L'analyse des modèles de pages par les moteurs de recherche : les approches explorées par Google, Yahoo! et Google

Retour au sommaire de la lettre

Domaine :	Recherche	**Référencement**
Niveau :	Pour tous	**Avancé**

Tous les liens (backlinks) sont-ils égaux pour les moteurs de recherche ? A priori non, si l'on en croit les nombreux travaux de recherche que mènent Google, Yahoo! et Microsoft pour donner un poids différent à un backlink en fonction de la zone où il se trouve dans la page web : un contenu rédactionnel "pointant" vers une page distante aura ainsi une force plus importante que lorsque le lien se trouve au fin fond d'un footer. Petite revue d'effectifs des différents travaux actuellement... référencés dans ce domaine à travers le monde...

Avant l'explosion de la Toile, les travaux sur l'IR (*Information Retrieval* : extraction d'information) portaient sur des ensembles de pages issus des supports imprimés, c'est-à-dire des textes bruts faiblement ou pas du tout structurés ou hiérarchisés. Le développement du World Wide Web a constitué un tournant radical, grâce à la création d'un nouveau type de documents, reliés entre eux par une structure hypertexte, et structuré par le code HTML. Les moteurs de recherche ont cherché très tôt à exploiter les informations issues de ces deux structures pour améliorer leurs algorithmes. Mais force est de constater que cette exploitation est restée longtemps très limitée...

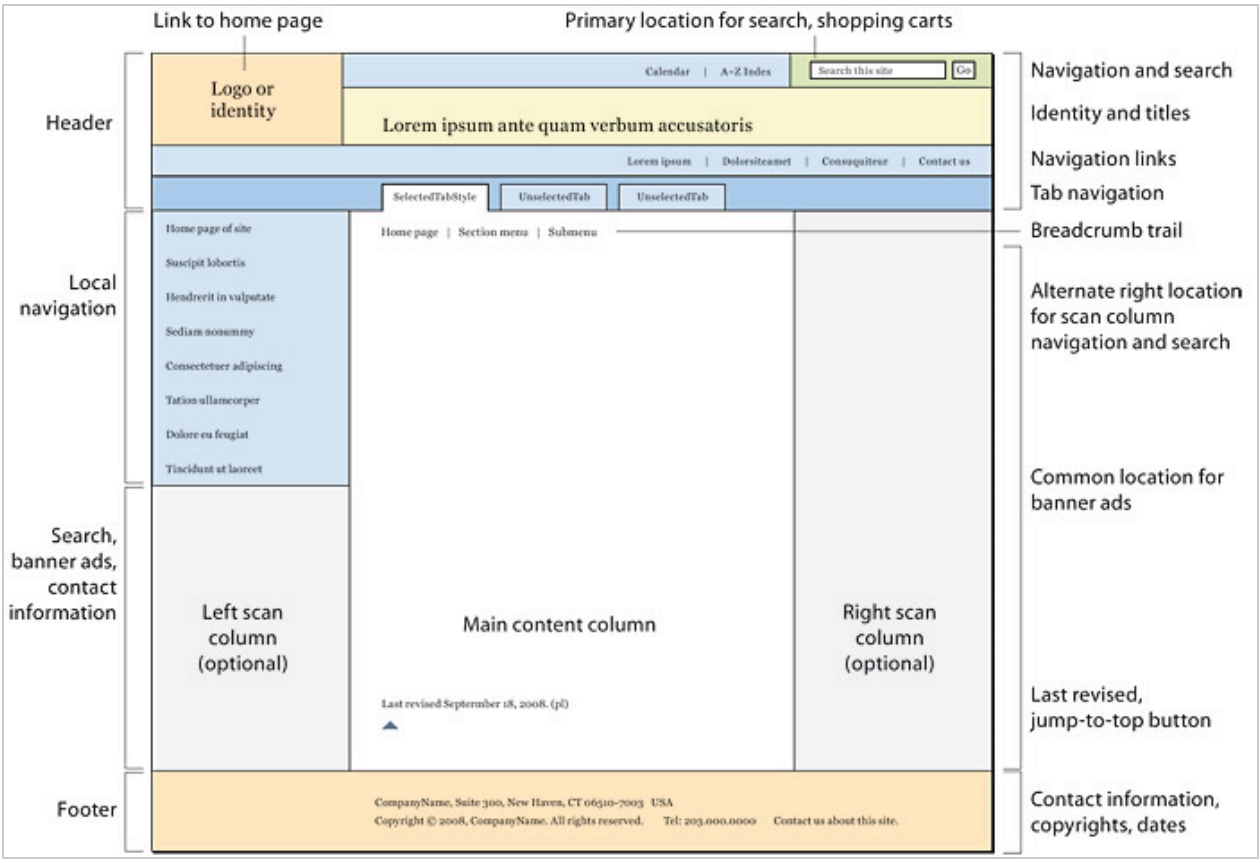
Depuis quelques années, on voit fleurir des articles scientifiques et des brevets décrivant des améliorations des algorithmes des moteurs à partir de critères issus de l'analyse des "*templates*" (modèles de page). Ces innovations ont été peu commentées, pourtant leur utilisation potentielle dans les moteurs conduit a des conséquences immédiates sur la manière de concevoir le code des pages web, leurs contenus, et les blocs de navigation. Cette utilisation par les moteurs conduit également à devoir affiner certaines méthodes de référencement !

