LA GEOGRAPHIE DES *AGREGATS* DE DOCUMENTS SUR LE WEB

Agrégat – assemblage hétérogène de substances ou éléments qui adhèrent solidement entre eux.
Syn. : agglomérat, conglomérat.
Fig. « agglomérats de raisonnements » (Proust).
(Petit Robert de la langue française)

Le web représente un « e-cosystem » ([1]) documentaire relativement inédit. Les principes de son organisation restent en effet mal connus, même si depuis quelques années on commence à percevoir certaines propriétés génériques de cet espace hypertexte ouvert et dynamique. A première vue, les documents web et leurs liens sont inextricablement mêlés. Ils forment un système complexe qui ne semble plus rien devoir à l'organisation physique des réseaux, se développant de lui-même au niveau le plus abstrait de toute cette série de couches de machines et de protocoles qui forment l'Internet. Ce système apparemment gouverné par le hasard présente pourtant des « régularités fortes » ([2]) qui permettent d'esquisser sa géographie. On sait qu'elle n'est pas régulée par ce type d'ordre qui est celui des structures arborescentes mais qu'elle agrège différentes zones de pôlarité qui concentrent la connectivité hypertexte. Maîtriser les lois de distribution et d'évolution de ces zones constitue un enjeu majeur de la recherche scientifique actuelle dans les champs des technologies associées aux moteurs de recherche, aux outils de veille ou encore à la sécurité sur les réseaux. C'est aussi un enjeu technologique et industriel comme le montrent les principes sur lesquels sont basés *Google* ([3]) ou des projets comme

1 B.-A. Huberman, *The Laws of the Web, patterns in the ecology of information*, The MIT Press, Cambridge, Massachussets, London, England, 2003.
2 Cf. HUBERMAN B., PIROLLI P., PITKOW J., LUKOSE R, “Strong Regularities in World Wide Web Surfacing”, *Science*, 280, pp.95-97, 1998.
3 Et son algorithme de *PageRanking*.

IBM-WebFountain ([4]).

Graphe et topologie

Pourtant cette géographie ne semble qu'à peine esquissée car des éléments aussi essentiels que la taille du web ou sa vitesse de développement constituent encore aujourd'hui des problématiques de recherche. Par exemple, en juillet 1994, *Lycos* dénombrait près de 54000 pages sur le web. En 1997 ce nombre a été estimé à 320 millions, à 800 millions fin 1998, l'on aurait dépassé le milliard début 2000 et les 3 milliards en 2002. Sans compter le « deep-web » que Bergman ([5]) évalue à l'équivalent de 550 milliards de pages en 2000. Et aujourd'hui? Il semble même risqué d'avancer des chiffres depuis cette date et les moteurs de recherche eux-mêmes ne visent plus à l'exhaustivité ([6]). Ces outils classiques qui fonctionnent souvent par requête et annuaire conjugués permettent de moins en moins de le recenser : à eux tous ils en couvraient à peine 12% en 2000 et ce chiffre serait à diviser par deux chaque année ([7]). Ces chiffres peuvent être discutés, notamment parce que les technologies d'*information retrieval* sur le web ont évolué depuis lors mais on retiendra que le taux de couverture du web en 2003 doit être voisin de 1,5% selon le modèle d'évolution que proposent Lawrence et Lee Giles ([8])! L'idée que l'on se fait du web à partir de l'utilisation d'un navigateur et d'un moteur de recherche avec ses résultats classés par liste représente donc une vision très limitée d'une possible géographie du réseau.

A grande échelle il est donc encore difficile d'apercevoir un principe général d'organisation tant du point de vue de la « taille » du web que des principes de son organisation. Pourtant, les premières études sur la cartographie du web à l'aide de la théorie des graphes ont permis d'isoler certains principes structurants et réguliers, autant de « lois » d'une géographie qui se construit depuis quelques années ([9]). Ces propriétés sont essentiellement de types statistique et topologique, assurant au réseau des régularités observables à différents niveaux. En considérant le web comme un vaste graphe où les pages (ou les sites) représentent des *noeuds* et les liens hypertextes des *arcs* entre ces noeuds ([10]), il devient possible de proposer un

4 Un partie significative des scientifiques qui ont travaillé sur la topologie du web ces dernières années participent au projet de « big blue ». Le projet *IBM-WEBFOUNTAIN* est en effet issu d'au moins deux projets scientifiques complémentaires et qui tous les deux sont basés sur une modélisation du web comme un graphe et sur "la théorie des agrégats" que l'on peut attribuer à J. Kleinberg : *the Clever Project* et le "Grand Central Station project". Robert Morris, directeur du centre d'Almaden, et Robert Carlson, responsable duu développement de *Web Fountain*, ont réuni une équipe dont on peut identifier quelques-uns des membres principaux. Tout d'abord Kumar, Raghavan, Rajagopalan, Sivakumar et surtout Chakrabarti, des chercheurs indiens très productifs dans le domaine de la modélisation du web comme graphe et des agrégats (probablement issus de l'un *des Indian Institut of Technology*). On citera aussi Tomlin, Gibson et surtout Tomkins à l'origine du projet sientifique "web Graph Structure" (selon le magasine *Fortune* "Tomkins soon found he could spot trends in public opinion and popular culture as they emerged and watch them migrate around the world like wildfire"). J. Kleinberg semble aussi associé au projet. Pour une bibliographie assez exhaustive des chercheurs, on pourra consulter les bibliographies produites pour le projet *TARENTe*.

5 BERGMAN M.-K., *The Deep Web : Surfacing Hidden Value*, Bright Planet Company, july 2000.

6 Les responsables du développement technologique de la plupart des moteurs de recherche préfèrent aujourd'hui parler de « content quality » en focalisant leur offre sur le recensement de sources web fiables et structurées.

7 LAWRENCE S., LEE GILES C., "Accessibility of Information on the Web", *Nature*, 400, july 1999. Ou encore : LAWRENCE S., LEE GILES C., "Searching the World Wide Web", *Science*, vol.280, April 1998.

8 Les deux auteurs travaillent aujourd'hui à la conception d'un moteur de recherche de seconde génération à Princeton.

9 Cf. KLEINBERG J.M., KUMAR R., RAGHAVAN P., RAJAGOPALAN S., TOMKINS A. S., Department of Computer Science, Cornell University IBM Almaden Research Center, "The Web as a graph: measurements, models, and methods", (http://www.almaden.ibm.com/cs/k53/clever.html).

