

## TOPOLOGIE DU WEB ET VISUALISATION DE L'INFORMATION : UN ETAT DE L'ART SCIENTIFIQUE ET TECHNIQUE

*Cette étude se propose de dresser un bilan des recherches et réalisations actuelles interrogeant l'articulation entre topologie et sémantique sur le Web d'une part, et la question de la visualisation de l'information d'autre part. Ceci dans un cadre visant à rassembler faits et concepts permettant de penser cet objet, en vue d'instrumenter son utilisation.*

*Les deux domaines étant très jeunes, même s'ils se réfèrent à des traditions anciennes, la plupart des sources sont disponibles sur le Web.*

### 1. TOPOLOGIE DES RESEAUX HYPERMEDIAS OUVERTS

Le Web peut se définir techniquement comme un réseau hypermédia, ouvert, au sens où son évolution en termes de contenu et de structure n'est pas dictée par une autorité homogène. En effet, la création et la suppression d'éléments et d'hyperliens peuvent se faire par plusieurs entités humaines ou informatiques à n'importe quel moment. C'est une première et fondamentale différence avec les autres univers documentaires, tels les bibliothèques ou les réseaux documentaires centralisés (par exemple, l'aide de Windows est un réseau documentaire, mais dont la gestion relève d'une seule autorité). Malheureusement, il existe une tradition importante de la documentique papier et de gestion de la connaissance textuelle, mais dont les résultats et hypothèses ne peuvent souvent pas se confronter à un univers dynamique et ouvert d'une échelle telle que celle du Web.

En effet, Bergmann [BER 00] estimait en 2000 le nombre de documents sur le Web à  $6.10^{11}$ , en incluant le *deep Web* i.e. les bases de données générant automatiquement des pages. Les pages indexables par les moteurs de recherche classiques à base de crawlers<sup>1</sup> sont estimées à un milliard en 99 par Lawrence – ce chiffre étant en augmentation de type exponentielle. On sait par ailleurs [LAW 99] que les moteurs de recherches n'indexent que 16% des documents du Web de surface. Techniquement, ces « documents » se définissent par un fichier ayant une URL ou adresse unique, ce qui ne suffit pas à les faire concorder à la notion traditionnelle de document. Ces fichiers peuvent être de multiples formats, l'économie du Web étant encline à produire des standards nombreux et quelques normes. La seule donnée qui ne change pas est donc la façon dont sont reliés les documents.

C'est à partir de là que l'on peut envisager de construire l'instrumentation de la lecture sur le Web, ce qui nécessite des outils conceptuels pour penser la topologie du réseau. En particulier, dans le champ des mathématiques, on prospecte depuis les années 60 un outil intéressant. En effet, les réseaux de documents peuvent se modéliser à l'aide d'un modèle intuitif et puissant : les

<sup>1</sup> Robot logiciel parcourant le Web à partir d'un ensemble de points d'entrée en suivant les hyperliens présents dans chaque page.

graphes. Chaque document est un nœud, lié par des arcs orientés (même parfois nommés) aux autres.

### **La théorie des graphes découvre la topologie**

Une littérature anglophone abondante s'est développée autour du traitement du Web comme un graphe, surtout à partir des articles de Kleinberg [KLE 98+99a+99b]. L'autre grand nom du domaine de la théorie des graphes est Barabasi, qui s'intéresse aux modèles mathématiques expliquant les propriétés de certains types de graphes [BAR 99].

Mais tout commence avec le modèle de graphe aléatoire d'Erdős et Rényi [ERD 60] qui décrit chaque nœud avec une probabilité  $p$  d'être en relation avec un des autres. Ce modèle conduit à une probabilité qu'un site ait  $k$  nœuds suivant une loi de Poisson :  $P(k) = e^{-\lambda} \cdot \lambda^k / k!$  avec  $\lambda = N \cdot (N-1) \cdot p^k \cdot (1-p)^{N-1-k}$ . Ce qui revient à dire que la probabilité de trouver un nœud très connecté diminue de manière exponentielle avec le nombre de liens attachés à ce nœud. On s'attendait donc, vu que le Web est un graphe orienté (les arcs ont un sens de parcours), à ce que les nœuds aient une répartition de liens entrants  $k_{in}$  et sortants  $k_{out}$  suivants une loi binomiale, qui convergerait en une distribution de Poissons pour un nombre de nœuds important.

#### *Barabasi et l'invariance d'échelle*

Barabasi et al. ont montré [BAR 99b] que les échantillons collectés par leurs soins sur le Web ne suivaient pas cette loi, mais plutôt une loi de puissance en  $P(k) = k^{-G}$  où  $G=2.45$  (*out*) et  $2.1$  (*in*) - résultat confirmé par le groupe de recherche IBM à Palo Alto. Intuitivement, on peut résumer ceci en disant qu'on trouvera peu de nœuds très connectés, et beaucoup de nœuds peu connectés, et ceci quel que soit l'échelle considérée (le nombre de nœuds). C'est pour cette raison qu'on les appelle *scale-free* ou à invariance d'échelle. Des expériences au niveau des « tuyaux » du web (routeurs, serveurs,...) ont donné des résultats similaires et montrent que la structure physique du réseau repose, elle aussi, sur un petit nombre de nœuds très connectés ou *hubs*.

La recherche en théorie des graphes s'est alors tournée vers d'autres disciplines proposant des réseaux à étudier, qui pourraient donner d'autres modèles expliquant les graphes empiriques venant du Web. Une branche de la sociologie a décrit en particulier les réseaux sociaux (S. Milgram, 1967), et des biologistes modélisent également de cette manière les interactions chimiques entre gènes et protéines. Milgram a décrit ainsi une propriété des réseaux sociaux sous le nom de *small world phenomenon* : les liens sociaux sont très denses en certains endroits, où s'agrègent un groupe de personnes, et il n'existe que très peu de liens connectant ces agrégats entre eux. Ainsi, il existe toujours un très court chemin d'un nœud à un autre<sup>1</sup>. Intuitivement, ceci représente assez bien l'idée qu'on peut se faire du comportement des réseaux en loi de puissance. A partir des échantillons récoltés, et à l'aide d'outils statistiques, Barabasi et Réka<sup>2</sup> ont inféré les propriétés du Web, et vérifié que le chemin moyen entre deux documents passe par  $d$  arcs, ou clics de souris, où  $d = 0.35 + 2.06 \log(N)$ . Soit pour un Web de  $10^9$  documents, une moyenne de 19 [HUB 01]<sup>3</sup>. Ce qui relève bien du *small world*, vu le nombre de documents. Cependant, dans le cas du

<sup>1</sup> Ainsi, entre deux individus pris au hasard, le chemin moyen le plus court est de 6 connaissances directes.

<sup>2</sup> [BAR 99a+01]

<sup>3</sup> IBM, Compaq, AltaVista ont mesuré une distance empirique de 16 sur un échantillon de 200 millions dont ce modèle prédisait 17.

Web, le chemin n'est pas garanti, puisque l'existence d'un lien dans un sens ne garantit pas l'inverse.

La propriété de *small world* permet déjà de déduire un certain nombre de propriétés topologiques. En particulier, on montre par simulation que la disparition de 80% des nœuds pris au hasard d'un réseau de ce type n'engage pas la structuration de ceux qui restent. Ceci explique l'extraordinaire robustesse de ces systèmes. D'un autre côté, une suppression ciblée des principaux *hubs* ruinerait l'organisation du réseau. Dans le cadre des théories de la diffusion (en épidémiologie, marketing), on a également montré que, pour les réseaux à invariance d'échelle, une information circulant sur un tel réseau ne serait jamais perdue. Par exemple, un virus ne sera jamais tout à fait éradiqué : il persistera. Ces corollaires sont vérifiés dans la réalité par les éditeurs d'anti-virus et par les spécialistes du *hacking*.

### *Barabasi et le modèle évolutif*

L'étape suivante vient du questionnement sur la loi de puissance, libre d'échelle, suivie et par le Web, et par l'architecture physique du réseau. Ceci a amené à remarquer, d'abord, que les modèles théoriques existants considéraient des réseaux figés, ce qui n'est pas le cas du Web, ni d'Internet comme structure physique. Ensuite, les réseaux présents dans la nature et nos deux intéressés sont un cadre d'observation du phénomène *d'attachement préférentiel*. Ainsi, un nœud nouvellement venu aura tendance à s'attacher à un nœud déjà existant et bien connecté. La motivation peut venir de l'intérêt de se lier à un site connu pour générer du passage sur sa page (cas du webmestre), ou de connecter toute nouvelle installation à un *hub* garantissant une large bande passante (cas de l'ingénieur réseau).

Un modèle d'évolution est alors proposé [BAR 01] où le réseau est construit en ajoutant à chaque pas de temps un nœud avec une probabilité de s'attacher aux  $i$  autres déjà présents, et qui suit une forme d'attachement préférentiel :  $P_i(k) = k/\sum_i k_i$ . La simulation permet de vérifier que le réseau est bien invariant en échelle, avec une loi de puissance d'exposant  $G=3$ . Ce modèle simule assez bien l'apparition d'une forme de hiérarchie de la connectivité entre nœuds, comme sur le Web. On notera que le phénomène d'attachement préférentiel est typique des systèmes compétitifs (chaîne alimentaire par exemple). Ce modèle évolutif simple permet de prédire l'apparition de nouveaux liens : on sait que  $dk/dt = mP_i(k)$  où  $m$  est le nombre de liens initial du nœud ; on en déduit  $k(t)=t^{1/2}$ , soit que l'apparition de nouveaux liens suit une loi de puissance d'exposant  $1/2$ .

Malheureusement, les réseaux réels sont bien souvent gouvernés, au niveau de la distribution de probabilité des liaisons, par des modèles non-linéaires. Ceci a été montré par J. Mendes et d'autres dans la même veine, prédisant ainsi que de tels réseaux auraient des déviations dans la trajectoire de la loi de puissance, ainsi que des phénomènes de seuil maximum.