Dans cet article, nous allons nous intéresser à trois approches explorées par Bing, Yahoo et Google, et qui illustrent bien cette tendance actuelle. Nous commencerons par un avertissement habituel dans ce genre d'articles : le fait que les laboratoires de recherche travaillent dans ces directions, ou qu'un brevet ait été déposé par un moteur ne doit pas conduire à penser que ces méthodes sont réellement utilisées ou implémentées de la façon décrite (même si dans certains cas, l'utilisation de ces approches dans Bing et Google est parfaitement avérée, on le verra plus loin dans l'article). Par contre, avec le recul, on peut toujours conclure que la multiplication des travaux dans un domaine révèle un vrai problème rencontré par les moteurs, ce qui est en soi intéressant à connaître. Et une convergence entre les solutions trouvées révèle probablement une évolution à venir des moteurs.

Quels sont les problèmes que l'analyse des templates permet de résoudre ?

L'analyse de la structure des pages permet d'obtenir deux familles d'améliorations dans les algorithmes des moteurs. Tout d'abord une évaluation plus précise et plus pertinente du poids des liens, c'est à dire une amélioration des algorithmes de la famille du Pagerank de Google. Ensuite une meilleure identification des zones recelant du contenu original dans les pages.

Sur ce dernier point, on peut remarquer que les pages web sont composées classiquement de "blocs" ou "segments". On trouve en général un "header" (ou "en tête") et un footer (pied de page) que l'on retrouve sur la plupart des pages du site, ensuite des blocs de navigation, ou des menus, des blocs réservés à la publicité, aux liens sponsorisés, aux liens partenaires.



Un exemple de modèle de page classique
(illustration extraite de Web Style Guide 3e édition :
http://www.amazon.com/exec/obidos/ASIN/0300137370/webstyleguidecom).

Ce qui est intéressant à noter, c'est qu'il n'est pas rare que l'essentiel des mots clés présents sur une page web ne soient pas liés au contenu spécifique à cette page mais à des blocs ou des menus qui se retrouvent sur de multiples pages ou des blocs annexes sans rapport direct avec la page. Dans ces conditions, les approches classiques comme le *Cosinus de Salton* pour déterminer la proximité sémantique entre une requête et un document peuvent être fortement perturbés par la présence d'une majorité de termes "hors sujet" dans la page !

Pour détecter et comprendre la structure d'une page web, le plus simple est d'utiliser les informations que le code HTML recèle, en s'appuyant notamment sur la détection des balises définissant des "zones" dans la page, comme les balises <table>, , <form> etc...

