Two Visions of the Web: from Globality to Localities

Fabien PFAENDER Mathieu JACOMY Guilhem FOUETILLOU

COSTECH Laboratory
University of Technology of CompiËgne
France

fabien.pfaender@utc.fr mathieu.jacomy@gmail.com guilhem.fouetillou@utc.fr

1. Introduction

Nous nous proposons de mettre en perspective l'approche de RTGI et l'approche maintenant classique des moteurs de recherche. Ces deux approches reposent sur une compréhension commune du web comme réseau ouvert, hétérogène et contenant une très grande masse de données. Tout corpus de pages hyperliées peut être calculé à la fois comme collection de documents numériques et comme graphe où les pages sont les noeuds et où les liens sont les arcs. Ainsi, les techniques utilisées par RTGI comme par les moteurs de recherche reposent sur la théorie des graphes et sur l'analyse de contenus calculables. Dans les deux cas, la mise en oeuvre passe par l'extraction et l'indexation des données du web. Nous nous proposons de etc. (plan)

2. Appréhender les formes du web : le cas des moteurs de recherche

Cette partie se donne pour objectif de mettre en perspective les moteurs de recherche produisant des listes, et les techniques permettant de produire des cartes. Deux aspects complémentaires du web comme graphe sont traités : les techniques de calcul et la représentation graphique.

2.1. Principes de calcul de graphes du web par les moteurs de recherche

Nous allons voir qu'à l'origine, les formes observables du web ont permis de concevoir des algorithmes aujourd'hui classiques, sur lesquels repose aujourd'hui une grande partie de la technologie des moteurs de recherche. Cependant ce sont l'observation et l'analyse du comportement des internautes qui ont guidé l'évolution de cette technologie. Sachant que la projection sous forme de graphe est un fondement de l'approche classique du web, nous pointons les divergences entre l'exploitation des formes du web et la production empirique de hiérarchies.

2.1.1. L'enjeu des premiers algorithmes de classement des ressources issues du web : mesurer le support de la navigation hypertextuelle

Les premiers algorithmes des moteurs de recherche, visant à restituer les ressources les plus pertinentes pour une requête donnée, sont construits pour exploiter des propriétés du web comme graphe. En particulier, les graphes peuvent être manipulés aisément sous forme de matrice, et les moteurs tirent parti de cette façon de coder les données extraites du web. Dans son article fondateur [Authoritative sources in a Hyperlinked Environment, 1998], Kleinberg l'énonce clairement :

Our methods seem to apply fairly broadly, to structures that are implicitly, as well as explicitly, linked. [...] We will show some interresting conections between our algorithms and the spectral properties of certain matrices derived from the link structure of the underlying environment ; it is through these connections that we will be able to develop some insight into their behavior, and to prove certain convergence properties.

L'algorithme HITS de Kleinberg exploite les structures bipartites des graphes du web pour restituer une mesure de l'autorité. La mesure d'autorité n'est une mesure de la pertinence que si les structures bipartites sont une propriété globale du web. Kleinberg sent que c'est le cas et il le justifie ainsi :

We claim that an environment such as the WWW is explicitly annotated with precisely the type of human judgment that we need in order to formulate a notion of authority. Specifically, the creation of a link in the WWW represents a concrete indication of the following type of judgment : the creator of page p, by including a link to page q, has in some measure conferred authority on q

Le succès de son algorithme et de ses dérivés lui donnera raison. Cependant, la dimension sociale du lien hypertexte n'est pas complètement analysée à l'heure actuelle et pour cette raison, il est difficile d'évaluer un algorithme au moyen d'arguments sociologiques. Par contre, il est possible d'observer le web pour en dégager des invariants, comme ses propriétés bipartites, et de construire les algorithmes appropriés tout en analysant ces invariants en sciences humaines.