10 Théorie des graphes dont on fait remonter l'origine au mathématicien suisse Euler.

ou plusieurs modèles généraux de cet espace et d'en calculer les régularités statistiques. La théorie des graphes permet ainsi d'appréhender des phénomènes qu'une trop grande complexité semblait masquer, comme ce fut aussi le cas dans les champs de la sociologie, de l'économie ou de la bilbiométrie. De surcroît, la méthode pour recueillir les données sur le web qui serviront à construire le graphe paraît simple techniquement : il faut « lâcher » un robot-*crawler* (autrement dit un petit programme) ([11]) à partir d'un ou plusieurs points d'entrée (des adresses web) et il va alors se déplacer de lien hypertexte en lien hypertexte sur le réseau en envoyant le résultat de ses pérégerinations à une base de données où sont stockés, a minima, les adresses des noeuds rencontrés et leurs liens respectifs. Ensuite, à partir de la base de données, plaçons sur un plan les noeuds représentant les documents web (pages ou sites, au choix) et les arcs indiquant la présence de liens (et éventuellement leur direction orientée). On verra ainsi apparaître (rapidement ou non selon la « distance » des points d'entrée) un graphe unique densément peuplé de noeuds dont le nombre moyens de liens pour chacun est d'environ 5 (2,7 liens entrants et 2,1 sortants en moyenne sur le web) ([12]). Certes, les méthodes d'extraction des données, les types de projection ou de calculs appliqués aux données peuvent être très complexes et varier grandement. Ce n'est pourtant pas là un signe de faiblesse ou d'incertitude : chaque mode de calcul appliqué aux données recueillies par le robot enrichit la description du web et de son organisation complexe en faisant varier certaines propriétés topologiques de l'espace représenté sous forme de graphe.

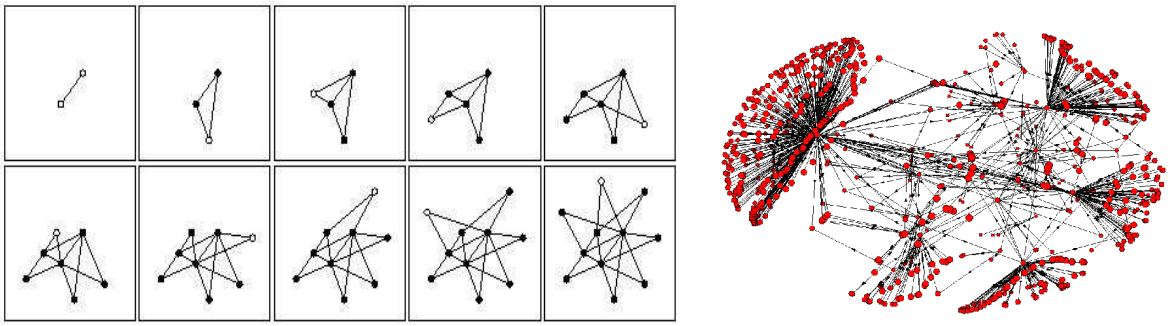

Fig.1 – A gauche, l'évolution schématique d'un graphe où les noeuds semblent s'agréger au fur et à mesure que de nouveux apparaissent. A droite, un graphe extrait du web à partir de cinq points d'entrée (robot-crawler «Orlinski » et visualisation sous *Pajek*). A cette échelle, on voit encore mieux le principe de la modélisation à base de graphes. Ces graphes ont des propriétés statistiques et mathématiques mais leur visualisation peut aussi beaucoup apporter au travail d'analyse, ne serait-ce qu'à titre d'outil d'exploration.

Noeuds dominants et loi de puissance

On a ainsi pu estimer le diamètre du web ([13]), en calculant notamment la distance moyenne (passage par des noeuds intermédiaires) entre deux noeuds pris au hasard. Elle varie de 5 à 7 avec un graphe non-orienté, les choses devenant plus compliquées avec un graphe orienté car le chemin n'existe alors pas toujours (la moyenne pouvant aller jusqu'à plus de 500 quand il existe)([14]). Il a aussi été proposé

11 Il en existe un certain nombre disponibles sur le web, comme *WebSphinx*.
12 Cf. note 14.
13 BARABASI A.-L., ALBERT R., “Emergence of Scaling in Randon Networks”, *Science*, vol.286, october 1999.
14 Andrei BRODER[1], Ravi KUMAR[2], Farzin MAGHOUL[1], Prabhakar RAGHAVAN[2], Sridhar RAJAGOPALAN[2], Raymie STATA[3], Andrew TOMKINS[2], Janet WIENER[3] 1: AltaVista Company, San Mateo, CA. 2: IBM Almaden Research Center,San-Jose,CA.,3:Compaq-Systems-Research-Center,PaloAlto,CA.(http //

un premier modèle d'organisation topologique à grande échelle . Le fameux « noeud papillon » issu aussi d'un travail de modélisation à base de graphes montre que le web est un espace « vectorisé » avec un centre d'attraction (un coeur très interconnecté) vers lequel pointent des milliards de liens et duquel en partent autant ([15]).

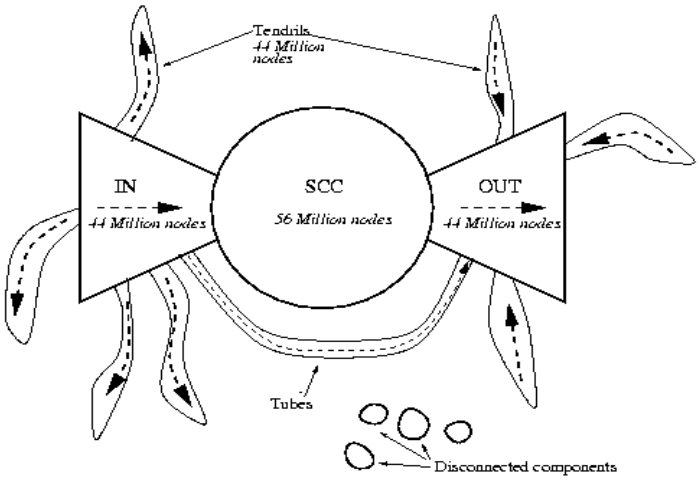

Fig.2 – Le « noeud papillon » extrait de l'étude de Broder et al. Il s'organise en quatre «zones » définies par des propriétés topologiques : le centre est un « strongly connected component » regroupant près de 25% des pages web. Elles sont très connectées entre elles. La partie gauche regroupe des pages qui pointent vers ce coeur, tandis que la partie droite regroupe des pages qui sont pointées par le coeur. Le web est donc un espace très vectorisé selon ce modèle. Au delà, ce sont des «tendrills » qui fonctionnent comme des espaces hypertextes relativement indépendants. On peut même imaginer, comme le montre lee bas du schéma, des agrégats de documents web totalement indépendants du reste du web.