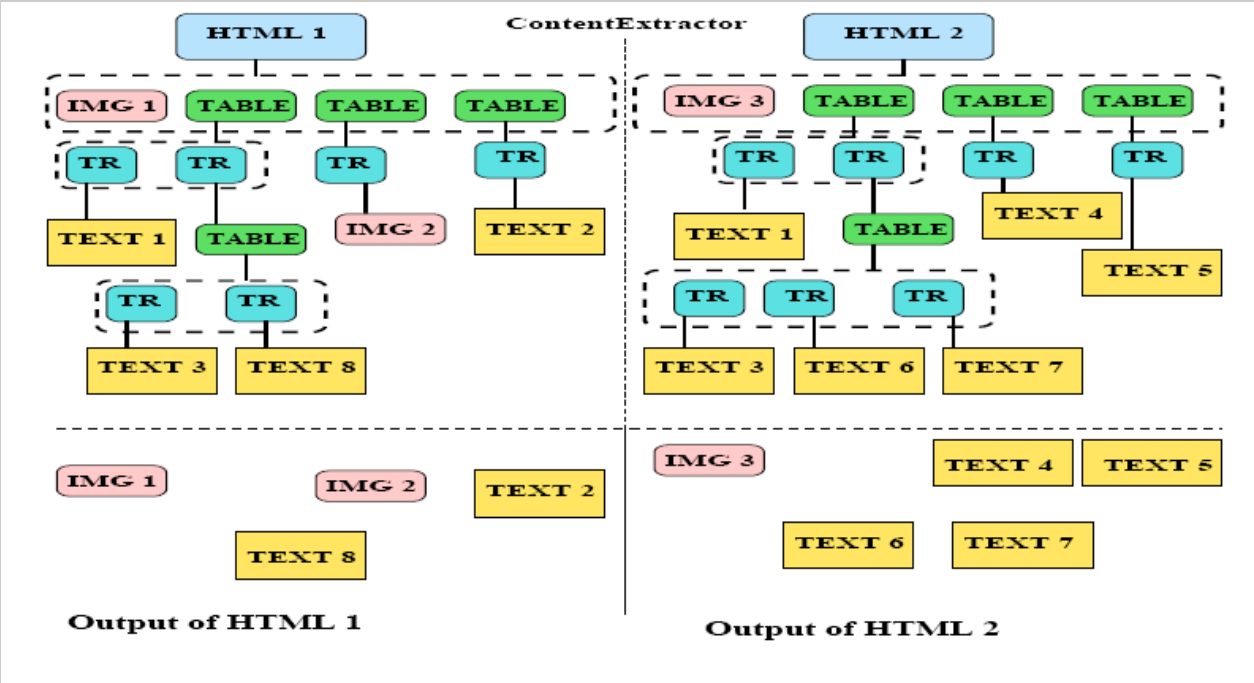


Schéma montrant un exemple d'analyse de contenu d'une page
en s'appuyant sur la structure DOM (hiérarchie du code HTML de la page)

Le Block Level PageRank pour corriger les défauts du Pagerank

Dans l'algorithme original du Pagerank, décrit dans l'article de 1998 de Page et Brin, tous les liens étaient traités de manière égale. Peu importait l'emplacement de ce lien sur la page ou sa visibilité, chaque lien était susceptible de transmettre une quote-part égale du pagerank de la page de départ à la page de destination.

Cette vision où tous les liens sont égaux entre eux semblait avoir une certaine légitimité dans les années quatre-vingt dix, à une époque où beaucoup de pages web étaient statiques et composées à la main par les webmasters. Quelques années plus tard, la plupart des pages sont composées par des programmes, à partir de modèles de pages. Ce qui signifie que la proportion de liens créés volontairement et directement par des humains sur les pages web a fortement diminué. Qui plus est, le web est devenu marchand, et le théâtre d'enjeux importants : une proportion non négligeable des liens présent sur les pages sont en fait des liens publicitaires ou des liens "partenaires" qui mettent à mal la neutralité des liens entre les pages web.

Dans ce contexte, l'utilisation de l'algorithme du Pagerank originel pose quatre types de problèmes :

- Tous les liens ne sont pas neutres : il faut pouvoir traiter les liens publicitaires autrement. D'où les injonctions de Matt Cutts : placez un nofollow sur les liens publicitaires et assimilés, ou redirigez les par une redirection 301 vers une page bloquée par un Robots.txt. Tous les liens publicitaires perturbent le calcul d'un pagerank, pas uniquement les liens achetés spécifiquement pour faire augmenter le pagerank...

