1. Principaux apport du papier

L'article propose une approche pour reconstruire la structure d'une page correspondant à la perception visuelle humaine. Cette approche définit **un modèle de page web** et **spécifie un algorithme** qui partitionnement successivement les éléments d'une page web conformément au modèle définit à priori.
Cette approche est indépendante de la représentation sous-jacente du document comme html. Elle fonctionne même si la structure est très différente de la structure de mise en page.

2. Fil conducteur :

Une page web présente plusieurs structures sémantique**.**
Les navigateurs web offrent une représentation en 2 dimensions d'une page web ainsi que de nombreux indicateurs visuels qui permettent de distinguer les différentes structures d'une page.
Les contenus sémantiquement proches sont principalement regrouper ensemble, la page est ainsi divisé en région par des **s**éparateur visuels implicites ou explicites.

3. Approche :

- Vision basé sur la représentation visuelle:

L'approche propose un segmentation du contenu basé sur la représentation visuelle. On reconstruit une nouvelle structure à partir du DOM dont les nœuds représentent des blocs. Ces blocs sont des regroupement d'élément du DOM en agrégation sémantique cohérente (Les éléments de la nouvelle structure n'ont pas nécessairement de correspondance avec les éléments du DOM).

Proposition d'un modèle de page web :
Pour cela, on définit un nouveau modèle d'une page web définit par le triplet : **Ω = (Θ, ϕ, δ)**.

- Θ est un ensemble fini d'objet ou de sous page web.
- Φ ={φ1, φ2, φ3} est un ensemble de séparateurs visuels. Chaque séparateur visuel possède un poids indiquant la visibilité et tous les séparateurs dans le même Φ ont le même poids.
- δ est la relation entre deux blocs dans Θ et δ = Θ x Θ → ϕ U {Null}.

Indicateur de cohérence :
Pour chaque bloc visuel (ou objet) identifié, on attribut un degré de cohérence (DOC). C'est un indicateur de la granularité de la cohérence entre les blocs. Le *DOC* d'un bloc enfant est toujours plus petit que le bloc parent.

La détection des blocs se fait sur des propriétés telles que la couleur, la taille ou la police des éléments du DOM

Seuil de cohérence :
Le **degrés de cohérence permis** (PDOC) indique le seuil de granularité satisfaisant. Il indique la granularité de la structure en sortie. Ainsi le processus de segmentation est stoppé seulement si le DOC du bloc traité est inférieur au PDOC

- L'algorithme VIPS :

L'algorithme déduit cette nouvelle structure par la structure du DOM et des indicateurs visuels qui sont récupérés du navigateur web.
Depuis la racine du DOM commence un **processus d'extraction de bloc**. Pour chaque nœud on vérifie si il peut constituer un nouveau bloc ou non .La décision est prise en vérifiant *la nature du nœud* (, <table>) courant ou *la distance visuel* avec le nœud parent ou frère (dans certain cas) :

- Si le nœud est un séparateur l'algorithme marque les frères directs du nœud courant comme des parents du nouveau bloc
  ex : , <p>, <table>
- Si la fonction de *distance visuel* est valide, le nœud courant est ajouté au bloc parent direct. Sinon l'algorithme crée un nouveau bloque et spécifie le bloque courant comme parent.
- Pour chaque nouveau bloc, on associe un DOC

Note : La fonction de distance visuel est le nombre de différence de style visuel du nœud du DOM : largeur, hauteur, taille de la police, couleur d'arrière plan, etc …

Les auteurs proposent une catégorisation des nœuds :

- Inline Node : nœuds qui cause une nouvelle ligne dans la page
- Line-break node :
- Invalide node :
- Partially node :
- etc …

Ces définitions conjointement utilisés avec les indicateurs visuels permettent de définir des pour la segmentation des pages.

Exemple :

1. Si tous les enfants sont des nœuds sont des nœuds de type virtual text node placé dans

Une fois les blocs extraits l'algorithme effectue un processus de **détection des séparateurs visuels** (implicites ou explicites).

1. L'algorithme **calculer des séparateurs** visuel horizontale et verticale. Un *séparateur visuel implicite* correspond à une ligne (verticale ou horizontale) de part et d'autre de la pool qui ne traverse pas de bloc. Un *séparateur visuel explicite* est une catégorie de balise HTML. Comme par exemple la balise <hr> .
2. Puis l'algorithme attribut un poids pour chaque séparateur.
   Ex : poids élevé pour des blocs plus éloignés géographiquement, couleur différente …

L'intérêt de calculer des séparateurs est de distinguer les blocs ayant une sémantiques différentes.