On peut aussi travailler à une échelle très réduite, quelques milliers de noeuds seulement, et chercher ainsi à isoler des propriétés typiquement locales ou à comprendre comment s'organisent les chemins d'un noeud donné à un autre, ou choisis au hasard dans le corpus (*random walk*). Tous ces graphes ont une propriété commune : ils ne sont pas réguliers avec un nombre fixe de noeuds dès le départ et doté chacun du même nombre de liens. Mais ce ne sont pas non plus des graphes où les liens sont distribués aléatoirement car, dans ce cas de figure, nous aurions une répartition homogène des noeuds suivant qu'ils s'approchent ou s'éloignent d'une moyenne, significative pour ce type de réseau (*random network*). Non, le web et la distribution de sa connectivité hypertexte sont en réalité réglés par une loi de puissance (*power-law*) où certains noeuds concentrent le trafic en termes de liens ([16]).

www.alamden.ibm.com/cs/k53/www9.final/) "*Graph Structure in the Web".*

15 Ibid.

16 A.-L. BARABASI y voit le principe essentiel de la topologie du web : *linked*, *the new science of network*, Perseus Publishing, 2003.

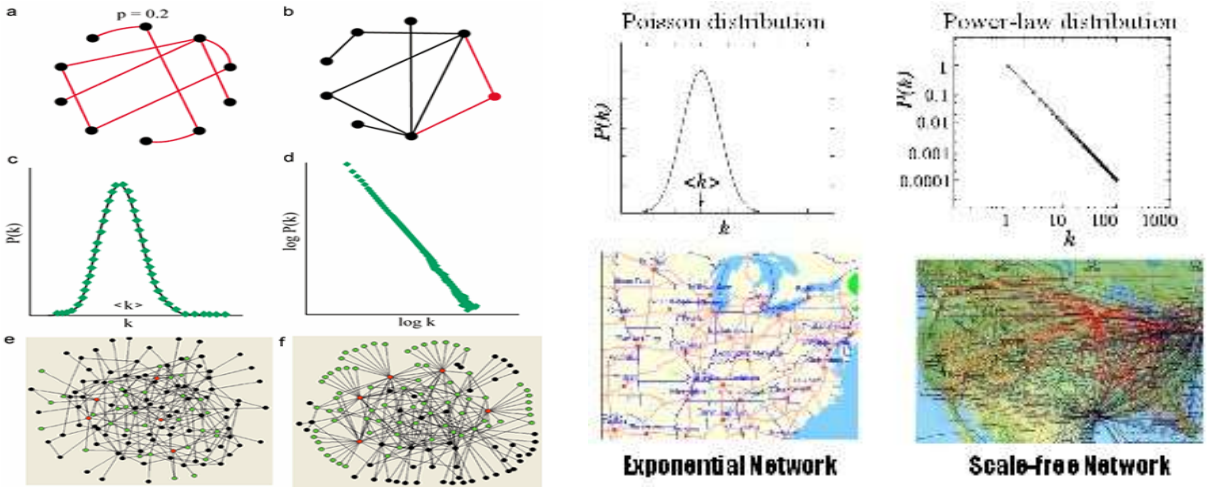

Fig.3-4 – Sur la partie gauche des deux figures, on voit que dans les « random networks » la distribution de la connectivité peut être mesurée par des moyennes pour l'ensemble du système, sous la forme d'une courbeou d'une « loi de Poisson ». On peut illustrer ce principe par la distribution, par exemple, des autorautes sur un territoire. En revanche, les réseaux régis par une «loi de puissance » ressemblent beaucoup plus à la distribution des lignes aériennes avec quelques gros « hubs » qui concentrent la connectivité.

Pour schématiser, nous sommes proches d'une loi de Zipf-Paretto où 20% des noeuds attirent et génèrent 80% des liens du réseau. On repère ces noeuds statistiquement mais aussi visuellement dans une organisation topologique où ils font saillance. C'est à ce noeuds dominants que J. Kleinberg attribuent des scores de *Hub* et *d'Authority* ([17]) à partir d'un algorithme ([18]) qui traite statistiqument leurs propriétés topologiques. Les *Hubs* génèrent beaucoup de liens sortants tandis que les *Authorities* fonctionnent comme des références auxquelles un grand nombre de liens renvoient. On comprend ainsi que la topologie du web est gouvernée par un maillage de noeuds dominants qui attirent ou diffusent la connectivité, les deux mouvements contraires tendant à se renforcer au cours du temps ([19]).

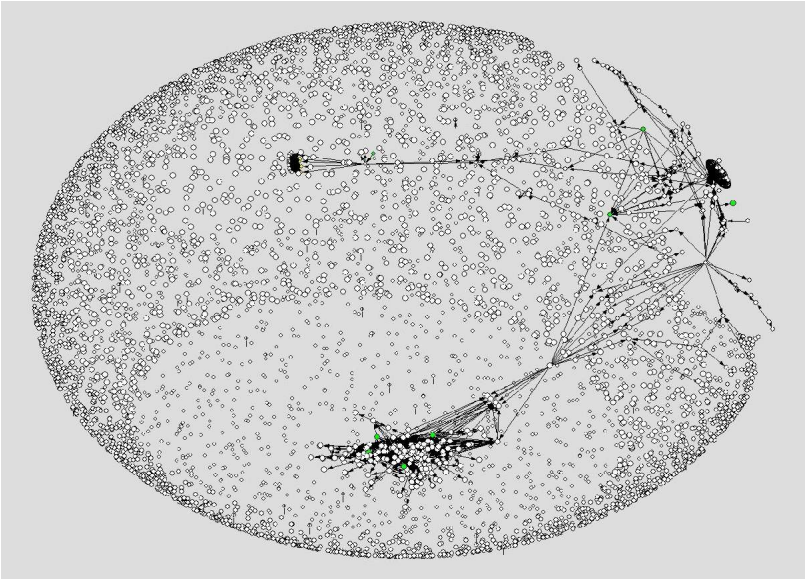

Fig 5 – Voici un graphe visualisé avec *Pajek* et produit à partir de *Human-Links*, une application de *knowledge management* asur