Les derniers modèles à invariance d'échelle proposés [BAR 01] tiennent compte du phénomène de compétition pour la connectivité. On introduit à cet effet une compétence  $\eta_i$  pour chaque nœud, qui représente ses chances de récolter des liens, ce qui donne  $P_i(k) = \eta_i \cdot k / \sum_i k_i \cdot \eta_i$  i.e. la connectivité de chaque nœud est donnée par  $k_i(t) \sim t^{\beta(\eta)}$  où  $\beta(\eta)$  est croissante. En somme, des nœuds performants peuvent rejoindre le réseau tard et gagner tout de même plus de liens que d'autres plus anciens. Physiquement, cela rappelle le succès de Google, qui en très peu de temps a pris la place de nœud le plus connecté, bien qu'il soit arrivé bien après AltaVista, Yahoo, eXcite ou d'autres.

Une voie actuellement suivie par Barabasi et al. [BAR 02] est de tester si l'on peut appliquer le modèle de Bose-Einstein pour la condensation des gaz, très étudié en physique nucléaire. Cela revient à remplacer la compétence de chaque nœud par une distribution dépendante du système. Pour ce modèle, on retrouve

le phénomène d'attachement préférentiel pour certaines distributions. Mais l'on ne sait pas encore lesquelles peuvent être utilisées, certaines conduisant à un *winner-take-all* - soit, le meilleur nœud s'arroge la plupart de la connectivité, d'autres ne produisant pas de « gagnant » incontestable.

La théorie des graphes a donc produit de remarquables modèles d'explication du Web, et s'enrichit actuellement chaque année des recherches dans les domaines les plus variées qui attestent toutes de la puissance du modèle (génomique, réseaux de rivières, réseaux sociaux, chaînes alimentaires, physique nucléaire, etc). Comme on l'a vu pour les virus ou les hackers, certains ont déjà compris l'intérêt de connaître la topologie du réseau. Du côté de la recherche scientifique, Kleinberg a suscité une vague de recherches afin d'utiliser ces propriétés pour aider l'humain à appréhender la structure du Web comme espace documentaire.

### **Topologie et sémantique hypertextuelle**

L'articulation entre topologie et sémantique sur le Web est une question qui commence d'être explorée. Cependant, les verrous technologiques (principalement la taille du Web et l'hétérogénéité des formats) et le manque de théorie dans le domaine de la sémiologie des interfaces et de l'interaction, ainsi que de résultats pratiques de la théorie des supports hypermédiés, ont jusqu'à présent restreint le champ des études. En effet, l'on ne sait pas faire grand-chose en analyse sémantique textuelle par ordinateur: la plupart des systèmes robustes sont basés sur des techniques statistiques exploitant le découpage de l'écriture en mots, soit une certaine spatialité du support.

Quant à la sémantique des images, animées ou non, seule la phénoménologie des objets temporels semble tenir la route<sup>1</sup>, mais les résultats n'ont pas permis une approche multi-échelle. D'ailleurs, l'« intelligence » des outils de l'analyse computationnelle en général se limitent en fait *par essence* à une réorganisation syntaxique et ne permettent aucun « calcul du sens », comme promis par certains participants au projet du « Web Sémantique ». Les outils robustes sont donc ceux qui permettent une réorganisation mettant en évidence des régularités fortes susceptibles d'instrumentaliser toute *lecture* du Web.

Les études présentées ici utilisent donc seulement l'analyse textuelle pour induire une certaine régularité thématique de ce que l'on est pour le moment condamné à nommer « documents Web », et dont on se contente pour le moment, faute d'une sémiologie consistante<sup>2</sup>. Ainsi, les documents sont téléchargés et l'on en extrait le texte, perdant ainsi toute spatialité porteuse de sens (sans compter les images, la vidéo, le son) sauf le découpage en mots, qui sont analysés suivant deux grands principes : La récurrence d'un mot dans un document induit son importance dans le discours véhiculé ; la présence dans un document mais pas dans d'autres d'un mot permet de caractériser ledit document. On considère à la limite le texte comme un signal qu'on cherche à caractériser, en éliminant le bruit<sup>3</sup>. Cette approche est caractéristique de la théorie de l'information utilisée en informatique. Des modèles linguistiques plus complexes font leurs preuves, mais particulièrement dans le cadre de systèmes fermés, comme celui des documents techniques par exemple, où la langue varie moins. Le Web ne permet pas pour l'instant ce type d'analyse<sup>4</sup>.

<sup>1</sup> Cf. les normes MPEG 7 et 21

<sup>2</sup> On peut imaginer que des études similaires, mais concernant les routines de navigation sur le Web, à celle de K. Lynch [LYN 60] sur la représentation mentale ou carte mentale que se construisent les usagers d'une cité pourraient nourrir ces réflexions.

<sup>3</sup> D'où l'utilisation de *stop-lists* pour supprimer les mots jugés « insignifiants » à ce niveau d'analyse, tels les conjonctions de coordination par exemple.

<sup>4</sup> Je n'ai pas parlé des problèmes supplémentaires liés au multilinguisme ambiant, par exemple.

Du point de vue topologique, nos connaissances sont basées sur des modèles mathématiques solides, éprouvés, et leur interprétation dans un cadre sémantique ou sémiologique est l'objet du présent chapitre.

Le premier travail articulant les différentes facettes des documents du Web est sûrement celui de Pirroli, Pitkow et Rao [PIR 96]. Les auteurs présentent la notion de *découverte* d'information, prévoyant déjà de l'instrumentaliser par le biais des informations disponibles pour les systèmes de navigation : topologie, méta-données, voie et fréquence d'usage, similarité textuelle. Ces éléments fournissent encore aujourd'hui la base de l'algorithme de classement des résultats de Google (le PageRanking).

### *Kleinberg et la théorie des agrégats*

En 1998, J. Kleinberg [KLE 98] développe un ensemble d'algorithmes pour extraire de l'information à partir de la topologie, avec pour perspective la *broad topic distillation*, soit la « distillation » de groupements thématiques de documents, par le biais de la mise en évidence de sources d'information faisant autorité. A partir de l'observation selon laquelle les nœuds du réseau les plus connectés forment une structure d'*agrégat*, il propose les notions de *hubs* et *authorities* pour désigner respectivement les nœuds pointant vers un grand nombre d'autres, et ceux étant pointés par le plus grand nombre. Ces notions ne se bornent pas à un décompte des liens entrants et sortants, elles incluent la perspective selon laquelle il advient dans ce type de réseau que les hubs et autorités se *renforcent mutuellement* par le biais de leur relation topologique. Ainsi, un hub pointera beaucoup d'autorités, et les autorités seront pointées par beaucoup de hubs. L'hypothèse faite ici est que les faces topologique et sémantique du réseau se recouvrent.

L'idée sous-jacente est de résumer un « thème » représenté sur le Web (parmi les exemples : bouddhisme zen, cryptographie, avortement<sup>1</sup>) par un petit groupe de documents de qualité, ceux-ci contenant à la fois des informations indispensables sur le sujet, et des liens vers d'autres pages. Ceci est réalisé en pratique par l'algorithme dit de repérage par renforcement mutuel, base du système HITS développé par Kleinberg et al. [KLE 99] et qui trouve suite dans le projet IBM ARC, puis CLEVER<sup>2</sup>. Chakrabarti et al. [CHA 98a+99c] travaillent également sur le projet, et ont publié des papiers faisant extension aux travaux de [KLE 98]. On note également une collaboration de certains d'entre eux avec Kleinberg sur [KLE 99b]. Fondamentalement, l'algorithme utilisé ne change pas, mais s'enrichit de plusieurs variantes mettant en jeu des informations « sémantiques ». En voici le principe :

#### *L'algorithme de détection par renforcement mutuel: premier niveau*

On commence par rentrer les mots-clefs de la requête dans un moteur de recherche travaillant uniquement sur le texte, et l'on récupère les  $N$  (usuellement autour de 200) premiers résultats. Ce lot de base des pages potentiellement pertinentes n'inclut pas forcément le cœur du thème, mais il y a des chances que les voisins sur le graphe soient intéressants – nous sommes dans un « petit monde ». On augmente alors le graphe en procédant à un crawl<sup>3</sup> pour trouver les voisins dans un sens, et l'on a recours à une astuce pour l'autre sens : on demande à un moteur de recherche de trouver des pages pointant vers une URL donnée. Ceci peut être fait récursivement jusqu'à différentes profondeurs. Le but

<sup>1</sup> Les thèmes recherchés sont relativement larges, et laissent de côté toute requête *spécifique*, comme par exemple de retrouver des paroles de chanson à partir d'une partie (« we all live in a yellow submarine »).

<sup>2</sup> < [www.almaden.ibm.com/cs/k53/clever.html](http://www.almaden.ibm.com/cs/k53/clever.html) >

<sup>3</sup> On parse le contenu des documents pour trouver les liens, que l'on suit pour télécharger les voisins, et ce de façon récursive, jusqu'à une certaine profondeur donnée au départ.

étant de récupérer un lot de petite taille, relativement centré sur le sujet, on s'arrête en pratique à 2-3 niveaux, la taille du lot augmentant exponentiellement avec la profondeur.

On extrait alors par calcul les hubs et autorités les plus importants. En fait, la difficulté tiendra au filtrage, soit à essayer de démêler les nœuds universellement populaires (google, amazon) des nœuds également bien connectés, mais centrés sur le sujet. Pour cela, il s'agit d'exploiter le fait que ces nœuds seront de gros hubs, mais ne feront pas partie du cœur de hubs et autorités.

