Extraction de la structure logique d'une page web

Plusieurs approche dans l'extraction de la structure logique :

1. Analyse syntaxique
2. Par apprentissage statistique
3. Partitionnement basé sur des indicateurs visuelles

Caractéristiques de HTML :

- Langage de description de données semi-structuré.
- Décrit la structure et la présentation des données.
- Pas de relation sémantique explicite (avec HTML 4)
- Normalisation syntaxe et de la sémantique par le W3C, mais encore une implémentation hétérogènes dans les différents navigateurs web.
- De plus les concepteur de page néglige les spécifications.
- Les pages web possède un contenu hétérogènes

Solution par analyse syntaxique :
Pas d'article lu sur l'analyse syntaxique des pages web. Mais cette dernière ne semble pas adapté au contenu web. En effet, la négligence des développeurs, limite l'extraction de la structure logique de la page. Et surtout, la **signification des balises est implicite** et connu que par l'application qui à généré le document.

Solution par apprentissage statistique :
Dans les travaux lus, on veut faire correspondre le schéma d'un document d*in* à un document d*out* conforme à ce schéma. Utilisation d'une méthode d'apprentissage statistique basé sur les réseaux de Bayésien.
Il apparaît que les que les nœuds internes sont moins bien reconnu que les feuilles. L'extraction des relations entre les éléments pose encore problèmes.

Solution par partitionnement sémantique basé sur des indicateurs visuelles :
Le processus de segmentation des pages par des indicateurs visuelle semble efficace. En effet, principalement par la nature de html. C'est un langage qui inclus une description structurelle et de présentation. De plus, contrairement aux approches basées sur l'analyse de tag, une interprétation syntaxique, cette approche n'est pas limité par la mauvaise utilisation des balises par le concepteur de page web.