17 Sachant qu'un même site ou une même page peuvent se voir attribués les deux valeurs simultanément.
18 HITS, puis suivront une lignée de variantes comme ARC ou SALSA.
19 Hits accélère-t-il e le tempsde *distillation topologique*?

laquelle nous avons greffé l'algorithme de Kleinberg. On voit ainsi apparaître, en relief, des zones denses constituées de*Hubs* et *d'authorities*. En bas, la concentration visible représente un « agrégat » lié à « l'Eglise de Scientologie ». En haut et à droite, une autre zone de densité consacrée au « mouvement anti-scientologie ».

Ces 20% de noeuds sont-ils prépondérants pour leur ancienneté? Pour leur « qualité » ([20]) (comme des sites maintenus régulièrement à jour)? Comment s'organise, si elle existe, leur hiérarchie supérieure? En observant dans le temps la façon dont évoluent certains graphes, on pourra comprendre comment ils attirent à eux les nouveaux noeuds (*preferentiel attachment*) ([21]), certains modèles de simulation confirmant d'ailleurs l'idée qu'ils finiront par dominer le web (ou qu'un seul même le fera) ([22]). Ce sont ces noeuds qui en font un espace très interconnecté : ils lui assurent son unité et la garantie de chemins courts toujours possibles entre deux noeuds secondaires. Mais ils sont aussi le « talon d'Achille » du réseau, non seulement parce qu'ils permettent aux virus de se propager rapidement, mais aussi parce qu'en les neutralisant c'est l'unité et la structure mêmes du web qui se trouvent atteintes en raison de la position topologique de ces noeuds dans le graphe ([23]).

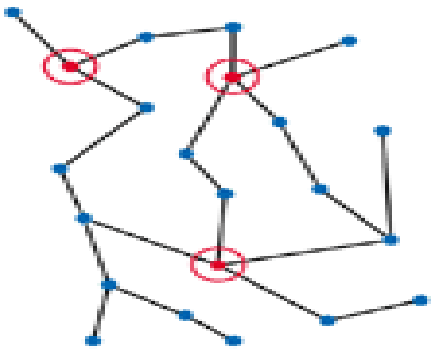

Fig 6 – Si les 20% de noeuds qui concentrent la connectivité jouent un grand rôle dans l'organisation du web, c'est qu'ils lui assure son unité et la présence de chemins courts entre un point et un autre, quels qu'ils soient. D'un autre côté, en cas d'attaque ciblée, leur disparition conduit à la fragmentation du réseau en multiples parties non reliées entre elles.

Les agrégats

Hubs, *Authorities* et l'ensemble des noeuds secondaires qui y sont reliés semblent se regrouper spontanément en agrégats. L'étude de leur morphologie interne constitue aujourd'hui un enjeu majeur de la recherche, notamment pour toute une série d'innovations technologiques qui pourraient en découler. On peut supposer qu'ils représentent au niveau local les briques élémentaires de l'architecture du web comme système documentaire. Les sonder, les isoler puis les « extraire » n'est pas chose aisée tant la méthodologie de leur analyse suppose la maîtrise de différentes variables encore peu testées. Mais on comprend que la démonstration expérimentale de l'existe d'agrégats structurés de document débouchera, notamment, sur l'étude de leurs relations réciproques, par juxtaposition à un même

20 Concept de « fitness » développé par R. Albert. Cf BARABASI A.-L., *linked*, *the new science of network*, Perseus Publishing, 2003.
21 Ibid.
22 Voir le chapitre de *Linked*. Ibid note 20.
23 Voir les deux articles de Barabasi : BARABASI A.-L., « The Physics of the Web » in *PhysicsWorld*, vol.14, Issue 7, IOP Publishing Ltd, 2001 et BARABASI A.-L., ALBERT R., JEONG H., “Error and Attack Tolerance of Complex Networks”, *Nature*, 406, pp.378-382, 2000.

« niveau » de sondage du web et par intégration successives dans des *clusters* de plus en plus larges. C'est là que se forgera l'essentiel de la « géographie du web ». Ce travail préliminaire a déjà été grandement mené ces dernières année, notamment par J. Kleinberg ([24]) et d'autres chercheurs qui se sont focalisés sur la question des propriétés inhérentes des agrégats comme sous-graphe du web. De façon complémentaire, ne seconde série de travaux a porté sur les technologies ou les algorithmes nécessaires à leur identification ou à leur exploration ([25]). A partir de l'ensemble de ces travaux, on peut dire que la question des agrégats se trouve réglée par deux types de problématiques complémentaires, celles des degrés d'échelle du web auxquelles on peut se placer et celle, qui se loge dans la première, des rapports entre le « sens » (constitué par ces constellations lexicales de mots clef présents dans les documents) et les données topologiques produites par la distributiondes liens hypertextes. En d'autres mots, il semble bien que les modes de corrélation entre contenu et topologie ne sont optimum *qu'à un certain degré d'échelle* du web. Et c'est à ce niveau-là d'abord (et peut être seulement) que résident les clefs qui nous feront concevoir le web comme un vaste système de documents organisés en agrégats, en espérant au delà construire les outils de son exploration.

