# Introduction

- Article : **Extracting Content Structure for Web Pages based on Visual Representation**
- Auteurs : Deng Cai, Shipen Yu, Ji-Rong Wen et Wei-Ying Ma.
- Principaux apports :
  - **Modèle page web**
  - **algorithme d'extraction de structure d'une page web**

# Approche

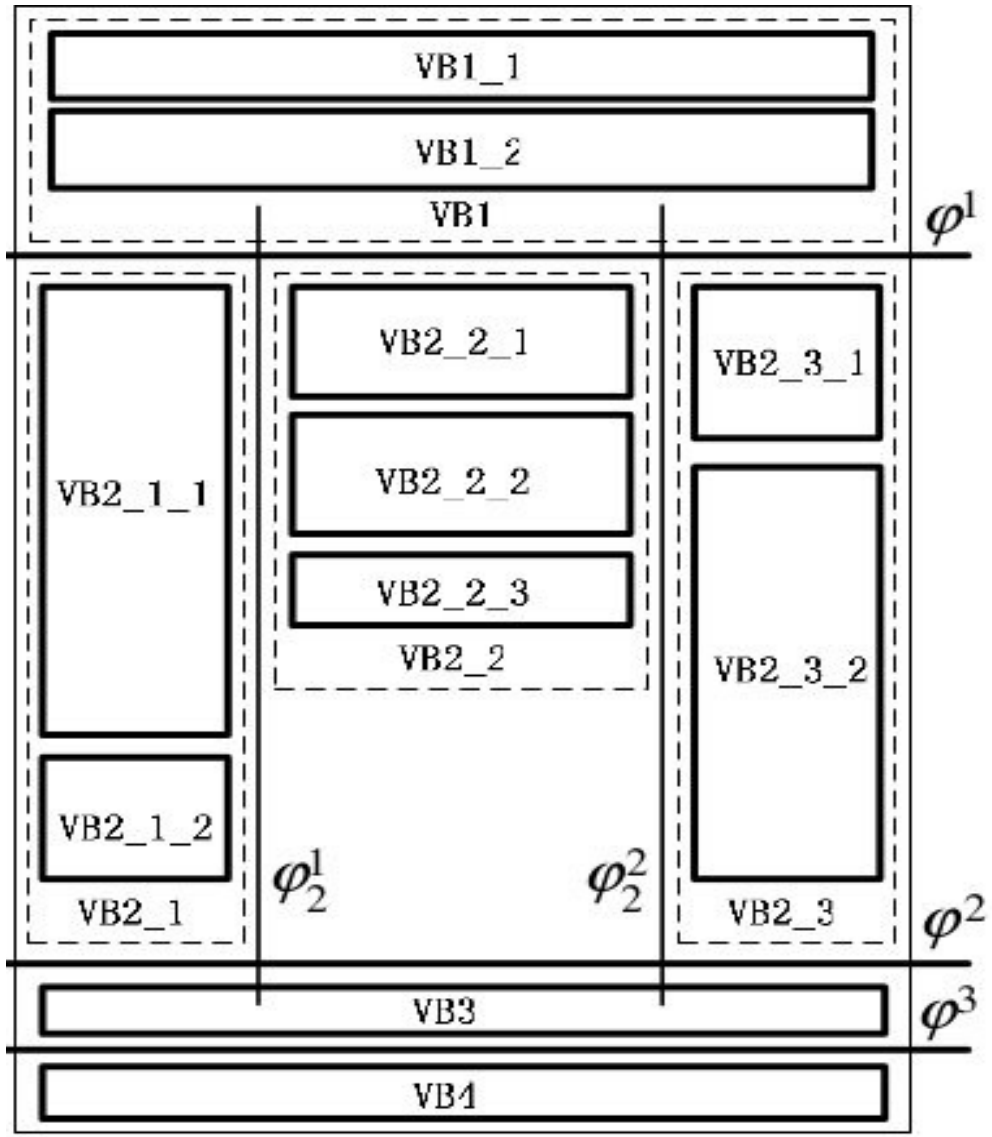
- Restructuration page web suivant une vision basé sur la représentation visuel

# Modèle page web

$$\Omega = (\Theta, \phi, \delta)$$

- **Θ est un ensemble fini d'objet ou de sous page web**
- **Φ ={φ1, φ2, φ3} est un ensemble de séparateurs visuels**
- **δ est la relation entre deux blocs dans Θ**

# Modèle page web




$$O = (VB1, VB2, VB3, VB4)$$

$$\Phi = \{\varphi^1, \varphi^2, \varphi^3\}$$

$$\delta \begin{pmatrix} (VB1, VB2) \\ (VB2, VB3) \\ (VB3, VB4) \\ else \end{pmatrix} = \begin{pmatrix} \varphi^1 \\ \varphi^2 \\ \varphi^3 \\ NULL \end{pmatrix}$$

# Synthèse VIPS

DOC : Degrés de cohérence attribué à chaque nouveau bloc construit

PDOC : Degrés de cohérence permis. Indicateur un degrés de cohérence accepté.

