dÃ©tecter la sÃ©mantique des objets dans une page. La dÃ©couverte des objets repose sur prÃ©traitement similaire Ã [3] dÃ©coupant la page.

annotation basÃ© sur la localisation Les auteurs proposent d'annoter les objets comme Ã©tant informatifs. Cette dÃ©cision est prise d'aprÃ¨s la taille et la position de chaque objet dans la page [6].

annotation basÃ© sur la rÃ©pÃ©tition Les auteurs proposes un partitionnement des objets comme Ã©tant informatif ou non en calculant une valeur d'entropie pour chaque bloc [5].

0.3 Adaptation

La consultation de fonds documents de type lÃ©gataires prÃ©sente une grande quantitÃ© d'information. Leurs consultations impliquent une surcharge cognitive pour les lecteurs. Les auteurs de l'article [7] propose une approche pour rÃ©duire cette surcharge en appliquant des adaptations de structure et de prÃ©sentation du contenu suivant des prÃ©fÃ©rences des utilisateurs. Ces prÃ©fÃ©rences prennent en compte les informations qui intÃ©ressent le lecteur et leur prÃ©sentation. Cette acquisition se fait par l'analyse de son comportement au travers l'interface de navigation des documents.

Bibliographie

[1] Chia-Hui Chang and Shao-Chen Lui. IEPAD : information extraction based on pattern discovery. In *Proceedings of the 10th international conference on World Wide Web*, pages 681–688. ACM, 2001.

[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] Chaw Su Win and Mie Mie Su Thwin. Informative content extraction by using eifce [effective informative content extractor].

[7] Corinne Amel Zayani, Ikram Amous, and André Péninou. Adaptation visuelle de documents légataires. *Document numérique*, 12(1) :11–29, 2010.