Les questions d'échelle d'analyse du web ne sont pas clairement établies et il n'existe pas encore, à notre connaissance, de schéma articulé de leurs niveaux respectifs. Certains travaux sont consacrés à la modélisation du web à grande échelle ([26]), d'autres à des niveaux plus locaux ([27]) et ils dépendent parfois tout autant de la problématique adoptée que de la puissance des systèmes de calcul ([28]). On peut cependant tenter d'organiser un schéma approximatif des niveaux d'échelle du web qui rend compte en premier lieu de trois « couches » constitutives. Ainsi, à grande échelle, le web est d'abord lui-même un agrégat d'un point de vue purement topologique. Il tire son unité et son organisation du rôle que jouent des noeuds capitaux comme *Yahoo*, *Google*, *Amazon* ou *Microsoft*, autant de sites auxquels tous les autres sont reliés directement ou indirectement. Ces sites représentent un peu plus de vingt-pour-cent du web et forment le fameux « central component » du modèle en noeud papillon proposé par Broder et al. ([29]). A échelle moyenne ou réduite, cet espace est composé de milliers d'agrégats que J. Kleinberg estimait en 1997 à 100.000 ([30]). En deçà encore, se trouve la troisième couche, celle du « deep-web » et de l'univers des bases de données en ligne dans laquelle la plupart des robots-crawlers ne peuvent entrer ([31]).

---

24 En particulier, l'un de ses premiers articles très connus sur le sujet : KLEINBERG J., “Authoritative Sources in a Hyperlinked Environment“, Proceedings of the ACM-SIAM Symposium on Discret Algorithms, ACM Press,1998.

25 En explorant, en particulier, des rapports entre graphe topologique et analyse du contenu des pages web. Davison, « Unifying Text and Link Analysis », *IBM*, Palo-Alto, 2003.

26 Cf. note 14.

27 Voir KLEINBERG J., GIBSON David, RAGHAVAN Prabhakar, “Inferring Web Communities From Link Topology”, 1998. et KUMAR R., RAGHAVAN P., RAJAGOPALAN S., TOMKINS A., “Trawling the Web for Emerging Cyber-Communities”, (www8).

28 A grande échelle notamment : un graphe extrapolé de l'ensemble du web suppose d'intégrer plusieurs centaines de millions de noeuds et quelques milliards de liens!

29 cf. note 14.

30 KLEINBERG J., GIBSON David, RAGHAVAN Prabhakar, “Inferring Web Communities From Link Topology”, 1998.

31 M. Bergman présente en 2000 une technologie spécifique pour attaquer les bases de données en ligne, via la société « Bright Planet ». Cf. BERGMAN M.-K., *The Deep Web : Surfacing Hidden Value*, Bright Planet Company, july 2000.

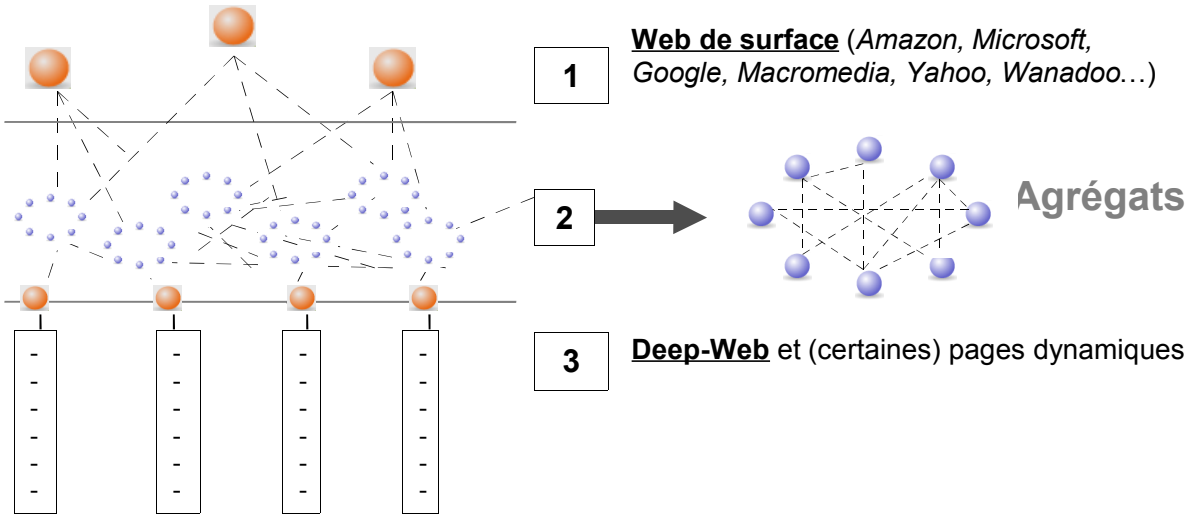

Fig 7 – Représentation schématique du web conçu comme un système de couches successives. Sonder les agrégats suppose donc une technologie qui se focalise sur la couche deux en « ignorant » l'importante connectivité développée dans la couche supérieure.

C'est à cette échelle moyenne ou réduite du web qu'il est possible de prendre en compte l'analyse du « contenu » des noeuds pour essayer de comprendre, notamment, comment l'organisation topologique en agrégats peut être corrélée à celle de la concentration-dispersion de mots-clefs spécifiques. C'est cette question capitale qui se trouve à l'horizon de la possible géographie du web comme espace documentaire inédit, à mi-chemin entre système d'information et topologie du graphe. La thématique des agrégats de documents web ouvre en effet une série de problématiques théoriques et de champs expérimentaux dans lesquels se logent quelques-uns des apports essentiels de la littérature scientifique actuelle, des *focus-crawlers* aux outils cartographiques de navigation sur le web.

Des liens et des mots

Epérimentalement, l'observation des agrégats, ou plutôt leur *extraction* n'est pourtant chose aisée. Certes, la théorie des graphes a constitué un pas décisif dans l'exploration de la structure topologique du web. Mais il s'en faut, et de loin, que l'on ait encore compris comment le « sens » est distribué dans cet espace, les méthodes de corrélation restant à inventer pour certaines. Les techniques de fouille de données textuelles sont pourtant nombreuses, et parfois très évoluées. Dans la plupart des cas, il s'agit de relever le contenu d'un noeud (page web), de le filtrer en sélectionnant essentiellement les « mots pleins » et enfin de les transformer morphologiquement pour les rendre comparables et regroupables. A minima, un tel

traitement permet d'obtenir un classement statistique de distribution à l'échelle d'un document ou d'un corpus. On peut, évidemment, raffiner le processus en tenant compte, par exemple, des spécificités des pages web : les balises *META* (content ou description) du code HTML peuvent fournir des indications pour pondérer le poids de certains mots-clefs, ou encore les balises *HREF* qui font porter le lien hypertexte sur certains mots ou leur entourage (technique de « l'anchor-window »).