En 1998 également, Sergey Brin et Larry Page publient l'article qui décrit leur futur moteur de recherche numéro 1, Google [The anatomy of a large-scale hypertextual web search engine]. Google utilise l'algorithme PageRank pour hiérarchiser les pages. Encore une fois, la justification intuitive que proposent les auteurs s'appuie sur les formes du web :

PageRank can be thought as a model of user behavior. We assume there is a 'random surfer' who is given a web page at random and keeps clicking on links, never hitting 'back' but eventuelly gets bored and starts on another random page. The probability that the random surfer visits a page is its PageRank. [...] Another justification is that a page can have a

high PageRank if there are many pages that point to it, or if there are some pages that point to it and have a high PageRank. Intuitively, pages that are well cited from many places around the web are worth looking at.

Encore une fois, les auteurs s'appuient sur les propriétés du web pour construire leur algorithme. Nous insistons sur le fait que malgré l'argumentation, ce ne sont pas les usagers mais les formes du web servant de support à la navigation qui permettent de conceptualiser l'algorithme. La dimension sociale du web est aujourd'hui exploitée sans être théorisée, et comme en 1998 des outils de dimension sociale enrichissent le web sans dimension théorique, uniquement sur la base d'intuitions. Ce sont les formes du web, fussent-elles sociales, et non l'analyse des usages, qui sont à l'origine des algorithmes des moteurs de recherche.

2.1.2. Aujourd ?hui, les moteurs se sont émancipés des formes du web

L'utilisation massive des moteurs de recherche par le grand public a changé la situation. Pour un moteur, l'efficacité des résultats renvoyés est un objectif à part entière : les algorithmes n'ont pas besoin de se référer aux formes du web. Certes, le modèle du surfeur aléatoire est encore très répandu. Cependant les moteurs, pour être le plus efficace possible, ont recours à diverses optimisations qui rendent les algorithmes de plus en plus difficilement interprétables. Les propositions pour faire évoluer le PageRank sont d'abord mathématiques, et les propriétés du web intuitées sont secondaires [An analytical comparison of approaches to personalizing PageRank, T. Haveliwala et al., 2005, http ://nlp.stanford.edu/pubs/comparison.pdf]. Le BlockRank [S. D. Kamvar, T. H. Haveliwala, C. D. Manning, and G. H. Golub. Exploiting the block structure of the web for computing PageRank. Stanford University Technical Report, 2003.] prend en compte l'importance des sites par rapport aux pages, et le Modular PageRank [G. Jeh and J. Widom. Scaling personalized web search. In Proceedings of the Twelfth International World Wide Web Conference, 2003.] donne une plus grande importance aux sites qui apparaissent en haut des résultats des moteurs. Dans ces deux cas, le référent n'est plus le web mais le comportement de l'internaute. Citons encore le Topic-sensitive PageRank [T. H. Haveliwala. Topic-sensitive PageRank. In Proceedings of the Eleventh International World Wide Web Conference, 2002.], qui classifie le web d'après le répertoire OpenDirectory : la structure du web n'est certainement pas conforme à ces répertoires, et donc cet algorithme ne se réfère plus aux invariants du web.

Les algorithmes permettant de hiérarchiser les ressources n'ont pour légitimité que leur efficacité. La valeur empirique de ces technologies, qui ont eu le succès que l'on connaît, nous oblige à nous poser la question suivante : y a-t-il un écart entre les hiérarchies des moteurs de recherche et les formes observables du web ? Pour répondre à cette question, nous allons maintenant aborder la question de la représentation des formes du web.

2.2. representing the web as a graph : manifesting shapes

Les modèles du web ont été construits dans le prolongement de la théorie des graphes. L'utilisation de cet appareillage théorique a des conséquences sur la vision ou l'appréhension que l'on a du web aujourd'hui à travers les différentes interfaces qui nous sont proposées au titre desquelles les navigateurs, les moteurs de recherche et leur liste ou encore les dispositifs synoptiques dont les cartes sont l'archétype.