# Synthèse VIPS

1) Extraction de blocs visuels :
    1) Depuis la racine DOM, parcours chaque nœud du DOM.
    2) Teste *la nature du nœud* et *la distance visuelle* du nœud parent :
        1) Test vrai : ajoute le nœud au bloc parent
        2) Test faux : crée un nouveau bloc
    3) Pour chaque nouveau bloc associe un DOC

# Synthèse VIPS

2) Détection des séparateurs visuels :

1) Pour chaque pool de bloc calcule :
    1) Séparateurs implicite : ligne de pixels horizontales ou verticales qui ne coupent pas de bloc
    2) Séparateur explicite : ensemble de balise HTML, ex <HR>
2) Calcule un poids pour chaque séparateur.

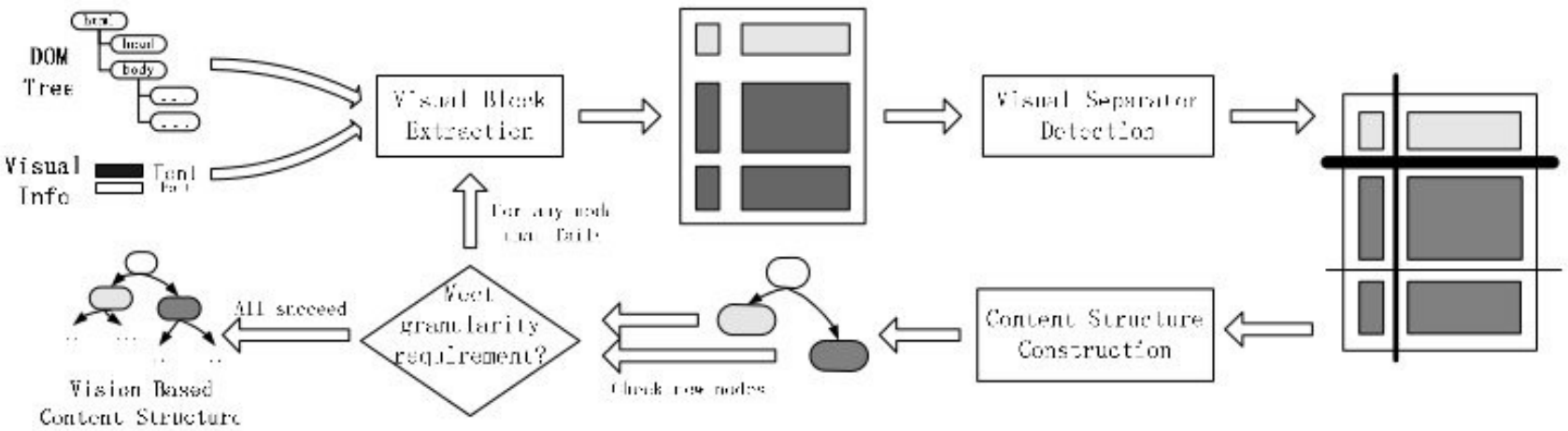
Ex : poids élevé pour des blocs plus éloignés géographiquement, couleur différente …

# Synthèse VIPS

3) Construction de la structure :

1) L'algo fusionne les blocs ayant le poids des séparateurs les plus faibles, jusqu'à ce qu'il rencontre des séparateurs de poids fort
2) Pour chaque nouveau bloc, attribut un DOC
3) Test que le DOC est inférieur au PDOC sinon recommence le processus d'extraction.

# Synthèse VIPS



# Résultats

| **Jugement** | Nombre de page |
|---|---|
| Parfait | 86 |
| Satisfaisant | 50 |
| Echoué | 4 |

« 97 % des structures sont correctement reconnu »

# Conclusion

- Les expérimentations montrent que l'algorithme fonctionnent bien
- Par rapport à notre problématique, on extrait la structure du document mais on ne connaît pas le rôle des éléments