Mathématiquement, on code d'abord le graphe augmenté ( $N$  nœuds) en matrice d'adjacence  $M$  et l'on considère deux vecteurs unités  $h$  et  $a$  de taille  $N \times 1$  représentant les scores de hubs et d'autorité des nœuds. Ces vecteurs seront normalisés à chaque itération de sorte que  $\sum_{i=1..N} \{h(i)\}^2 = 1$  et  $\sum_{i=1..N} \{a(i)\}^2 = 1$ . En désignant  $h^j$  et  $a^j$  les vecteurs aux étapes  $j$ , on effectue récursivement les opérations suivantes :  $a^{j+1} = Ma^j$  et  $h^{j+1} = h^j M^T$ , ce qui correspond à affecter à chaque nœud un score d'autorité représentant l'importance du flux de « passage » entre les temps  $0$  et  $j$  vers ce nœud et à partir des  $j$  premiers voisins. De même pour les hubs. Au bout de quelques itérations (5-10 en moyenne), les scores se cristallisent et l'on peut alors repérer les plus importants. En prenant typiquement les 15-20 premiers hubs et autorités, nous avons le cœur du thème par rapport au web entier.

Il est à noter que les résultats d'algèbre linéaire classique nous prouvent que ce processus converge à l'infini vers  $a^*$  et  $h^*$ , vecteurs propres principaux de  $M^T M$  (appelée matrice de co-citation) et  $MM^T$  (matrice de couplage bibliographique), respectivement. L'algorithme est assez efficace, et robuste, puisqu'il ne suppose (presque) rien sur le format des pages, ni sur leur organisation a priori. D'autres travaux dans cette veine tenteront de le combiner à des heuristiques essayant de deviner le « type » des pages, pour supprimer les liens commerciaux par exemple. Ces éléments semblent toutefois trop sujets à changement, puisqu'ils essayent de modéliser un usage humain que le seul fait de décrire changera forcément. L'efficacité de cet algorithme vient sans doute de l'invariance d'échelle, qui pour certaines valeurs du coefficient de la loi de puissance (proches de 2 [KLE 00]) fait que l'on peut trouver des courts chemins et inférer le comportement global à partir d'une connaissance très locale, ici fournie par un moteur de recherche standard.

#### *L'algorithme de détection par renforcement mutuel: second niveau*

Pratiquement, on s'est bien vite rendu compte que cet algorithme retourne des cœurs comportant souvent des variations sur le thème, en particulier pour des sujets polémiques comme l'avortement. Il s'agit alors de reconduire une analyse des liens du cœur pour produire des groupes homogènes ou clusters, dont on voudrait qu'ils soient centrés sur une opinion sur le sujet. Cette direction se base également sur la présomption selon laquelle la structure des liens du Web renferme une information sur le contenu, quel que soit l'échelle considérée.

Des méthodes usuelles de *data-mining* comme l'ACP<sup>1</sup>, et la mise à l'échelle multidimensionnelle<sup>2</sup> permettent de décomposer un graphe en clusters. Ces méthodes utilisent une matrice  $M$  de  $\mathcal{M}_{N,N}(\mathbb{R}_+)$  contenant l'information de similarité entre nœuds, ainsi qu'un vecteur de  $\mathcal{M}_{N,1}(\mathbb{R}_+)$  tiré de  $M$  pour chaque nœud, représentant sa similarité vis-à-vis de chacun des autres. On utilise alors les premiers vecteurs propres non principaux de  $M$  pour définir un espace à peu de dimensions dans lequel on projette les représentations des nœuds. Un panel de visualisations et d'algorithmes permet alors de clusteriser densément dans les espaces de peu de dimensions. L'algèbre linéaire de base nous assure ici que la décomposition sur les  $k$  premiers vecteurs propres de  $M$  produit une distorsion

<sup>1</sup> Analyse en Composantes Principales

<sup>2</sup> multidimensional scaling

minimale sur les projections dans  $k$  dimensions. Ces méthodes ont été testées par Larson et Pirolli et Pitkow en 1996/97.

Des méthodes plus exotiques existent, comme par exemple l'analyse spectrale de graphes (non-orientés) qui nous livre ce résultat intéressant : Chaque vecteur propre de la matrice d'adjacence peut se voir comme un vecteur de poids pour chaque nœud, et l'on remarque que pour un vecteur propre donné, les nœuds ayant des scores négatifs sont très peu connectés au nœuds ayant des scores positifs. Une méthode similaire est celle de la mise à l'échelle des centroïdes, également basée sur l'analyse en vecteurs propres, permet de clusteriser par des méthodes géométriques.

Vient alors la partie peut-être la plus surprenante, où Kleinberg [KLE 98] montre sur plusieurs exemples une méthode de clustering de la structure des hyperliens ayant des implications pour le regroupement en sous-thèmes du cœur du domaine. L'algèbre linéaire nous permet de dire que  $M^T M$  et  $MM^T$  ont les mêmes valeurs propres, et leurs vecteurs propres  $w_i$  peuvent être choisis de sorte que  $Mw_i(M^T M) = w_i(MM^T)$ . De cette façon, chaque paire de vecteurs propres  $[a_i^* = w_i(M^T M); h_i^* = w_i(MM^T)]$  à la propriété suivante : pré-multiplier par  $M$  (resp. post-multiplier par  $M^T$ ) laisse les composantes en  $a$  (resp.  $h$ ) parallèles à  $a_i^*$  (resp.  $h_i^*$ ). En d'autres termes, apporter la contribution des voisins pointants ou des voisins pointés donne cette relation de renforcement mutuel, concrétisée au niveau du modèle mathématique par la propriété ci-dessus. De plus, en appliquant  $M^T M$  (resp.  $MM^T$ ), on multiplie les composantes en  $a$  (resp.  $h$ ) par  $|\lambda_i|$ , ce qui nous donne précisément la mesure de ce renforcement entre hubs et autorités. Mais, contrairement au vecteur propre principal, les autres ont des composantes négatives et positives, ce qui nous permet de distinguer au sein de chaque paire  $[a_i^*; h_i^*]$  les composantes les plus positives et négatives i.e. des lots de hubs et d'autorités densément liés. Pratiquement, on prend les  $c$  valeurs les plus négatives et positives de  $a_i$  et de  $h_i$  pour avoir deux lots.

Plus la valeur absolue du vecteur propre considéré est grande, plus les lots sous-jacents seront connexes. Ce classement en fonction des vecteurs propres  $a$ , dans un certain nombre de cas, un sens au niveau de l'orientation sémantique des lots par rapport au sujet de départ. En effet, on remarque que la décomposition selon les vecteurs ayant les plus grandes valeurs absolues tend à générer les mêmes lots – ce qui veut dire que l'on obtient souvent beaucoup moins de lots que de vecteurs.

Par cette méthode, Kleinberg a obtenu les deux composantes pro- et anti- du domaine de l'avortement, par exemple. Cela vient intuitivement du fait que les pro- et les anti- forment des communautés plutôt fermées et donc ne sont que peu connexes entre elles, alors que les membres de la communauté forment un sous-graphe fortement connexe.

#### *Heuristiques annexes et conséquences théoriques pour l'interface*

Dans [KLE 98], se dégage déjà la volonté de restitution à l'utilisateur de la structure hyperliée des documents. On y voit d'ailleurs l'amorce d'un problème théorique : la définition d'un « site Web ». Dans le langage courant, on suppose que c'est l'unité signifiante du point de vue de l'utilisateur, qui se représente bien souvent un *agrégat* de documents à partir d'un semblant d'autorité ou plutôt d'*authorship*<sup>1</sup>, c'est-à-dire d'homogénéité dans le discours et dans la forme qui se dégage au fil des « pages ». Ici, Kleinberg classe les hyperliens en *intrinsèques* (ceux reliant des documents dont le nom de domaine est le même, par exemple « utc.fr ») et *transverses* (ceux reliant les documents de deux domaines différents), supprimant les premiers du graphe, justifiant ce choix par la présomption qu'ils ne servent qu'à naviguer à l'intérieur du « site ». Il existe bien sûr des « sites », soit un lot de pages reliées thématiquement et émanant d'une

<sup>1</sup> Le terme est mien, mais n'arrivant pas à le traduire, je le conserve tel quel.

même autorité, distribuées sur plusieurs serveurs, et dans ce cas, on représentera un nœud par nom de domaine sur le graphe<sup>1</sup>.

En définitive, cette heuristique est une manière de se passer de définition technique explicite d'une unité documentaire signifiante. En effet, un document (au sens informatique de fichier) particulier, dont l'URL serait du type « <http://www.abc.org/de.xml> » sera un nœud du graphe, même si ses liens ne seront comptés que s'ils sortent du domaine « abc.org ».

On peut d'ailleurs légitimement se demander si, du point de vue technique<sup>2</sup>, le Web ne formerait pas qu'un seul *hyperdocument*, distribué sur la structure physique d'Internet. Si l'on considère les choses de cette manière, on ne peut alors que se questionner sur les représentations mentales utilisées par les internautes – ce que des études d'usages couplées à une volonté de théorisation pourrait peut-être éclaircir. Le manque de concepts pour penser ce qu'est un hyperdocument nous interroge ici particulièrement, puisque nous sommes dans une perspective d'instrumentation. En effet, nous ne pouvons rendre à l'utilisateur une représentation qu'en dégagant des unités signifiantes à représenter, afin de pouvoir instancier une sémiologie graphique adaptée. Ceci se fait en faisant correspondre par des relations les unités signifiantes de la pensée à celles de la représentation. Nous élargirons ces éléments dans la partie visualisation ci-après.

De toute façon, le fichier reste pour le moment l'unité ou *grain* le plus immédiat et abordable, plus du point de vue technique, que de la structuration du discours dans l'hyperdocument, puisque son existence est une information de granularité qui va varier considérablement d'un emballer de contenu à l'autre – et qui n'en est, bien souvent, pas l'« auteur »... On peut alors penser aux hyperliens comme une unité plus petite, puisqu'il y en a qui sont internes au document, et qui peuvent servir à relier un paragraphe à l'autre<sup>3</sup>. Mais de la même manière, certains fichiers seront plus ou moins structurés : les publications professionnelles le seront, alors que la plupart des pages personnelles ne le seront pas, bien qu'elles renferment souvent un contenu non négligeable.