Pourtant, comprendre la ou les façons dont ces mots-clefs et leurs rapports sémantiques éventuels peuvent être corrélés aux propriétés topologiques des agrégats constitue encore un enjeu de taille pour la recherche. On pourrait par exemple essayer de traiter sous forme de graphes indépendants les deux aspects, liens d'un côté, mots-clef de l'autre, pour ensuite les superposer. On peut aussi distribuer directement sur le graphe les données statistiques d'occurrence des mots-clefs : apparaîtront ainsi des zones de concentration, puis de dispersion pour chacun d'eux. Mais ce ne sont que des premières étapes car il faudrait pouvoir, à terme, associer à certaines propriétés générales du graphe des propriétés d'ordre sémantico-statistiques. Quelles sont, par exemple, les spécificités sémantiques des Hubs, ou des Authorities, dans un agrégat ([32])? Pourrait-on décomposer un agrégat en différentes sous -parties « géographiques » en fonction de la façon dont certains mots-clefs ont tendance à s'y associer en « formules » ([33])? Comment et quand passe-t-on sémantiquement et/ou topologiquement d'un agrégat à un *autre*?

Pour l'heure, des sondages ponctuels et localisés indiquent que les agrégats concentrent bien contenu thématique et connectivité. A priori, la corrélation est forte entre données topologiques et sémantiques si l'on considère le voisinage immédiat d'un ensemble de documents. Pour le dire autrement, les documents qui « parlent de la même chose » se trouvent bien localisés dans la même région sans que l'on puisse pour l'instant proposer un modèle cohérent des rapports entre analyse du contenu et graphe topologique. Et ce modèle est d'importance pour la construction d'une géographie valide qui déterminera *in fine* la pertinence des concepts et des outils développés pour maîtriser l'incertitude native du web. La discussion sur les modèles topologico-sémantiques du web reste ouverte. On peut déjà en envisager de deux types, sans pouvoir dire pour l'heure s'ils sont mutuellement exclusifs ou alors coexistants suivant les situations. Dans un premier cas, on peut supposer que le web, au niveau local, est constitué sur toute sa « surface » d'un maillage relativement homogène de noeuds dominants (*Hubs* et *Authorities*) où se sont essentiellement les différences de « contenus » (concentration/dispersion de mots-clefs) qui feront apercevoir des agrégats et leurs frontières respectives. Dans ce cas là, seul la distribution du « sens » permet de « découper » en ensembles homogènes la structure du web. Dans l'autre cas, les agrégats résultent de propriétés topologiques identifiables *à-priori* sur lesquelles peuvent alors faire porter des méthodes d'analyse de contenu pour comprendre à quelle particularité thématique répond une concentration de connectivité et de noeuds. Dans ce second

32 Un crawl expérimental sur le domaine de « l'infovis » a montré que les *Authorities*, en termes de contenu, pouvaient remplir plusieurs fonctions : *authorities* « pédagogiques » où est présenté synthétiquement le champ de l'infovis, *authorities* constituées de textes ou de sites de référence (en général d'auteurs) et *authorities* constituées, comme un hub, de signets web particulièrement pertinents pour le domaine.

33 Une analyse d'une série de sites issus d'un crawl expérimental sur le domaine de « l'anti-mondialisation » a montré que certains mots-clefs des balises META s'associaient en « formules » de type : thème+nom d'une association ou d'un syndicat ou d'un parti+nom d'une ville ou d'un rassemblement (exemple : « anti-mondialisation+SUD+Gênes). Dans d'autres cas on trouvait thème+problème spécifique+développement d'un concept (exemple : « anti-mondialisation+vivisection+protection de la nature).

cas, le terme d' »agrégat » prend tout son relief car la manipulation expérimentale de structures topologiques correspondrait aussi, au niveau local, à la manipulation d'unités organisées de « sens ». Mais peut être faudrait-il aussi envisager l'idée que les modes de corrélation entre topologie et contenu des documents divergent selon les *web localities* ([34]) étudiées, alternant au fil des données ramenées par un même crawl.

C'est dire combien il faut admettre méthodologiquement la relativité des modèles de corrélation proposés. Et il ne s'agit pas de la réduire mais de l'exploiter en accumulant théoriquement les modèles et expérimentalement les indices. La pluralité des modèles disponibles enrichit les possibilités d'exploration du web si l'on admet que l'agrégat n'est jamais donné à l'observation mais toujours préalablement calculé à partir, non pas d'un, mais d'une série d'algorithmes. C'est pourquoi il faut intégrer le principe de l'alternance des modes de corrélation aux outils dont on se dote pour extraire des « agrégats ». C'est tout l'intérêt, expérimentalement, de disposer d'un jeu de « filtres » différents dans des tâches d'exploration et de sondage. L'extraction de données topologiques et sémantiques via un robot-crawler représente en effet une phase de *scanning* du web. L'application (pour l'heure non-automatisable) du jeu de filtres représente quant à elle une phase de *tuning* où il s'agit de « caler » sur les données, à un moment donné, le ou les types de calcul qui stabiliseront l'agrégat sous une forme de corrélation optimale. Il faut ici imaginer un curseur théorique qui permet de figer ensemble, en une forme relativement singulière, propriétés topologiques du graphe et distribution des mots-clefs dans l'espace.

Des chantiers, en cascade

Eprouver différentes méthodes de corrélation entre « sens » et lien hypertexte, entre contenu et connectivité, n'est d'ailleurs qu'un point de départ, à peine les directions cardinales du territoire à cartographier. Toute une série de chantiers de recherche restent à développer, ou à ouvrir une fois stabilisée la morphologie générale de l'agrégat. Ce dernier représente, par exemple, un principe d'organisation des données, exploitables pour constituer un ensemble de ressources hiérarchisées (Hubs/Authorités et l'ensemble des noeuds secondaires) et, éventuellement, organisées « localement » pour peu que l'on puisse déterminer des sous-ensembles sémantiques dans la structure générale du graphe. De là, l'importance d'expérimentations renouvelées sur différents agrégats thématiques pour comparer les phénomènes et les principes d'organisation. Une typologie serait donc alors possible, contribuant à enrichir la description de la topologie du web. De là, aussi, une réponse possible à la maîtrise du web comme espace documentaire, par exemple dans le domaine de la documentation ou de la veille sur certains sujets. Mais un agrégat se détermine aussi par ce qui le différencie d'un autre, ou plutôt de plusieurs autres qui en sont à la fois les voisins et les repères. Du point de vue topologique, on peut penser à des zones de « faible densité » hypertexte *entre* agrégats alors que ces derniers concentrent la connectivité, ou plus simplement à des variations de densité dues à des pratiques ou des traditions de liaison