- Tous les liens n'ont pas la même signification : quelle est l'importance respective d'un lien situé dans le footer, dans le menu horizontal du header, dans un bloc de navigation en haut de colonne de gauche, dans un bloc de liens connexes en bas de colonne de droite, ou figurant au milieu de la page en pleine zone de texte.

- Les liens des blocs de navigation, présents sur de nombreuses pages, ne doivent pas être traités comme les liens figurant dans le contenu spécifique à une page donnée

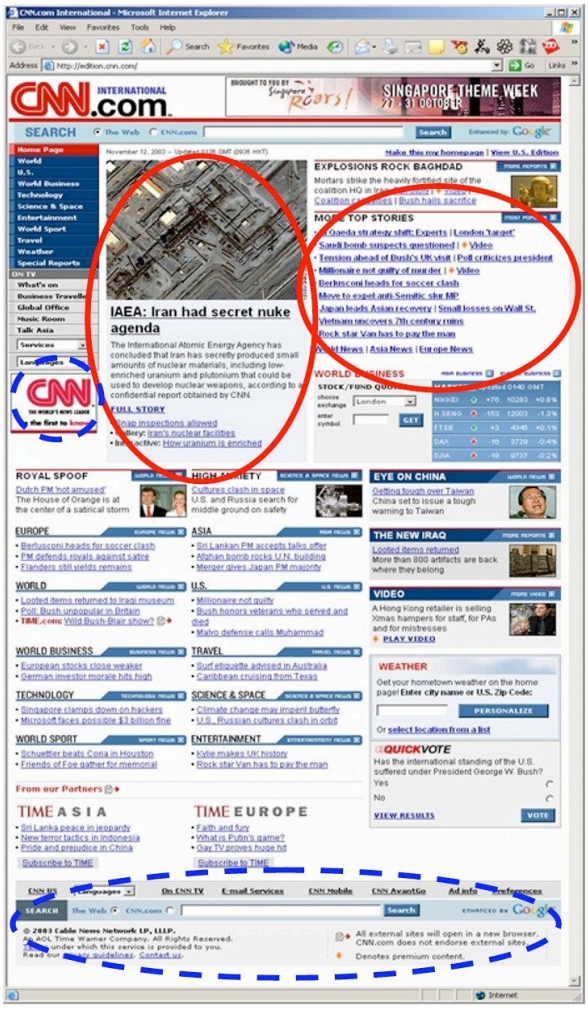
- Tous les liens ne sont pas aussi visibles les uns que les autres et aussi cliqués les uns que les autres !

Sur ce dernier point, on relira la partie de notre article sur le pagerank paru en juin 2009 dans la lettre professionnelle d'Abondance, et notamment la partie consacrée au modèle du surfer "poursuivant un objectif" opposé au modèle du surfeur aléatoire.

Une solution à ce problème a été proposée dès 2004 par des chercheurs du laboratoire de Microsoft à Pékin (c'est l'un des laboratoires les plus importants travaillant sur les outils de recherche).

L'équipe dirigée par Jirong Wen a déterminé une méthode pour détecter sur les pages la présence de blocs de code identifiables (*header, footer*, menu, blocs de navigation etc...). Ils ont ensuite modifié la "formule" de calcul du Pagerank pour tenir compte de la nature différente des liens présents dans chacun de ces blocs. Ils ont baptisé cette nouvelle formule le **Block Level Pagerank** (BLPR, à ne pas confondre avec le BlockRank, qui est une méthode permettant de calculer le PR d'une nouvelle page à partir des liens des pages situées à proximité, sans avoir à recalculer la matrice entière des liens du web).

Le BLPR est une méthode de calcul d'un critère de popularité par les liens qui prend en compte les relations hypertextes non pas entre les pages mais entre les blocs. Comme dans le graphique ci-dessus, la méthode permet aussi de repérer dans les pages les zones qui sont liées à une rubrique ou une thématique précise.