Une autre heuristique annexe de Kleinberg, non testée, consiste en la mise en place d'un seuil maximum (usuellement 4-8) sur le nombre de liens entre pages d'un même domaine. Ici aussi, l'on est confronté à une volonté de deviner le sens des pratiques des créateurs de documents, dans le but de ne retenir que des liens « conférant autorité ».

D'une manière générale, il existe beaucoup de variantes ou de généralisations du renforcement mutuel. Parmi les plus intéressantes, on peut citer les travaux de Chakrabarti et al. [CHA 98a] qui utilise le texte du voisinage de l'hyperlien<sup>4</sup> dans le fichier de départ pour pondérer les liens, selon la proportion de mots du voisinage se retrouvant dans la requête. Cette technique de pondération est appelée *anchor window weighting* et donne des résultats dépendant fortement du choix de l'étendue du voisinage<sup>5</sup>. [CHA 99c] teste un panel d'innovations sans dévoiler leur implémentation, parmi lesquelles la diminution de score des pages similaires, l'assimilation des pages d'un domaine (hostname) à un seul point d'entrée, et une identification des zones dans les pages où l'on a une information pertinente. Les résultats sont meilleurs que dans [KLE 98], et battent Yahoo ! Categories (ontologies humaines) et AltaVista sur une évaluation par des

<sup>1</sup> La définition du site donnée ici est celle de l'homme de la rue, et bien qu'intuitive, ne se prête guère au repérage automatique, pour la simple raison que les méta-données de la plupart des documents, incluant l'auteur, ne sont soit pas mises à jour, soit mal ou pas du tout renseignées.

<sup>2</sup> i.e. de la structure des hyperliens

<sup>3</sup> C'est le système des ancres nommées : les URL de type « abc.htm#par1 » renvoient à une balise <A NAME=par1></A> qui est cachée à l'utilisateur en HTML 4 . Ceci existe également pour une variété de formats : en Flash, dans les scripts java, perl, cgi, php,...

<sup>4</sup> i.e. le texte avant et après la balise HTTP <a href=URL> texte </a>

<sup>5</sup> Ceci est développé dans [BRI 99]

utilisateurs humains. Cependant, le banc d'essai ne paraît pas garantir une fiabilité intrinsèque au système, beaucoup de paramètres devant sans doute être réglés suivant l'évolution des usages du Web. Par exemple, l'apparition d'un algorithme de type anchor window dans Google a changé les habitudes des webmestres avertis, qui sont en compétition pour la connectivité et pour le référencement.

D'autres techniques intéressantes ont été publiées par Bharat et Henzger [BHA 98], ce qui les amena directement à une embauche en tant que Principal Scientist et Research Director de Google Inc. Il s'agit notamment de l'expansion de requête: on récupère les premiers résultats d'un moteur de recherche par term-matching basique, comme à la première étape de l'algorithme de Kleinberg, et l'on télécharge les  $k$  premières pages, avant de récupérer les  $N$  premiers mots de chaque. On calcule alors une fréquence pour chaque mot dans ce lot, et l'on extrait les  $L$  premiers (usuellement :  $k=30$ ,  $N=1000$ ,  $L=100$ ). Ainsi, on espère trouver un vocabulaire intéressant, composé de termes utilisés fréquemment pour parler du sujet de la requête. On l'utilise alors, soit pour pondérer les arcs, soit pour filtrer les nœuds jugés non pertinents (à l'aide d'une fréquence seuil), soit encore les deux à la fois.

Ce papier pose un nombre imposant d'autres variantes, introduisant à tous les niveaux de l'algorithme une pondération « sémantique » ou topologique. Ainsi, le poids des liens d'un « domaine » (hostname) sont divisés pour que la somme fasse 1 avant le calcul des hubs et autorités. Pendant le calcul ou après, on seuille par différentes valeurs (médiane dans le corpus de base, dans le corpus augmenté, dans la requête expansée,...) pour rejeter les nœuds impertinents, afin notamment d'accélérer la stabilisation des scores. Ceci est compréhensible, puisque le nombre de nœuds considéré dans le lot augmenté est plus grand que celui de [KLE 98].

Une question très importante posée par ces variantes est de savoir comment synthétiser toutes ces versions qui sont en fait des paramétrisations de l'algorithme de base. Il serait merveilleux de pouvoir les fondre en un corps de règles robuste, i.e. dépendant peu du jeu de données. Pour cela, seule une campagne de tests grandeur nature peut permettre de séparer les règles robustes des particulières. On peut également imaginer que si l'on disposait d'une archive (immanquablement réduite et dépassée) d'un thème sur le Web, cela permettrait de passer plusieurs tests en conservant le même jeu de données. Après, des tests entre jeux de données différents permettraient de découvrir les paramètres dépendants des usages. Ceci a été très vite perçu par les membres du projet IBM, qui ont dès le départ programmé un *serveur de connectivité* leur permettant d'avoir à disposition un sous-graphe du web pour les essais, ce qui fait gagner un temps considérable en évitant le téléchargement à chaque étape.

#### *SALSA et PageRank, deux inspirés*

Dans [LEM 01], deux chercheurs de l'université de Haïfa proposent un algorithme (SALSA) à base d'outils stochastiques comme les chaînes de Markov, qui modélise un déplacement aléatoire sur le Web, et dont ils prouvent qu'il suit la même métaheuristique que celui de Kleinberg. Les résultats sont, là encore meilleurs que ceux de Kleinberg, et ce sans une paramétrisation extensive.

Finissons cette revue par l'algorithme de Page, assise du succès de Google. Celui-ci permet de classer les résultats du premier niveau de recherche en texte intégral. Comme décrit brièvement dans [BRI 98], les pages les plus pointées auront un rang élevé, et seront capables de conférer de l'autorité aux pages qu'elles pointent. On a une propriété de transitivité. Ensuite, la taille des caractères est évaluée pour pondérer l'importance des mots vis-à-vis de la

requête. Plusieurs autres métadonnées sont utilisées, comme le titre de la page et les commentaires de l'auteur dans la balise HTML 4 <meta content> et <met description> par exemple. La pondération est également affectée par un système de type anchor window. La dernière version de Google intégrerait également un profilage utilisateur, afin de lui rendre l'expérience capitalisée par d'autres utilisateurs ayant envoyé des requêtes similaires, changeant ainsi le classement en faveur des pages visitées par les prédécesseurs. Depuis un an, les sites payant une redevance à Google sont également favorisés<sup>1</sup>.

En définitive, le PageRank est une heuristique exploitant les idées sous-jacentes de [KLE 98], mais dans une perspective qui ne produit pas une instrumentation du parcours, puisqu'en effet l'utilisateur se retrouve bien souvent avec un problème d'abondance: il ne regarde que les 10 premiers sites proposés sur les quelques millions jugés pertinents par le système.

### *Conclusion*

De nombreux chercheurs ont contribué à l'innovation dans le domaine de l'analyse topologique du Web, et je n'ai pu rendre leur apport qu'à un petit nombre. Les références bibliographiques proposées dans la suite de ce travail permettent de (re)trouver sans peine les contributeurs manquants, vu que le domaine est encore assez ténu et connexe.

Nous avons vu ici l'importance de la structure hyperliée du Web, notamment parce qu'elle permet de proposer un point de départ à l'exploration par l'utilisateur d'un domaine, mais aussi en ce que sa connaissance est nécessaire pour bâtir une sémiologie des documents numériques et de l'interaction.

Nous allons maintenant nous pencher sur les représentations de cette structure, à travers la revue d'un domaine récent, qui est celui de la visualisation d'information ou InfoVis. Celui-ci s'est développé et nourri d'un aller-retour entre les questionnements posés par la topologie des réseaux et leur exploration.

---

<sup>1</sup> Pour les dernières nouvelles, se reporter à <<http://www.searchenginewatch.com/>>.

## 2. INTERFACE ET VISUALISATION DE L'INFORMATION

*Cette section est incomplète. Les éléments présentés ici ne sont que l'amorce d'une réflexion. La bibliographie permettra de compléter le tableau.*

Ce champ de connaissances a pour objet la création d'outils informatiques qui exploitent le système visuel humain afin d'aider l'exploration ou la compréhension de données. L'interaction avec une représentation visuelle adéquatement conçue nous permet de former des modèles mentaux nouveaux, qui nous permettent d'effectuer certaines tâches jusqu'alors malaisées, voire impossibles.

Les traditions plus anciennes comme le champ des IHM, de la psychologie cognitive, de la sémiotique, du design graphique, de la cartographie et de l'art ont toutes proposées des représentations spatiales pour des jeux de données abstraits<sup>1</sup>. La synthèse de ces réalisations reste encore à compléter pour en tirer de nouvelles méthodologies et techniques, en accord avec les possibilités inconnues jusqu'alors de l'interaction avec un support numérique. Le but des anciennes comme des nouvelles étant de produire de la connaissance par le biais de l'apparition de schémas de pensée nouveaux, modelés par les propriétés cognitives inhérentes à l'usage des outils créés.

### **GraphVis**

Ce domaine traite de la représentation des graphes, et a dès ses débuts été une partie importante de l'InfoVis. Les algorithmes produits ont eu pour but pendant longtemps de minimiser le nombre d'intersections et la longueur des arcs des représentations.

On peut citer le travail de C. Ware qui a remis en perspective [WAR 00] les enjeux de la discipline, notamment en posant la question du coût cognitif des choix de briques signifiantes de la représentation. En ce sens, il se situe dans le prolongement de la sémiologie graphique de J. Bertin [BER 67]. Plusieurs autres auteurs ont depuis suivi cette voie visant à retrouver des règles de présentation définissant les meilleures manières de choisir les variables graphiques de la représentation.