La dernière phase est la **construction de la structure**. L'algorithme fusionne les blocs ayant les séparateur de plus petit poids jusqu'à ce qu'il rencontre les séparateurs d'un certain poids. Pour chaque nouveau bloc on lui attribut un DOC. Si il est supérieur au PDOC, l'algorithme reviens au processus d'extraction.

4. <u>Résultats :</u>

La méthode à été testé sur un corpus de 140 page web, les plus consultés d'après Yahoo. Les structures obtenues ont été soumis à un jury humain.

Juge	**Nombre de page**
Parfait	86
Satisfaisant	50
Échoué	4

On constate donc que 97 % des structures sémantiques sont correctement reconnu. Il n' y a pas de résultat sur les ressources utilisés par l'algorithme.

5. <u>Intérêt d'avoir lu ce papier :</u>

Ce papier propose un algorithme qui permet un partitionnement d' une page en bloc sémantique cohérent, les expérimentations montrent que la méthode fonctionne plutôt bien.
Au regard de notre problématique, ces travaux nous donne un algorithme pour extraire la structure d'une page web, mais ne nous permet pas de connaître les rôles associés à chaque éléments de la structure.

L'article propose une méthode pour d'extraction de structure d'une page web basé sur une **représentation visuelle.** Cette méthodologie vise à aider des applications web de recherche d'information et d'adaptation de contenu.

L'approche recherche à reconstruire la structure d'une page web correspondant à la perception humaine, ce qui **permet de mieux refléter la structure logique.**
Ce papier propose un algorithme VIPS (VIsion-based Page Segmentation) qui exploitent les caractéristiques de mise en page pour un partitionnement de la page en niveau sémantique.
En partant de la racine du Dom, l'algorithme parcours chaque niveau de DOM et le segmente en **bloc visuel** en se basant sur des règles heuristiques. Chaque bloc visuel est rangé dans un pool correspondant à son niveau dans le DOM. Pour Chaque pool crée un nouveau processus de **détection de séparateurs** est lancé. Ce dernier crée des séparateurs correspondant au ligne de pixel horizontale ou verticale de part et d'autre de la pool et qui ne sont pas coupé par un bloc. Pour chaque séparateur ont lui attribut un poids. L'intérêt de trouver des séparateur entre les blocs d'une même pool est distinguer les blocs ayant une sémantique différente.
Une fois les étapes d'extraction de bloc et détection des séparateurs effectuer, VIPS lance un processus de reconstruction du contenu en fusionnant les blocs ayant des séparateurs de faible poids.

L'algorithme ne prend en compte seulement de la structure statique de la page et que les éléments visibles de la page. L'algorithme peut être amélioré en prenant en compte d'autre heuristique basé sur les balises HTML5 et les normes d'accessibilités.

Ce papier propose un approche pour analyser la structure logique d'une page web en s'appuyant sur un **partitionnement sémantique** de la structure du DOM. L'algorithme simule la compréhension d'une page web par un humain en s'aidant des représentations visuelles.

Extraction de la structure d'une page web basée sur la représentation visuel de celle ci.
Processus ascendant pour comprendre de manière automatique comment un utilisateur comprend l'agencement d'une page web.
Propose une méthode pour découper une page en fonction de *repère visuel* implicite et/ou explicite. Découpage selon des critères humains.

Page web n'est pas une unités d'informations atomiques. Contient des informations hétérogènes.
Partitionnement des informations contenu dans une page, facilite l'adaptation de contenu pour les terminaux mobiles qui ont un petit écran, **facilite la navigation globale d'une page.**
Les utilisateurs visualises les pages web à travers un navigateur web, ils ont ainsi une représentation en deux dimension d'une page web. possède des repères visuelles pour aider à distinguer les différentes parties de celle-ci.
Le contenu sémantiquement liée est regroupé ensemble. La page entière est découpé en région pour chaque contenu différent en utilisant des séparateurs explicites et implicites (ex : ligne, couleur

ect...).
Le travail s'appuie donc sur ce constat pour regrouper les éléments sémantiques liées entre eux en bloc en s'appuyant sur les représentations visuelles.

Partitionnement : http://fr.wikipedia.org/wiki/Partitionnement_de_donn%C3%A9es