2.2.1. Le graphe au coeur du web

La projection de la structure du web dans un graphe est considérée comme naturelle. La présentation en graphe et l'utilisation de ces derniers pour modéliser et spatialiser le web correspond à une expérience perceptive et cognitive très ordinaire qui est celle des navigateurs. De manière très simple, le navigateur affiche des pages et permet en cliquant sur les liens hypertextes d'afficher d'autres pages qui apparaissent être liées de manière causale les unes aux autres. Les boutons avance et recule sont une manière basique de se déplacer dans ce qui devient un espace [1]. La forme naturelle de cet espace composé de pages et de liens causaux entre ces pages est le graphe. Les pages y sont des noeuds et les liens, des arcs. L'espace du graphe devient ainsi représentable mathématiquement et l'on peut alors lui appliquer les nombreux algorithmes et propriétés qui lui sont associé. Toutefois il est également possible de le spatialiser, c'est à dire le présenter notamment graphiquement [2][3] et ce bien que ce mode de présentation soit encore largement inusité du grand public. Cette ambiguïté entre structures logique et spatiale conduit à deux utilisations différentes des graphes.

2.2.2. Double utilisation des graphes

Suivant l'objectif, on peut utiliser le graphe soit de manière analytique soit de manière synoptique. L'utilisation analytique du graphe se sert des propriétés mathématiques de ce dernier pour en réduire la complexité de manière automatique et suivant des propriétés connues. D'un ensemble de ressources web, l'analyse algorithmique permet de sélectionner un petit nombre de ressources que l'on restitue ensuite à l'utilisateur. Cette méthode est automatisable dans la mesure où il ne s'agit que de calculs formels. C'est notamment la solution adoptée par les moteurs de recherche. Ceci non seulement parce qu'elle représente un gain de temps évident mais surtout parce que cela correspond à un modèle du web globale sur lequel un calcul peut être appliqué. L'utilisation synoptique est utilisée à deux fins. La première est une utilisation analogue à la méthode analytique en ce qu'elle permet d'outiller une réduction de complexité. Au contraire d'une analyse automatique, l'utilisation synoptique de graphes aide un utilisateur humain à détecter des régularités et faits saillants qu'il ne saurait percevoir

autrement [4]. La réduction de complexité est donc visuelle. De même cette méthode est employée par la suite pour visualiser toujours sous forme de graphe le résultat de cette réduction de complexité. Cette analyse est exclusivement manuelle et requiert un utilisateur humain. Elle fait l'impasse sur le modèle en s'ouvrant à de nouvelles interprétations de ce dernier, en fonction de l'endroit du web étudié.

2.2.3.Du graphe à la liste et à la carte

La double utilisation des graphes conduit en conséquence directe à deux spatialisations des ressources différentes. On considère en premier le cas du graphe mathématiquement représenté et l'utilisation qui en est faite par les moteurs de recherche, c'est à dire proposer les ressources les plus pertinentes pour une requête donnée. Puisque l'analyse est analytique et que le modèle qui a servi à l'analyse est global, la spatialisation qui en résulte possède ces deux caractéristiques. Non pas qu'il soit obligatoire de retrouver l ?analytique et le global, mais la construction d'une visualisation n'est que la conséquence des concepts qui ont présidés à son élaboration ; en l'occurrence il s ?agit d ?une analyse globale de la pertinence, fruit d'un calcul de graphe. La visualisation la plus évidente dans ce cas est la liste de résultats. La liste est une structure logique. Elle s'inscrit dans le plan mais ne fait appel à une métrique particulière de la 2D. Il n'y a pas de principe de distance entre les éléments. Seule compte l'unité spatiale en une colonne ou ligne qui permet de voir la liste au milieu d'autres visualisations. Les éléments de la liste ont tous une taille uniforme comme on peut le voir sur la figure. Ainsi aucun élément de la liste ne possède