### **Mapping**

La réflexion de Bertin sur la sémiologie des représentations graphiques papier présente une approche reconduite dans de nombreux domaines. En effet, il est curieux de remarquer que les méthodes utilisées par l'atelier de l'IRCAM, par exemple, pour créer de nouveaux instruments revient à *mapper* des variables gestuelles classées suivant leurs caractéristiques cognitives, à des variables informatiques, classées suivant les effets sonores produits. Il en est de même pour les interactions avec une interface graphique : il s'agit d'instancier une sémiologie de l'interaction. Soit faire correspondre, à l'aide de combinaisons, pondérations, en général linéaires, des variables de base pour exprimer les variables du vis-à-vis, qui répètera l'opération dans le sens inverse afin de

<sup>1</sup> L'InfoVis se préoccupe particulièrement de la représentation de données qui n'ont pas de représentation spatiale implicite, par opposition à la visualisation scientifique.

boucler la boucle sensori-motrice. C'est dans l'invariant sensori-moteur (même action, même sensation) que va se créer la représentation mentale nouvelle<sup>1</sup>.

### ***Théorie du support: conséquence pour le design***

Comme dit B. Stiegler dans le tome 2 de *La technique et le Temps*, repris par B. Bachimont au sujet des hypertextes, « comprendre un texte, [...] c'est être capable d'en écrire sa lecture ». En ce qui concerne un réseau hypermédia, si l'on veut instrumenter sa lecture, il semble naturel de vouloir inscrire son parcours. Des outils en ce sens ont été développés, comme le bien connu Nestor<sup>2</sup> qui permet de naviguer en ayant sous les yeux le graphe du parcours, ainsi que des outils d'annotation graphiques.

### ***Réalisations en InfoVis***

Un excellent site maintenu par M. Dodge<sup>3</sup>, qui a écrit également des papiers et livres sur la « cartographie du cyberspace », recense les réalisations les plus importantes.

On y trouve notamment des réalisations de Tamara Munzner comme son algorithme H3 qui *mappe* un réseau de 100 000 nœuds sur une sphère, utilisant pour cela une géométrie hyperbolique, qui permet de constituer une représentation de type fisheye<sup>4</sup>. L'astuce consiste à projeter l'espace hyperbolique sur un cercle, ce qui a pour effet de donner une mesure de distance exponentielle à partir du centre. De cette façon, on peut choisir un nœud-focus et avoir une vision du voisinage qui décroît en taille avec la distance vis-à-vis du focus. L'échelle exponentielle permet de représenter des réseaux très importants ou plutôt d'en avoir une vision d'ensemble.

[MUN 00b] présente une étude de Microsoft sur l'outil XML3D, browser intégrant une représentation 3D du graphe visualisé, et des outils contextualisants (nœud suivants et précédents) et linguistiques (requêtes), destiné aux webmestre professionnels.

On trouve aussi les réalisations de CAIDA<sup>5</sup>, consortium d'analyse des tuyaux d'Internet et notamment leur outil de visualisation de graphes en JAVA : Otter.

### ***Conclusion***

Si pour l'utilisateur l'interface EST la donnée même, au travers de ses invariants de manipulation, alors la structure hyperliée du document (le Web) doit lui être rendue par cette interface pour instrumenter son activité cognitive. Il pourra ainsi produire de nouveaux concepts en inscrivant sa connaissance.

<sup>1</sup> Cette partie me vient de la fréquentation de C. Lenay

<sup>2</sup> <<http://www.gate.cnrs.fr/~zeiliger/nestor.htm>>

<sup>3</sup> <<http://www.cybergeography.com>>

<sup>4</sup> ou focus+context : on peut interagir avec un focus en suivant l'évolution du contexte.

<sup>5</sup> <<http://www.caida.org>>



**BIBLIOGRAPHIE COMMENTEE :  
TOPOLOGIE ET USAGES DU WEB, INTERFACES, VISUALISATION ET  
SPATIALISATION DE L'INFORMATION**

**TOPOLOGIE DES RESEAUX HYPERMEDIAS OUVERTS**

*Web, connectivité, hyperliens*

- [ADA 99a] ADAMIC L.-A., HUBERMAN B.-A., "**Technical comment to 'Emergence of Scaling In Random Networks'**", *Xerox Parc*, 1999.  
Polémique dépassée dans [BAR 02]
- [ADA 99b] ADAMIC L.-A., HUBERMAN B.-A., "**The Nature of Markets in the World Wide Web**", *Xerox Palo Alto Research Center, CA*, 1999.
- [ADA 99c] ADAMIC L.-A., HUBERMAN B.-A., "**Growth Dynamics of the World Wide Web**", *Nature*, vol.401, p.131, September 1999.
- [BAR 02] BARABASI A.-L., REKA A., "**Statistical mechanics of complex networks**" in *Reviews of Modern Physics*, vol.74, 2002.  
Version longue et comportant les preuves mathématiques de [BAR 01]
- [BAR 01] BARABASI A.-L., "**The Physics of the Web**", 2001.  
Histoire des modèles de graphe et conséquences
- [BAR 00] BARABASI A.-L., ALBERT R., JEONG H., "**Error and Attack Tolerance of Complex Networks**", *Nature*, 406, pp.378-382, 2000.  
Application de la théorie des graphes
- [BAR 99a] BARBARASI A.-L., ALBERT R., "**Emergence of scaling in random networks**", *Science*, vol.286, 1999.  
Histoire du passage du modèle aléatoire à l'invariance d'échelle.
- [BAR 99b] BARABASI A.-L., ALBERT R., JEONG H., "**Diameter of the World Wide Web**", *Nature*, vol.401, p.130, 1999.  
Vérification de la propriété de small-world
- [BER 00] BERGMAN M., **the deep-Web : surfacing hidden value**, 2001 (Voir l'étude de la société *BrightPlanet* réalisée en juillet 2000 ([www.lexibot.com/press](http://www.lexibot.com/press)) et l'article de *Libération* « 500 milliards de pages oubliées dans les abysses du web » du mercredi 13 décembre 2000.)  
Présente la notion de *deep-web* i.e. les sites dynamiques (ASP, PHP,...) – dont le contenu est stocké en BDD – et son importance en terme de taille (les 60 + gros *deep sites* = 40x le web de surface). Cite S. LAWRENCE pour [LAW 99+98] qui montrent l'incapacité des moteurs à base de crawlers (trad), qui indexent 15% (en baisse) du web, et sont en décalage de 3-4 mois au moins sur le rafraîchissement des pages. De plus, ils n'indexent pas les pages ayant peu de liens.
- [BHA 98] Krishna Bharat, Monika Henzinger "**Improved algorithms for topic distillation in hyperlinked environments**" in *Proc. 21<sup>st</sup> Int'l ACM SIGIR Conference*, 1998  
*! Fait suite à [CHA 98a] !*  
Explique comment on trouve les in-links vers le base-set dans [KLE 98] : en cherchant sur les moteurs de recherche, les pages répondant à la requête « link :http://www.xyz.com/www/yyy/zzz ». On récupère ainsi plus de 2000 nœuds pointant. Mais cela pose le problème du download des pages en question pour éventuellement les intégrer à l'augmented-set.  
Propose des algos complémentaires pour améliorer l'algo de Kleinberg par rapport à 3 facteurs : 1) certains *host* se renforcent mutuellement : un paquet de docs a beaucoup de liens avec un seul doc sur autre host ; mais comme on suppose que les docs d'un host sont rédigés par la même personne, cela

donnerait trop de poids (on est dans un contexte où chaque host représente l'ensemble de ses sous-pages ?) 2) certains liens sont automatiquement générés : l'hypothèse que les liens sont générés par la citation humaine est donc violée dans ce cas 3) Il existe des nœuds bien connectés, mais non centrés sur le thème.

- analyse de la connectivité : diviser le poids d'un hôte sur ses différents liens ; défausser les nœuds non-connexes
- comparer les documents vectorisés (en entier) à une requête étendue construite à partir des 30 meilleurs éléments du base-set (sur connectivité, adresse url – pour ne pas loader les pages) dont on extrait ensuite 1000 mots qu'on compare à la requête par similarité. Le 25<sup>ème</sup> score de similarité est ensuite pris comme seuil minimum pour décider si les pages suivantes rencontrées lors de la construction du graphe étendu seront considérées comme centrées sur le thème ou non.
- pondérations à base de seuils (médians, médians dans l'augmented set,...) pendant ou après la construction du graphe étendu.
- Elagage du graphe étendu avant ou pendant (itératif) calcul des hubs+autorités sur des bases topo pour accélérer la convergence.

Résultats apparemment meilleurs que [KLE 98] sur une notation par des humains sur les listes délivrées par le système uniquement.