La méthode VIPS de Microsoft

L'approche inventée par l'équipe de Jirong Wen s'appuie sur l'identification de blocs, selon une méthode qu'ils ont appelé VIPS (*Vision based Page Segmentation*). VIPS s'appuie essentiellement sur l'analyse de la structure DOM de la page *(Document Object Model*), mais en exploitant en plus toutes les informations de mise en forme contenues dans le code html (taille de la police, graisse, couleur, existence de séparateurs etc...).

La structure DOM décrit l'architecture logique des documents. Une page HTML est ainsi décomposée en balises qui s'imbriquent les unes dans les autres et créent une arborescence logique. Une balise <TD> (cellule de tableau) est logiquement rattachée à une balise <TR> (ligne de tableau)

rattachée elle même à une balise <TABLE> rattachée à une balise <DIV> etc.

Cette approche permet d'extraire des informations exploitables pour hiérarchiser le contenu des pages web. Par exemple, la page ci-contre, extraite du site de CNN présentée dans l'article de Jirong Wen présente clairement des zones où l'information est intéressante (entourées en rouge) et des zones moins importantes comme le footer ou certains espaces d'autopromotion (entourés en bleu).

La méthode VIPS permet d'aller encore plus loin... En effet, l'analyse du code HTML permet, en identifiant les séparateurs importants et pertinents, de segmenter la page en grandes zones cohérentes. L'illustration ci-dessous montre le résultat pour la page de CNN.

Ensuite, d'autres critères sont appelés à la rescousse pour hiérarchiser l'importance des blocs comme par exemple :

- Le nombre et la taille des images dans le bloc,
- Le nombre de liens, le nombre de mots dans le texte des liens, situés dans le bloc
- Le nombre de mots dans le texte du bloc
- Le type d'interaction proposé à l'utilisateur, en se basant sur le nombre de champs input ou la présence de formulaires
- L'emplacement du bloc (header, footer, à gauche, à droite...).

Ce qui donne, dans le cas de page CNN, une hiérarchisation à quatre niveaux de l'importance des pages (voir l'illustration ci-contre).

On peut aussi, grâce à une analyse des liens entre les blocs de différentes pages, identifier des relations particulières entre les pages. Ainsi, on peut parfaitement relier, comme dans l'exemple ci dessous, tel bloc sur une page d'actualité avec la rubrique thématique associée. Ce qui permet de mieux comprendre la nature des blocs de navigation.

Des approches similaires chez Yahoo et Google ?

Microsoft a exploré cette voie dès 2003/2004 semble-t-il. Chez Yahoo et Google, les signes de l'existence de travaux sur l'analyse des templates sont plus récents, mais ils démontrent qu'aujourd'hui les trois principaux moteurs explorent ces approches nouvelles.

En juillet 2006, Google publie un brevet baptisé "*Segmentation de document en fonction des lacunes visuelles*". Il s'agit d'une méthode de segmentation des templates qui pousse plus loin que la méthode VIPS la modélisation de l'apparence visuelle d'une page dans un navigateur. Dans l'outil de Microsoft, on cherche avant tout à déduire de la structure le découpage en blocs cohérents et à situer grossièrement ces blocs dans l'espace à deux dimensions que constitue la fenêtre du navigateur (en haut, en bas, à gauche, à droite... ?). Le brevet de Google décrit une tentative pour repérer l'existence d'espaces vides à l'intérieur des blocs, ou d'autres séparateurs.

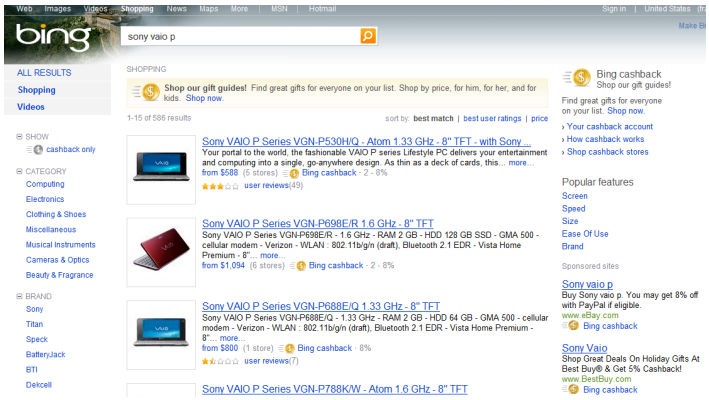
Le brevet de Google développe également une application de cette segmentation à l'identification d'informations géographiques reliées à un lieu dans une page. Cette approche permet de découvrir automatiquement des "attributs" qu'il est possible de relier à ce nom de lieu, sans confusion possible avec des commentaires faits sur d'autres lieux évoqués sur la même page. Cette même méthode est applicable pour identifier d'autres types d'attributs.