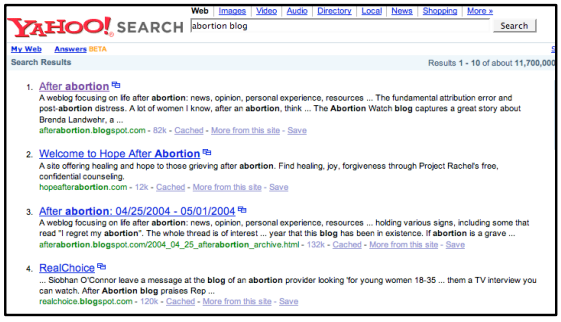

Figure 1Yahoo result list

plus d'importance qu'un autre, ils sont équivalents. Ce principe est important puisque sans cela la liste perdrait son caractère logique et deviendrait cartographique. De plus, l'ordre apparent des résultats classés du plus au moins pertinents ne résulte que d'une habitude de lecture du haut vers le bas. Rien dans la liste ne permet de supposer pareil classement. Yahoo a donc ajouté un numéro devant chaque ressource pour rendre ce classement apparent alors qu'il n'est que supposé dans la plupart des autres moteurs de recherche. Le fait que tous les résultats aient la même apparence et une distance égale sans métrique aucune est en accord avec l'idée de globalité véhiculée lors de l'analyse. La carte au contraire se fonde sur le placement en x et y de ses éléments et pas sur une logique où tous les éléments sont égaux. L'ordre logique n'est pas applicable, c'est le placement des éléments les uns par rapport aux autres dans un espace mesurable qui fait sens. Les éléments perceptifs saillants deviennent les sources d'un questionnement [5] qui n'est pas contraint par un sens de lecture. Les ressources s'individualisent par leurs différences de position. Cette spatialisation beaucoup plus libre pour la lecture et l'interprétation est conforme à une vision du web locale et non plus globale en laissant apparaître les singularités de cet espace. Elle correspond aussi à une évolution de l'utilisation perceptive des graphes en la dotant d'une métrique adaptée suivant la circonstance au caractère que l'on souhaite observer. Les cartes sont aussi beaucoup moins manipulatoires que les graphes et permettent assez peut d'interactions, ce pour respecter le critère synoptique et la métrique qui doivent être stables pour la lecture. Une carte est la finalisation d'un processus d'exploration visuel et gagne en globalité et stabilité ce qu'elle perd en souplesse et liberté d'interprétation [6]. En cela, elle se destine à un public d'utilisateurs finaux qui reçoivent ainsi une aide ainsi à la navigation, limitant du même coup leur désorientation.

3. Tensions entre les notions de localité et de hiérarchie lors de l'observation des formes du web

3.1. Méthodo et analyse.

Pour illustrer les différences fondamentales entre la liste et la carte, nous avons décidé de reconstruire le graphe des sites renvoyé par Google à une requête donnée. La thématique choisie pour notre requête a été l'avortement : 'abortion'. Nous avons choisi une thématique pour laquelle les sites renvoyés présentaient des points de vue différents. Nous nous sommes limités aux cinquante premières pages web retournées par Google, que nous avons catégorisées en trois sous groupes : top3, top10 et top50. Cette catégorisation découle du comportement des internautes face aux résultats des moteurs de recherche : les internautes cliquent rarement sur les résultats suivant les trois premiers et dépassent encore plus rarement la première page des 10 premiers résultats. Nous avons aussi catégorisé nos sites (les 50 pages appartenaient à 40 sites différents) selon le positionnement explicite de leurs auteurs par rapport à l'avortement : 5 catégories ont 'émergé' de ce travail :

- prolife : 6 sites contre le droit à l'avortement
- prochoice : 13 sites en faveur du droit à l'avortement
- neutral : 14 sites ne prenant pas position sur la question de l'avortement mais traitant du sujet
- catholic : 2 sites se présentant avant tout comme site catholiques et traitant de l'avortement
- after abortion : 5 sites proposant des ressources pour passer le moment de l'après avortement