- [BIA 01] G. Bianconi, A-L Barabasi, "**Bose-Einstein Condensation in Complex Networks**" *Phys. Rev. Lett.* 86, 5632, 2001  
Dernier modèle en date pour les réseaux
- [BOT 91] BOTAFOGO R.-A., SHNEIDERMAN B., "**Identify Aggregates in Hypertext Structures**" in the *Third ACM Conference on Hypertext*, San Antonio, ACM Press, pp.63-74, 1991.
- [BRA 96] BRAY T., "**Measuring the Web**", *Proceedings of the Fifth International World Wide Web Conference*, may 6-10, Paris, 1996. ! *Outdated* !  
Pose le problème de la définition d'un site web. Propose la solution sous la forme xxx.{mil, com}.{fr,..}. Etudie brièvement les stats de connectivité, taille, format.
- [BRE 00] BREWINGTON B. E., CYBENKO G., "**How dynamic is the Web ?**", in *www9*.
- [BRO 00] Andrei Broder<sup>1</sup>, Ravi Kumar<sup>2</sup>, Farzin Maghoul<sup>1</sup>, Prabhakar Raghavan<sup>2</sup>, Sridhar Rajagopalan<sup>2</sup>, Raymie Stata<sup>3</sup>, Andrew Tomkins<sup>2</sup>, Janet Wiener<sup>3</sup> 1: AltaVista Company, San Mateo, CA. 2: IBM Almaden Research Center, San Jose, CA. 3: Compaq Systems Research Center, Palo Alto, CA. "**Graph Structure in the Web**" in *www9* < <http://www.almaden.ibm.com/cs/k53/www9.final/> >
- [CAR 99] CARRIERE J., KAZMAN R., "**Webquery : Searching and Visualizing the Web Through Connectivity**", Univ. Waterloo, Canada, 1999.
- [CHA 00] CHAKRABARTI S., GIBSON D.-A., MCKURLEY K.-S., « **Surfing the Web Backwards** » in *www8*  
Introduit les *metadata* dans les composantes du surf. Notamment l'ancien (?) header http qui donne l'url-from. Remarque également que HTML 4.0 introduit la notion de lien nommé, avec TITLE, CONTENT etc. Pose également le problème des ressources (la majorité) abandonnées par leur auteur. Cite le RDF (Ressource Description Framework) et WebDAV, ainsi que la future norme http qui pourrait délivrer ces infos.  
Pose des specs pour un client webbrowser similaires à NESTOR – introduit la notion de contexte (d'où l'on vient pour aller là). Au-delà, teste brièvement l'impact du supplément d'info que sont les backlinks. Pose le problème du spamming de backlinks.
- [CHA 99a] S. Chakrabarti, B. Dom, D. Gibson, J. Kleinberg, R. Kumar, P. Raghavan, S. Rajagopalan, and A. Tomkins. "**Hypersearching the web**" in *Scientific American*, June 1999.
- [CHA 99b] S. Chakrabarti, B. Dom, D. Gibson, J. Kleinberg, S.R. Kumar, P. Raghavan, S. Rajagopalan and A. Tomkins, "**Mining the link structure of the World Wide Web**", *IEEE Computer*, 1999  
Sur l'importance de la structure hyperliée
- [CHA 99c] CHAKRABARTI S., DOM B., GIBSON D., KUMAR R., RAGHAVAN P., RAJAGOPALAN S., TOMKINS A., "**Experiments in Topic Distillation**", *IBM Almaden Research*

Center, CA, 1999.

! Fait suite à [KLE 98] !

Rappelle la notion de *topic distillation* i.e. étant donné un *broad topic* ou thème présent sur le web, on s'occupe de trouver un petit nombre de pages de haute qualité qui représentent bien le thème.

Contexte : le système CLEVER développé chez IBM qui fait suite à HITS, système de Kleinberg et al. Parmi les innovations : 1) Prendre en compte les pages qui sont sur un même site logique (def ?) comme des venant d'une même source et donc peu authority conferring entre elles. 2) Diminuer les scores des pages similaires à un hub déjà existant (critères ?). 3) Retourner seulement un seul point d'entrée vers une ressource (cf 1 ?). 4) Identifier des zones (homogènes ? de proximité ?) dans les pages et les analyser pour savoir lesquelles pointent vers des pages intéressantes.

Suit une étude d'usage entre CLEVER, Yahoo! categories (fait par des taxonomistes humains) et Altavista => CLEVER 1er.

Perspective : essayer de produire/mixer les résultats vers une structure d'arbre taxonomique.

- [CHA 98a] CHAKRABARTI S., DOM B., GIBSON D., KLEINBERG J., RAGHAVAN P. and RAJAGOPALAN S., "**Automatic Resource Compilation by analysing Hyperlink Structure and Associated Text**" in *Proceedings of the 1998 World Conference on the WWW, Internet, and Intranet*.

! Fait suite à [KLE 98] !

Système ARC Automatic Resource System basé sur une analyse locale des liens et du texte.

Algorithme type Kleinberg modifié : phase d'expansion du root-set reconduite i.e. on augment par ajout des voisins en profondeur  $P^{+2}$  ce qui conduit à un augmented-set de 200-3000 nœuds. Calcul du cœur modifié par la pondération des liens, basée sur le texte avoisinant. Après tests, l'algo est calé sur une valeur du voisinage de 50 octets de texte dans les deux sens, dans lequel on cherche le nombre d'occurrences des mots-clefs de la requête. Le poids modifié est de  $1+nb\_occur$  au lieu de 1. On obtient donc deux vecteurs avec des scores de Hub h et d'Autority a pour chaque page. On code une matrice d'adjacence M et l'on applique 5 fois les opérations :  $a \leftarrow Mh$  puis  $h \leftarrow M^T a$ . Cela suffit à garantir une stabilisation des scores h et a. On extrait alors les 15 plus importants.

Relate ensuite une étude utilisateur comparant ARC à Yahoo! et à InfoSeek ; critères retenus pour les descriptions humaines des listes proposées: focus sur le thème, étendue de la couverture (a-t-on des aspects manquants), potentiel de guidage vers d'autres pages bien centrées sur le thème. Commentaires utilisateurs : les hubs sont de meilleurs points de départ ; les listes sont utilisées comme des points de départ ; manque un contrôle de la *granularité du focus* sur le sujet ; le contexte donné par l'arborescence yahoo! est utilisé comme indice visuel ; un résumé de chaque lien proposé est très apprécié, même s'il est constitué des deux premières lignes de la page. Les hubs sont souvent moins choisis à cause de leur identité moins marquée.

- [CHA 98b] S. Chakrabarti, B. E. Dom, and P. Indyk. "**Enhanced hypertext categorization using hyperlinks**" in *Proceedings of ACM SIGMOD*, Seattle, WA, 1998

Propose le Document Object Model pour améliorer HITS en prenant en compte le texte et la structuration du doc hypertexte – à analyser.

- [CHE 01] CHEN T.-M., "**Increasing the Observability of Internet Behaviour**" in *Communications of the ACM*, January 2001, vol.44, n°1, pp.93-98.

- [ERD 60] Erdős, Rényi "**Publ. Math. Inst. Hung. Acad. Sci.**" 5,17 – 1960  
Première théorie des graphes – modèle aléatoire

- [FAL 99] FALOUTSOS M., FALOUTSOS P and FALOUTSOS C., "**On Power-law Relationships of the Internet Topology**" in *ACM SIGCOMM* 1999  
Découverte de la loi de puissance

- [GAR 99] Garofalakis M.N., Rastogi R., Seshadri S., & Shim K. (1999). "**Data mining and the Web: past, present and future**" in *Proceedings of the 2<sup>nd</sup> international workshop on Web information and data management*, 1999. 43-47

- [GIB 98] David Gibson, Jon Kleinberg, Prabhakar Raghavan. **"Inferring Web communities from link topology"** in *Proc. 9<sup>th</sup> ACM Conference on Hypertext and Hypermedia*, 1998  
 ! Fait suite à [KLE 98] !  
 Idée maîtresse : proposer des *aggrégats* à explorer plutôt que des « documents » seuls. Se placer dans une perspective de *découverte d'information* plutôt que de récupération.  
 On peut tirer des idées pour le design d'outils d'exploration de la notion de *hubs et autorités* se renforçant mutuellement. Ceux-ci forment le cœur d'une communauté. Grâce à l'algo HITS, on pourrait peut-être prouver le *silence* d'un domaine ou thème sur le web i.e le fait qu'il n'existe pas de communauté autour du thème.  
 Approche méthodologique à ruminer : le choix du lot de départ de l'algo, la vitesse de sédimentation des cœurs suivant le thème, non-convergence de l'algo en cas de multilinguisme.  
 On trouve le même cœur en prenant comme requête un thème ou des éléments de ce thème (individus, sous-disciplines,...) : c'est la *convergence de la généralisation*.  
 Le cœur à long terme est obtenu en suivant l'évolution sur plusieurs mois des cœurs.
- [HEN 00] HENZINGER M. R., MITZENMACHER M., HEYDON A., NAJORK M., **"On Near-Uniform URL Sampling"** (www9).
- [HAW 99] HAWKING D., CRASWELL N., HARMAN D., **"Results and Challenges in Web Search Evaluation"**, in *the Text Retrieval Conference*, 1999.
- [HEN 99] HENZINGER M. R., MITZENMACHER M., HEYDON A., NAJORK M., **"Measuring Index Quality Using Random Walks on the Web"** in *www8*
- [HUB 01] Bernardo A. HUBERMAN **"The Laws of the Web"**, *The MIT Press*, 2001  
 Résumé des régularités fortes rencontrées sur le Web, notamment dans les domaines de la topologie, des usages, du marketing.
- [HUB 97] HUBERMAN B., PIROLI P., PITKOW J., LUKOSE R., **"Strong Regularities in World Wide Web Surfacing"**, *Science*, 280, pp.95-97, 1998.  
 Intro à [HUB 01]
- [KLE 99a] Jon KLEINBERG **"The small-world phenomenon: An algorithmic perspective"** in *Cornell Computer Science Technical Report 99-1776*, 1999  
 Etudie les influences sur les algorithmes de recherche et de parcours des graphes de la propriété de small world – indispensable.
- [KLE 99b] Jon M. Kleinberg, Ravi Kumar, Prabhakar Raghavan, Sridhar Rajagopalan and Andrew S. Tomkins. **"The Web as a graph: measurements, models and methods"** in *Proceedings of the 5<sup>th</sup> International Computing and combinatorics Conference*, 1999
- [KLE 98] John Kleinberg **"Authoritative sources in a hyperlinked environment"** in *9<sup>th</sup> ACM-SIAM Symposium on Discrete Algorithms*, 1998. *Extended version in Journal of the ACM* 46(1999).  
 Présente des algos pour extraire de l'info des structures hyperliées, notamment pour *distiller des thèmes étendus (broad topics)* ou domaine de connaissance. Propose de chercher des pages faisant office d'autorité dans un domaine (broad topic) et qui représenteront les autres. Pour cela on utilise les hyperliens comme une forme de jugement latent de la part des humains qui les ont créés.  
 Distingue les liens navigationnels, commerciaux et les « citations ». Propose le modèle des aggrégats de hubs et autorités qui se renforcent en se citant mutuellement et représentent le *centre* d'une communauté sur le web. Soulève le problème du juste milieu entre popularité et intérêt des pages à rapatrier. Elimine par conséquent le classement par liens entrants seulement. S'éloigne également de la méthode du *clustering*, bien que l'algo proposé permette (?) de différencier différents sens et opinions sur les termes d'une requête.  
 Suit l'algo de calcul des hubs+authorities à partir d'une requête sur un moteur de recherche std (sans topo), sélection des 200 premiers résultats, augmentation du graphe par les voisins sur le web, puis codage en matrice d'adjacence du graphe, preuve de convergence des scores de hubs&auth par décomposition en vecteurs propres. Lien avec le *small-world phenomenon* en assumant que l'on fait un très gros trou dans le web, en particulier parmi les petits sites peu connectés, mais

que ca n'affecte que peu sa structure.