En 2007, un article de Chakrabarti, Kumar et Punera présentait une méthode étudiée par Yahoo (*Détection de templates à l'échelle d'un site entier par lissage isotonique*). Elle démontre que les chercheurs de Yahoo ont également tenté d'exploiter l'information issue du code HTML et en particulier de la structure DOM pour segmenter la page en différentes zones. Le "lissage isotonique" représente par ailleurs une approche originale pour pondérer subtilement en fonction de l'emplacement d'un élément les critères calculés au moment de l'indexation. L'étendue des recherches chez Yahoo! est révélée par de multiples articles signés par Kumar et/ou Punera sur ce sujet (voir Bibliographie).

En 2008 on découvre également un brevet publié par Yahoo! décrivant une méthode permettant d'analyser la structure DOM pour identifier les zones de contenu intéressantes dans les pages.

Ces méthodes sont elles réellement utilisées dans les moteurs de recherche ?

Les approches inventées par Microsoft sont officiellement utilisées dans la partie shopping de Bing, et dans le moteur Libra de Microsoft (http://libra.msra.cn)



Sur la rubrique "shopping" de Bing, les attributs qui apparaissent en regard de chaque produit sont collectés selon la méthode décrite dans l'article "*Object-level Vertical Search*" (voir Bibliographie).

Google utilise apparemment une méthode similaire au brevet sur les "lacunes visuelles" pour enrichir sa recherche locale.

Que faut-il en conclure d'un point de vue SEO ?

Il y a dix ans, les moteurs exploitaient peu les informations du code HTML, excepté certaines balises importantes comme par exemple le <title> de la page. De nombreuses balises servant uniquement à la présentation étaient purement et simplement ignorées dans le processus d'indexation. L'évolution récente démontre que les moteurs cherchent à présent à exploiter toutes les informations contenues dans le code des pages. Cela signifie que la manière dont les pages sont codées n'est plus aussi "neutre" qu'autrefois, et la qualité du contenu d'une page d'un point de vue SEO n'est plus uniquement liée au texte brut présenté, mais aussi à la manière dont est structuré, hiérarchisé et présenté ce contenu.

Cela ne signifie pas pour autant qu'un code valide ou structuré en utilisant des balises sémantiques donne un avantage en matière d'optimisation pour les moteurs de recherche. Google classe très bien des pages non valides. Par contre, Bing semble aujourd'hui tellement utiliser la structure DOM des pages pour extraire de l'information. Ainsi, des erreurs de code peuvent être pénalisantes, notamment des balises de structures mal fermées, ou une hiérachie dom mal respectée.

Par contre il semble de plus en plus indiqu de faire attention à créer un code simple et clair permettant de distinguer facilement la structure en blocs de la page. En effet, en règle générale, on ne cherchera pas à éviter que la nature d'un bloc soit identifié, mais au contraire à éviter une mauvaise identification des contours d'une zone dans un template.

Par ailleurs, on voit que les moteurs disposent de méthodes opérationnelles pour diffuser différemment le pagerank en fonction de l'emplacement des liens. Cela signifie que des différences de traitement peuvent apparaître entre des liens figurant dans une zone informative de la page, un lien en footer, et un lien dans un bloc ou un menu de navigation. Ces différences peuvent conduire à faire des choix plus subtils dans la construction de modèles de page. En général, les choix dictés par l'ergonomie seront également des bons choix pour le SEO, à quelques nuances près.