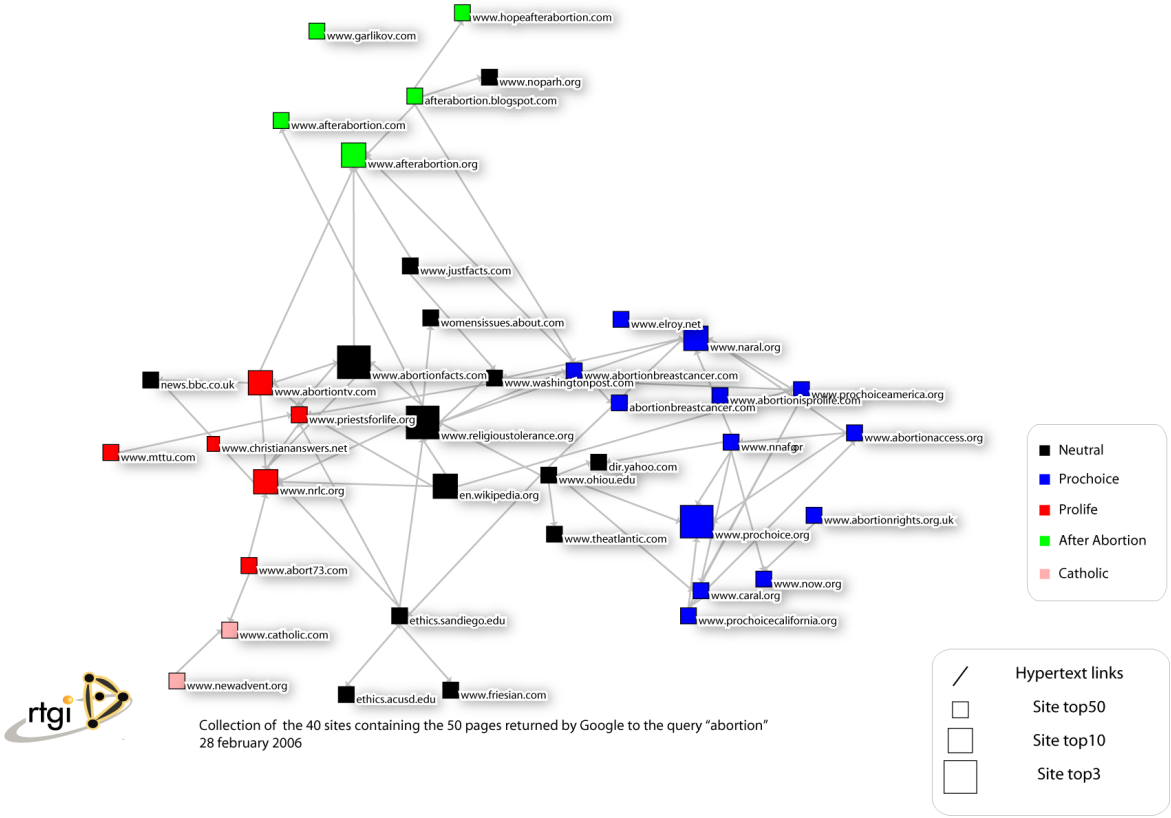

Figure 2 Collection of 40 sites containing the 50 pages returned by Google to the query "abortion", 28 february 2006

Le top3 de google propose au regard de ce classement, deux sites neutres et un site prochoice. Le top10 est complété par 1 site prochoice, 2 sites prolife, 1 site neutre et un site after abortion. Pour produire la carte de ces 40 sites, nous avons eu recours à un robot crawler capable de reconstituer la structure hypertextuelle d'un ensemble de pages. Pour chaque site nous lui avons fait visiter la page d'accueil, la page de liens, ainsi que toutes les pages traitant de l'avortement ceci pour ne pas avoir à explorer l'intégralité de sites non exclusivement dédiés à l'avortement (www.washingtonpost.com par exemple). A partir des données collectées par le crawler nous avons pu construire la carte présentée ici. Le placement des noeuds est obtenu par un algorithme qui positionne en proximité les noeuds partageant des liens.