Elargissement sur les *similar pages queries* et les travaux sur les réseaux de liens sociaux.

Etat de l'art sur le clustering : def d'une fonction de similarité+methode de prod de clusters. ACP, multidimensional scaling, dim reduction algos. Plus particulièrement pour les liens : partition spectrale des graphes, centroid scaling +refs biblio.

Point sur la distillation de domaines : En sélectionnant les pages suivant leurs décompositions selon le vecteur propre principal, on obtient plusieurs « sens ». En suivant les autres premiers vecteurs propres non principaux, on obtient pour chaque vecteur des « sens » ayant trait à des communautés spécifiques.

Point sur la *généralisation* des domaines : Quand la requête initiale est dirigée vers un domaine pas assez étendu, l'algo récupère des domaines connexes denses, qui ne sont pas forcément centrés au niveau intérêt sur la requête, mais plutôt sur des thèmes concurrents et plus densément reliés. Dans ce cas, l'algo a *diffusé* à partir de la requête originale. Les domaines récupérés forment souvent une généralisation de la requête.

Conclusion : l'utilisation des vecteurs propres non-principaux et le term-matching de base peuvent extraire des agrégats thématiques.

Principal reproche : on ne considère pas le nombre de liens unissant deux nœuds. Ceci est fait dans les extensions de ce travail : [BHA 98] , [CHA 98a] et [CHA 99c].

Idée pour la visualisation : on peut disp le graphe de sortie des h+a, et voir déjà s'il est connexe, ce qui montre peut-être dans certains cas l'existence ou pas de communautés différentes ayant des opinions ou des connaissances divergentes.

Problème technique soulevé : comment trouvent-ils les urls pointant vers les pages du root-set pour en faire l'augmented-set ? Par un crawl en profondeur P<sup>+20</sup> ? Par un graphe venant du crawl d'altavista ? Les deux ?

- [LAW 99] LAWRENCE S., LEE GILES C., (NEC Research Institute) « **Accessibility of information on the web** », *Nature*, vol. 400, juillet 1999.
- [LAW 98] « **Searching the World Wide Web** », Steve Lawrence, C. Lee Giles, *Science*, vol.280, April 1998
- [LEM 00] LEMPEL R., MORAN S., "The Stochastic Approach for link-Structure analysis (SALSA) and the TKC Effect" in *www9*.  
Présentation de la métaheuristique rapprochant [KLE 98] et SALSA, et l'approche à base de marches aléatoires sur le Web, modélisée par des chaînes de Markov. Propose également un correctif théorique sur les hubs et autorités – à reprendre.
- [MUR 00] MURRAY M., CLAFFY K.-C., "Measuring the Immeasurable : Global Internet Measurement Infrastructure", supported by the *National Science Foundation*, "The Internet Atlas" project, 2000.
- [SAR 00] SARUKKAI R. R., "Link Prediction and Path Analysis Using Markov Chains", (*www9*).
- [TU 00] TU Y., « **How Robust is the Internet ?** », *Nature*, 406, pp.353-354, 2000.

### *Usages, Navigation*

- [MAG 98] Paul P. Maglio and Teenie Matlock, "Metaphors we surf the web by" in *Workshop on Personalized and Social Navigation in Information Space*, Stockholm, Sweden, 1998  
Montre que la métaphore spatiale est naturelle chez les web-users, quelque soit leur « niveau d'expertise ».
- [PIR 96] PIROLLO P., PITKOW J., RAO R., « **Silk From a Sow's Ear : Extracting Usable Structures From the Web** », *Proceedings of CHI'96*, Palo Alto, CA, 1996.  
Présente la notion d'*information foraging* pour décrire les usages des humains sur le web par analogie aux stratégies de chasse des prédateurs. Le Web est vu

comme une mémoire externe (prothèse mémorielle). Résume l'InfoVis actuelle au mappage des propriétés des documents sur des items d'une représentation. Décrit plusieurs types de pages et les types de données disponibles pour décrire un « document » pour les systèmes de navigation : topologie, méta-données, voie et fréquence d'usage, similarité textuelle. Propose alors des voies pour prédire le degré d'intérêt de la page pour l'utilisateur – assez similaires au pageRanking de Google, dernière version, avec détection de profil utilisateur suivant les stratégies de recherche adoptées et système apprenant pour traiter les nouvelles démarches.

### Information Retrieval

- [BHA 01] Krishna Bharat, Bay-Wei Chang, Monika Henzinger, and Matthias Ruhl. "**Who Links to Whom: Mining Linkage between Web Sites**" in *Proceedings of the 2001 IEEE International Conference on Data Mining (ICDM)*, 2001, pp. 51-58
- [BRI 98] [Page et al 98] S. Brin and L. Page. "**The anatomy of a large-scale hypertextual web search engine**" In WWW Conference, volume 7, 1998  
Présentation de Google et particulièrement de l'algorithme de PageRank.
- [HEN 03] Monika Henzinger, Bay-Wei Chang, Brian Milch, and Sergey Brin. "**Query-Free News Search**". To appear in *Proceedings of the 12th International World Wide Web Conference (WWW2003)*, 2003  
Introduit l'utilisation d'un système de reconstruction de requêtes à partir d'un document pour en trouver des similaires.
- [HEN 02] Monika Henzinger, Rajeev Motwani, and Craig Silverstein. "**Challenges in Web Search Engines**". *SIGIR Forum*, Fall 2002  
Identifie les problèmes récurrents actuellement pour les moteurs de recherche.

## INTERFACE ET SYSTEMES D'INFORMATION

- [ABR 99] ABRAMS M., PHANOURIOU C., BATONGBACAL A.-L., WILLIAMS S.-M., SHUSTER J. E., "**UIML : an Appliance-Independent XML User Interface Language**" *www8*.
- [BAR 03] Barkowsky, T., Bertel, S., Engel, D., & Freksa, "**Design of an Architecture for Reasoning with Mental Images**" in *International Workshop on Spatial and Visual Components in Mental Reasoning about Large-Scale Spaces*. 01-02 Sept 2003, Bad Zwischenahn, Germany
- [BAR 01] Bertel, S. "**Benutzerunterstützung im World Wide Web mit Hilfe räumlicher Konzepte**" - Diplomarbeit 2001, *Department for Informatics, Universität Hamburg*  
Non-lue en entier, mais l'incipit est alléchant : contextualiser la navigation à l'aide de concepts spatiaux.
- [CHA 93] CHALMERS M., "**Using a landscape metaphore to represent a corpus of documents**" in *Spatial Information Theory: A Theoretical Basis for GIS*, vol.716, Springer Verlag, Berlin, 1993.
- [CHA 95] CHALMERS M., "**Design Perspectives in Visualizing Complex Information**" in *Proceedings, IFIP Third Visual Databases Conference, 27-29 march, Lausanne, Switzerland*, 1995.
- [COU 98] COUCLEDIS H., « **Worlds of Information: the Geographic Metaphor in the Visualization of Complex Information Systems** » in *Cartography and*

*Geographic Information Systems*, 25 (4) : 209-20.

- [DIE 95] DIEBERGER A., "**Providing Spatial Navigation for the WWW**", Georgia Institute of Technology, 1995.
- [FAB 91] FABRIKANT S.-I., BUTTENFIELD B., "**Formalizing Semantic Spaces for Information Access**", *Annals of the Association of American Cartographers*, 91(2), 2001, p.264.
- [GER 01] GERSHON N., PAGE W., « **What storytelling can do for Information Visualization** », *Communications of the ACM*, pp.31-37, August 2001, vol.44, number 8, ACM Press.
- [GHI 02] GHITALLA F., "**L'Age des cartes électroniques : les outils graphiques de navigation sur le web**" in *Communications et Langages*, n°131, A. Colin, avril 2002.
- [GHI 01] GHITALLA F., "**Arpenter le web : liens, indices, cartes**" in *Terminal*, n°86, L'Harmattan, 2001.
- [JOH 97] JOHNSON Steven, "**Interface Culture**", *Basic Books*, New York, 1997.
- [KEI 01] KEIM D.-A., "**Visual Exploration of large Data Sets**", *Communications of the ACM*, pp.39-43, August 2001, vol.44, number 8, ACM Press.
- [LAM 94] LAMPING J., RAO R., "**Laying out and Visualizing Large Trees using Hyperbolic Space**" in *Proceedings of the ACM Symposium on User Interface Software and Technology (UIST)*, Marina Del Rey, CA, nov.2-4, ACM Press, pp.13-14, 1994.
- [MUK 00] MUKHERJEA S., "**WTMS : A System for Collecting and Analysing Topic-Specific Web Information**", in *www9*
- [RAS 00] RASKIN J., **The human interface, new directions for designing interactive systems**, *ACM Press*, 2000.
- [ROS 99] ROSSI G., SCHWABE D., LYARDET F., "**Improving Web Information Systems with Navigational Patterns**" in *www8*
- [ROU 99] ROUSSEAU F., GARCIA-MACIAS J. A., VALDENI DE LIMA J., DUDA A., « **User Adaptable Multimedia Presentations for the WWW** » in *www8*
- [SHN 97] SHNEIDERMAN B., BYRD D., CROFT W. B., "**Clarifying Search, A User Interface FrameWork for Text Searches**", *D-Lib Magazine*, March 1997.
- [WAN 99] WAN E., ROBERTSON P., BROOK J., BRUCE S., ARMITAGE K., "**Retaining Hyperlinks in Printed Hypermedia Document**", Canon Information Systems Research Australia (*www8*).