34 Cf. J. Kleinberg, D. Gibson, P. Raghavan, “Inferring Web Communities From Link Topology”, In *Proc. of the 9th ACM Conference on Hypertext and Hypermedia* (HYPER-98), pages 225--234, New York, June 20--24 1998.

hypertexte différentes d'une communauté d'acteurs à une autre ([35]). Du point de vue sémantique, c'est tout le problème des seuils sur lesquels on joue pour intégrer ou non à un « domaine » des données moyennement ou faiblement représentées. A cet égard, on peut supposer que des ensembles lexicaux déterminés constituent des zones de densité et qu'ils épousent de plus ou moins grandes tailles, à l'image d'une sorte *d'amplitude sémantique*. C'est là, peut être, que l'on comprend le rôle majeur que joue la topologie des agrégats quand elle contraint à regrouper ensemble des mots-clefs que l'observateur n'aurait jamais songé à associer ([36]), ou à envisager des relations de voisinage entre une série d'agrégats qui rien ne semblait rapprocher ([37]). La géographie thématique du web n'est pas encore dessinée mais il y a fort à parier qu'elle n'épousera pas toujours les façons dont sont organisés les champs de savoirs dans une bibliothèque.

Et l'on peut aussi intégrer une observation temporelle des agrégats qui, une fois isolés, peuvent en effet être à nouveau sondés régulièrement. On pourrait imaginer un *tracker* qui fonctionnerait par *reporting*, indiquant chaque niveau de changement important. La durée de vie d'une page web étant en moyenne de cent jours ([38]), on imagine sans peine ce qu'un graphe animé (ou une carte) de tous ces changements pourrait nous apprendre les scénarios d'évolution temporelle des agrégats. Comment évoluent-ils? Si dissocient-ils ou, au contraire, peuvent-ils « fusionner »? Tous ces éléments topologiques et temporels représentent un cadre de contextualisation très pertinent pour toutes ces « informations » ou ces « documents » sur lesquels, bien souvent, on ne fait que « tomber » sur le web sans pouvoir les situer dans un territoire. C'est pourquoi l'étude expérimentale des agrégats sur le web pourrait aussi déboucher des procédures d'archivage des documents numériques, en termes de contenus ou de formats techniques mais aussi de position à l'intérieur d'un territoire qui les situe. Il s'agirait là des premières briques d'une « mémoire » à la fois topologique, thématique et finalement sociale des réseaux.

Des modèles d'univers

Il reste à articuler en une vision cohérente l'ensemble des premières données recueillies sur les agrégats, leurs relations réciproques, les principes de leur composition et de leur évolution. Le modèle devrait permettre d'articuler cette géographie de ces localités thématiques à dautres phénomènes plus généraux, comme la diamètre du web ou le rythme de son expansion. Peu existent encore aujourd'hui : ils sont essentiellement spéculatifs et s'appuient rarement sur des expérimentations à petite comme à grande échelle. Parmi eux, le plus cohérent avec une géographie des agrégats telle que nous venons de la décrire est un modèle de type *gravitationnel du web* [BEN 03]. Il faut supposer, si on le suit, qu'un lien hypertexte entre deux pages (ou un lien transverse entre deux sites) equivaut à un jeu de forces qui s'exerce entre ces deux pages. Les deux pages s'attirent ainsi l'une l'autre mais pas de façon égalitaire : seulement celle qui a la plus grande masse attire à elle la seconde. Et comment, alors, déterminer les masses

35 Idée que l'on doit à Eustache Diemert.

36 Cela s'est vérifié plusieurs fois, notamment en s'intéressant au domaine de la pêche en « surfcasting ». Dans la rapide lexicographie du domaine que nous avions construite pour analyser les balises META des pages web, figurait en bonne place le terme « turlute » à côté « d'appât » ou de « leurre ».

37 Cela a été le cas de l'apercevoir quand en travaillant sur le domaine des « aveugles et de la « cécité visuelle » nous avons commencé à trouver beaucoup de références aux industriels de l'alimentation animale (probablement à cause du thème récurrent dans ce domaine des « chiens pour aveugles »).

38 Cf. note 7.

respectives des pages ou des sites qui ne sont pas des entités physiques communes? Là, les options divergent, en retrouvant d'ailleurs encore la problématique de l'origine et de l'évolution des *Hubs* et des *Authorities* que nous avons rencontrée au niveau local des agrégats. Le calcul de la masse des objets web peut intégrer plusieurs facteurs : le nombre de liens entrants et/ou sortants, la « taille » du message (en octets par exemple), la description du contenu, l'ancienneté de l'élément, le nombre d'usagers qui y ont eu accès (mesure d'audience), ou plus probablement une combinaison subtile de tout cela. Des epxérimentations devront faire jouer ces facteurs en les combinant de façon originale, jusqu'à retrouver le principe explicatif de la topologie du web telle que l'on peut la sonder actuellement.

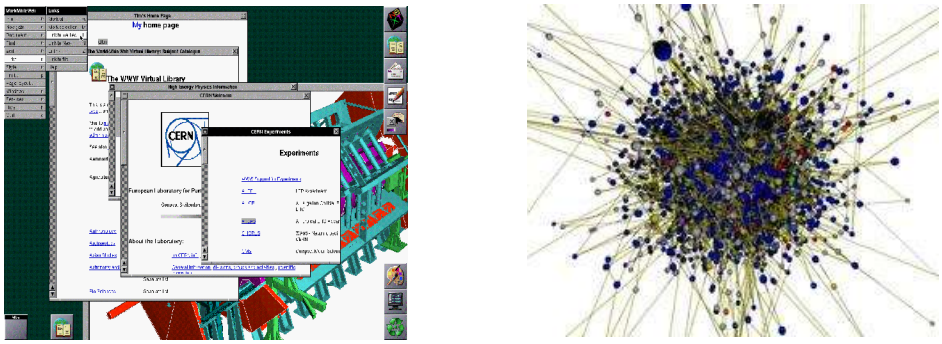

Fig 8 – Comment expliquer l'évolution de la topologie du web depuis la première page HTML diffusée par Tim Berners-Lee (à gauche)? (à droite : le travail exploratoire de Steven Coast en 2001 pour le projet *IP Mapping*)