Mais le temps où on pouvait impunément ajouter des liens dans des zones non visibles et peu cliquées des templates de pages est peut être bientôt totalement révolu. Peut être est-ce déjà en partie le cas ?

BIBLIOGRAPHIE

Le rôle exact du pagerank en juin 2009 par Philippe YONNET in La lettre professionnelle d'Abondance, Juin 2009

MSN Search utilise-t'il l'analyse au niveau des blocs ? par Philippe YONNET, Webmaster-Hub février 2005
http://www.webmaster-hub.com/publication/MSN-Search-utilise-t-il-l-analyse.html

Block Level Pagerank :

Block-level Link Analysis par Deng Cai1* Xiaofei He2* Ji-Rong Wen* Wei-Ying Ma* - Microsoft Research/Juin 2004
http://research.microsoft.com/en-us/people/jrwen/block-level_link_analysis.pdf

A propos de la méthode VIPS :

VIPS: a Vision-based Page Segmentation Algorithm - Technical Report
Deng Cai,Shipeng Yu,Ji-Rong Wen,Wei-Ying Ma - Nov. 1, 2003

VIPS: a Vision-based Page Segmentation Algorithm
Block based web search - Deng Cai ; Shipeng Yu ; Ji-Rong Wen ; Wei-Ying Ma - Microsoft Research/ Juin 2004
http://research.microsoft.com/pubs/69113/21.pdf

Object Level Ranking: Bringing Order to Web Objects
Zaiqing Nie,Yuanzhi Zhang, JiRong Wen, WeiYing Ma
http://www2005.org/cdrom/docs/p567.pdf

Object-level Vertical Search, Zaiqing Nie, Ji-Rong Wen
Web Search and Mining Group, Microsoft Research Asia, Beijing, China
http://research.microsoft.com/en-us/um/people/znie/cidr2007-nie.pdf

Microsoft Research Asia at the Web Track of TREC 2003.
http://www.dbs.informatik.uni-muenchen.de/~spyu/paper/microsoft-asia-web.pdf

Yahoo! :

Page-level Template Detection via Isotonic Smoothing,
Deepayan Chakrabarti, Ravi Kumar, Kunal Punera
http://www.cs.cmu.edu/~deepay/mywww/papers/www07-templates.pdf

Autres travaux :

Automatic Identification of Informative. Sections of Web-pages. Sandip Debnath1, 3, Prasenjit Mitra2, Nirmal Pal3, C. Lee Giles1,2,3 ...
www.personal.psu.edu/faculty/p/u/pum10/tkde05-final.pdf

The volume and evolution of web page templates. In Proc. 14th WWW (Special interest tracks and posters), pages 830–839, 2005.
D. Gibson, K. Punera, and A. Tomkins.
http://www2005.org/cdrom/docs/p830.pdf

Hierarchical topic segmentation of websites.
R. Kumar, K. Punera, and A. Tomkins. In Proc. 12th KDD, pages 257–266, 2006.
http://research.yahoo.com/files/paper_4.pdf

Learning to remove internet advertisement.
N. Kushmerick In Proc. 3rd Agents, pages 175–181,1999.

BREVETS

Microsoft :

Method and system for calculating importance of a block within a display page
Invented by Wei-Ying Ma, Ji-Rong Wen, Ruihua Song, Haifeng Liu
Assigned to Microsoft, US Patent 7,363,279, Granted April 22, 2008, Filed April 29, 2004
Google

Document segmentation based on visual gaps
Invented by Daniel Egnor, US Patent Application 20060149775, Published July 6, 2006, Filed: December 30, 2004

Yahoo! :

Techniques for approximating the visual layout of a web page and determining the portion of the page containing the significant content
Invented by Anandsudhakar Kesari, US Patent Application 20080033996, Published February 7, 2008, Filed August 3, 2006

Philippe Yonnet, ***Directeur Technique @Position (http://www.aposition.com) et président de l'association SEO Camp (http://www.seo-camp.org/)***

Réagissez à cet article sur le blog des abonnés d'Abondance :
http://abonnes.abondance.com/blogpro/2009/12/lanalyse-des-modeles-de-pages-par-les.html