## INFORMATION VISUALIZATION

### Theory

- [BAT 97] BATTY M., "**Virtual Geography**", *Futures*, Vol.29, n°4/5, Pergamon Press.
- [BER 67] BERTIN J., « **Sémiologie Graphique** », *Gauthiers-Villars - Mouton, Paris - La Haye*, 1967.
- [DOD 01] DODGE M., KITCHIN R., "**Mapping Cyberspace**", *Routledge*, London, 2001.

- [DOD 00] DODGE M., KITCHIN R., "**Exposing the 'Second Text' of Maps of the Net**" in *Journal of Computer-Mediated Communication*, vol.5, 2000.
- [GOO 97] GOODCHILD M., « **An interview with Michael Goodchild** », *Environment and Planning D : Society and Space*, vol.17, California, 1999. Voir aussi M. Batty « Virtual Geography », *Futures*, vol.29, n°4/5, Pergamon Press, 1997.
- [EIC 01] EICK S.-G., "**Visualizing online Activity**", *Communications of the ACM*, pp.45-50, August 2001, vol.44, number 8, ACM Press.
- [HAR 99] HARPOLD, T., "**Dark Continents : Critique of Internet Metageographies**", *Postmodern Culture*, January 1999.
- [HOD 99] HODGE D.-C., JANELLE D.-G., "**Information, Place and Cyberspace**", Springer-Verlag, Berlin/Heidelberg, 1999.
- [JIA 00] JIANG B., ORMELING, F. "**Mapping Cyberspace : Visualizing, Analysing and Exploring Virtual Worlds**", *The Cartographic Journal*, Vol.37, n°2, December 2000, pp.117-122.
- [KOP 00] KOPPEL J.G.S., "**No "There" there; Why Cyberspace Isn't Anyplace**", *The Atlantic Monthly*, Vol.286, August 2000, n°2, pp.16-18.
- [LYN 60] LYNCH K., « **L'Image de la Cité** », 1960, *M.I.T. Press*, 1976, *Dunod*.  
Présente les résultats d'une étude demandant à des usagers d'une cité de dessiner ce qu'ils pensent de leur ville – fournit un outillage conceptuel pour classer en unités significatives les éléments des plans mentaux des usagers (nœuds, frontières, point de repère, ...)
- [MAC 94] MACEACHREN A.E., "**The Use of Maps**", 1994.
- [MAC 94 a] MACEACHREN, "**Visualization in modern cartography**", *Pergamon*, 1994.
- [MAC 94 b] MACEACHREN, A.E., "**The Use of Maps**", *Guilford Press*, New-York, 1994.
- [MIN 99] Ming C. Hao, Meichun Hsu, Umesh Dayal, Adrian Krug, "**Visual Mining Large Web-based Hyperbolic Space Using Hidden Links** ", in *The Third International Conference on The Practical Application of Knowledge Discovery and Data Mining*, PADD'99, April.
- [MOL 70] MOLES A., « **L'Image, Communication Fonctionnelle** », *Dunod*, Paris, 1970.
- [MUN 00a] MUNZNER T., "**Interactive Visualization of Large Graph and Networks**" in *Ph.D. Dissertation*, Stanford University, June 2000 (graphics.Stanford.edu)  
La dissertation d'une chercheuse très connue, proposant 3 systèmes de visualisation de l'information. Finit sur une théorie générale de l'infoVis à partir de son expérience, qui met en place des concepts pour le design d'interfaces.
- [MUN 00b] MUNZNER T., RISDEN K., CZERWINSKI M., COOK D., "**An Initial Examination of Ease of Use for 2D and 3D Information Visualizations of Web Content**" in *International Journal of Human-Computer Studies*, vol. 53, pp. 695-714, 2000  
Présente l'outil XML3D, browser intégrant une représentation 3D du graphe visualisé, et des outils contextualisants (nœud suivants et précédents) et linguistiques (requêtes). Testé chez des webmasters pour le design de sites.
- [RIJ 99] RIJKEN D., "**Information in Space: Explorations in Media and Architecture**", *Interactions*, ACM Press, pp.44-57, June 1999.
- [SHN 99] SHNEIDERMAN B., CARD S.-K., MACKINLAY J.-D., "**readings in Information Visualization, using vision to think**", *Morgan-Kaufmann Publishers*, New-York, 1999.
- [WAR 00] WARE C., "**Information Visualization : Perception for Design**", *Academic Press*, San Diego, CA, 2000.

## COMMUNITIES

- [BRY 97] BRYNJOLFSSON E., VAN ALSTYN M., "**Electronic Communities : Global Village or Cyberbalkans?**" *Science*, 29/11/96, pp.1479-1480.
- [DIE 00] DIEBERGER A., DOURISH P., HOOK K, RESNICK P. and WEXELBLAT A., "**Social Navigation: Techniques for Building More Usable Systems**", *Interactions*, dec. 2000, Richard Morell/The Stock Market.
- [FAL 98] FALK J., "**The Meaning of the Web**", *The Information Society*, n°14, pp.285-293, 1998.
- [KUM 99] KUMAR R., RAGHAVAN P., RAJAGOPALAN S., TOMKINS A., "**Trawling the Web for Emerging Cyber-Communities**", in *www8*
- [STA 00] STAAB S., ANGELE J., DECKER S., ERDMANN M., HOTHO A., MAEDECHE A., SCHNURR P., STUDER R., SURE Y., "**Semantic Community Web Portals**" *www9*
- [UBO 01] Jeff UBOIS, « **Casting an Information Net** » in *UpsideToday*, 2001.

## URLOGRAPHIE

- <http://www.cs.cornell.edu/home/kleinber/> La page personnelle de Jon M. Kleinberg au Department of Computer Science, Cornell University
- techniques for analyzing and modeling link structure in the World Wide Web and related information networks; discrete optimization and network algorithms; algorithmic approaches to clustering, indexing, and data mining
- <http://www.cs.cornell.edu/Courses/cs685/2002fa/> La biblio du cours de Kleinberg sur les réseaux documentaires
- Résultat de la veille conduite par un spécialiste du domaine. Contient tous les papiers importants parus sur le sujet et leurs versions digitales. *Un Must absolu !*
- <http://www.casa.ucl.ac.uk> The Centre for Advanced Spatial Analysis
- Geographic information systems (GIS), computer-aided architectural design, spatial analysis and simulation, and methodologies for planning and decision support.
- <http://www.ccom.unh.edu/vislab/CWBio.html> La page personnelle de Colin Ware au Data Visualization Research Lab, Center for Coastal and Ocean Mapping, University of New Hampshire
- Page perso de Ben Shneiderman  
<http://www.cs.umd.edu/users/ben/>
- Base de Computer Science Papers @ NEC  
<http://citeseer.nj.nec.com/cs>
- Education Center on CS : InfoVis page  
<http://www.edcenter.sdsu.edu/faculty-fellows/fall2002/gawron/lingvizlinks.htm>
- InfoViz Conf.  
<http://www.infovis.org/>
- Une biblio sur les projets d'agents web  
<http://melanie.teamcircuitbreaker.com/lit/index.html>
- Barabasi et al., page de recherches sur les modèles de Réseau à Notre Dame  
<http://www.nd.edu/~networks/papers.htm>
- Le site du groupe Réseaux, Territoires et Géographie de l'Information à l'UTC  
<http://www4.utc.fr/~so03/>
- La page perso de Soumen Chakarbaty  
<http://www.cs.berkeley.edu/~soumen/>
- La page perso de Steve Lawrence  
<http://www.neci.nec.com/~lawrence/>
- La page perso de Tamara Munzner  
<http://graphics.stanford.edu/~munzner/>
- La page du projet IBM CLEVER  
<http://www.almaden.ibm.com/cs/k53/clever.html>
- La page de Monika Henzinger  
<http://www.henzinger.com/>
- Cybergéographie en français (!)  
[http://www.nicolas-guillard.com/cybergeography-fr/atlas/info\\_spaces.html](http://www.nicolas-guillard.com/cybergeography-fr/atlas/info_spaces.html)

**TABLE DES MATIERES**

**TOPOLOGIE DU WEB ET VISUALISATION DE L'INFORMATION :  
UN ETAT DE L'ART SCIENTIFIQUE ET TECHNIQUE ..... 1**

1. TOPOLOGIE DES RÉSEAUX HYPERMÉDIAS OUVERTS ..... 1

*La théorie des graphes découvre la topologie* ..... 2

        Barabasi et l'invariance d'échelle ..... 2

        Barabasi et le modèle évolutif ..... 3

*Topologie et sémantique hypertextuelle*..... 3

        Kleinberg et la théorie des agrégats..... 3

        L'algorithme de détection par renforcement mutuel: premier niveau..... 3

        L'algorithme de détection par renforcement mutuel: second niveau ..... 3

        Heuristiques annexes et conséquences théoriques pour l'interface ..... 3

        SALSA et PageRank, deux inspirés..... 3

        Conclusion..... 3

2. INTERFACE ET VISUALISATION DE L'INFORMATION ..... 3

*GraphVis* ..... 3

*Mapping* ..... 3

*Théorie du support: conséquence pour le design* ..... 3

*Réalisations en InfoVis* ..... 3

**BIBLIOGRAPHIE COMMENTÉE :  
TOPOLOGIE ET USAGES DU WEB, INTERFACES, VISUALISATION ET  
SPATIALISATION DE L'INFORMATION ..... 3**

    TOPOLOGIE DES RÉSEAUX HYPERMÉDIAS OUVERTS ..... 3

        Web, connectivité, hyperliens ..... 3

        Usages, Navigation..... 3

        Information Retrieval..... 3

    INTERFACE ET SYSTEMES D'INFORMATION ..... 3

    INFORMATION VISUALIZATION..... 3

        Theory ..... 3

    COMMUNITIES..... 3

**URLOGRAPHIE..... 3**

**TABLE DES MATIÈRES ..... 3**