Le modèle ne s'arrête pas là, car l'attraction liée à la gravitation se double d'un mouvement contraire et complémentaire nécessaire à l'explication de la géographie possible des agrégats. Sans ce mouvement contraire, l'univers web, quelle que soit sa taille, se concentrerait en un tout indistinct, une masse sans espace définissable. C'est en effet son expansion qui tend à écarter les une des autres ces particules que sont les pages ou les sites, éloignant en particulier de façon radicale les éléments non reliés entre eux ou très indirectement. Si l'on admet le doublement du web tous les six mois (ou quelle que soit l'estimation, du moment qu'il y ait bien expansion), on comprend l'importance du phénomène d'expansion (qui d'ailleurs, en théorie, diminue en fonction du vieillissement de l'univers). C'est donc entre expansion qui « écarte » les éléments (surtout peu liés) et gravitation (qui concentre les éléments en fonction de leurs masses respectives) que se situent des agrégats dont on commence seulement aujourd'hui à comprendre le rôle. Une question demeure pourtant : le nombre des agrégats augmente-t-il aussi vite que l'univers lui-même? Cela supposerait, évidemment, de pouvoir les dénombrer et d'être ainsi capable de suivre leur évolution, ce qu'ont probablement commencé à faire les experts d'IBM. De telles questions sur l'organisation de l'univers web montrent en tous les cas comment, aujourd'hui encore, modèle scientifique et fiction logique se conjuguent dans l'exploration des réseaux.

BILBIOGRAPHIE

[ADA 99a] ADAMIC L.-A., HUBERMAN B.-A., “Technical comment to ‘Emergence of Scaling In Random Networks’”, *Xerox Parc*, 1999.

[ADA 99c] ADAMIC L.-A., HUBERMAN B.-A., “Growth Dynamics of the World Wide Web”, *Nature*, vol.401, p.131, September 1999.

[BAR 03] BARABASI A.-L., *linked*, *the new science of network*, Perseus Publishing, 2003.

[BAR 01] BARABASI A.-L., « The Physics of the Web » in *PhysicsWorld*, vol.14, Issue 7, IOP Publishing Ltd, 2001.

[BAR 99b] BARABASI A.-L., ALBERT R., “Emergence of Scaling in Randon Networks”, *Science*, vol.286, october 1999.

[BEN 03] “Un modèle gravitationnel du web”, T. BENNOUAS, M. BOUKLIT, F. DE MONTGOLFIER, Les Journées “Graphes, Réseaux et Modélisation”, Paris, ESPCI, dec. 2003.

[BER 00] BERGMAN M.-K., *The Deep Web : Surfacing Hidden Value*, Bright Planet Company, july 2000.

[BOT 91] BOTAFOGO R.-A., SCHNEIDERMAN B., “Identify Aggregates in Hypertext Structures“ in the *Third ACM Conference on Hypertext*, San Antonio, ACM Press, pp.63-74, 1991.

[BRO 00] Andrei BRODER[1], Ravi KUMAR[2], Farzin MAGHOUL[1], Prabhakar RAGHAVAN[2], Sridhar RAJAGOPALAN[2], Raymie STATA[3], Andrew TOMKINS[2], Janet WIENER[3] 1: AltaVista Company, San Mateo, CA. 2: IBM Almaden Research Center, San Jose, CA. 3: Compaq Systems Research Center, Palo Alto, CA. (www.alamden.ibm.com/cs/k53/www9.final/ <http://www.alamden.ibm.com/cs/k53/www9.final/>) “Graph Structure in the Web”.

[CAR 97] CARRIERE J., KAZMAN R., « WebQuery : Searching and Visualizing the Web trough Connectivity » (http://www.cgl.uwaterloo.ca/Projects/Vanish/Webquery-1.html)

[CHA 99] CHAKRABARTI S., DOM B., GIBSON D., KUMAR R., RAGHAVAN P., RAJAGOPALAN S., TOMKINS A., “Experiments in Topic Distillation”, *Almaden Research Center*, CA, 1999.

[CLA 99] CLAFFY K.-C., “Internet measurement and data analysis: topology, workload, performance and routing statistics”, NAE’99 workshop.

[DAV 03] DAVISON B.-D., « Unifying Text and Link Analysis », *IBM*, Palo-Alto, 2003.

[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.

[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.

[GAR 99] GAROFALAKIS M. N., RASTOGI R., SESHADRI S., SHIM K. (Bell Laboratories), “Data Mining and the Web: Past, Present and Future”, (<http://www.bell-labs.com/user/minos/Papers/widm99.pdf>).

[HUB 03] HUBERMAN B. A., *the laws of the web, patterns in the ecology of information*, M.I.T. Press, 2003.

[HUB 97] HUBERMAN B., PIROLLI P., PITKOW J., LUKOSE R, “Strong Regularities in World Wide Web Surfacing”, *Science*, 280, pp.95-97, 1998.

[KLE 01c] KLEINBERG J., LAWRENCE S., « The Structure Of The Web », *Science*, vol.294, 30, november, 2001.

[KLE 98a] KLEINBERG J., CHAKRABARTI S., DOM B., RAGHAVAN P., RAJAGOPALAN S., GIBSON D. « Automatic Resource Compilation by analyzing hyperlink structure and associated text ».

[KLE 98b] KLEINBERG J., GIBSON David, RAGHAVAN Prabhakar, “Inferring Web Communities From Link Topology”, 1998.

[KLE 97] KLEINBERG J., “Authoritative Sources in a Hyperlinked Environment“, Proceedings of the ACM-SIAM Symposium on Discret Algorithms, ACM Press,1998.

[LAW 99] LAWRENCE S., LEE GILES C., “Accessibility of Information on the Web”, *Nature*, 400, july 1999.

[LAW 98] LAWRENCE S., LEE GILES C., ”Searching the World Wide Web”, *Science*, vol.280, April 1998.

[UBO 01] Jeff UBOIS, « Casting an Information Net » in *UpsideToday*, 2001.

[UTC 03] rapport de recherche interne sur le projet expérimental « TARENTe » inauguré en mars 2003 à l'Université de technologie de Compiègne.