Le caractère synoptique de la visualisation permet immédiatement d'appréhender les forces en présence. Alors que comme nous l'avons vu la structure de la liste rassemblait toutes les ressources sous une même catégorie, les distances permettent ici aux ressources de se distinguer les unes des autres, www.abortionaccess.org est intuitivement mis en opposition avec par exemple www.abort73.com.

Sur la carte, le positionnement des différents clusters prend le pas sur les critères de hiérarchie, ainsi le lecteur de la carte sera plus susceptible d'aller visiter des ressources dans chaque zone géographique dessinée par la cartographie qu'uniquement les ressources globalement les mieux classées.

Au delà de ces lectures intuitives de la carte, ce type de représentations qui révèlent la structure hypertextuelle véritable des sites présents sur la carte permet des lectures expertes qui se fondent sur la considération du lien hypertexte comme manifestation singulière d'un lien social *électif*. Ainsi on peut remarquer la présence des deux sites catholiques en proximité des prolife alors que ces sites ne se positionnent pas explicitement contre l'avortement, ce placement s'il est constaté par l'usager produira un effet de contexte qui influencera l'interprétation du site au moment de sa visite. Cas beaucoup plus intéressant et permettant de s'interroger sur l'importance du contexte (contexte hypertextuel dans ce cas), le site www.abortionfacts.com. Ce site se dit être un site neutre, raison pour laquelle il est noir sur la carte. Maintenant une lecture en précision de la carte nous apprend les choses suivantes : - son positionnement est frontière entre la zone des sites neutres et la zone des sites prolife. - il ne possède aucun lien sortant vers des sites neutres - il possède trois liens sortant vers des sites prolife Si ce site n'est pas situé dans la zone des prolife, c'est uniquement par la présence d'un lien de www.religioustolerance.org pointant vers lui, ce lien est en quelque sorte sa seule 'attache' au territoire des sites neutres, ses autres attaches étant prolife et pour une after abortion.

3.2. Les hiérarchies du web peuvent être repensées par rapport à la notion de localité

3.2.1. Une hiérarchie est à la fois l'expression d'un critère plus ou moins complexe et une façon fondamentale de restituer l'information.

D'un point de vue technique, un moyen fondamental de traiter une grande masse de données est de se doter d'une mesure, c'est-à-dire d'attribuer des scores aux données. Le simple fait de classer les données en fonction d'une mesure produit une hiérarchie, et pour cette raison hiérarchiser l'information est et restera une nécessité pour restituer des données. Cependant, il y a une grande différence entre l'utilisation des hiérarchies par un expert du web et par le grand public. L'expert peut croiser plusieurs mesures du web pour l'analyser, et dans ce cas il s'intéresse aux ressources dans leur ensemble : c'est l'expression d'un critère qui est l'aspect important de la hiérarchie. Au contraire, le grand public n'utilise en général qu'un seul critère et ne s'intéresse qu'aux données qui sont en haut de la hiérarchie. C'est le cas des moteurs de recherche, mais aussi de tous les « top 10 » que l'on trouve dans les sites contenant beaucoup de contenu (meilleures ventes, articles les plus consultés, les plus récents'). Dans ce cas, c'est la représentation en liste qui est l'aspect le plus important de la hiérarchie.

3.2.2. Les hiérarchies de la connectivité dépendent de sous-graphes

Quand on n'a qu'une partie d'un graphe, on ne peut pas savoir si les noeuds fortement connectés localement le sont aussi globalement. Il est assez simple de le comprendre. Prenons un graphe G non-orienté et intéressons-nous au degree des noeuds : le degree d d'un noeud v appartenant au graphe G=(V,E) et noté d(v—G) est le nombre de liens du noeud v dans le graphe G. Soient maintenant le graphe G et son sous-graphe G1. Nous nous intéressons à la relation R suivante :

$$\forall n, n' \subset G1\ (d(n)^{G1} \geq d(n')^{G1}) \Rightarrow (d(n)^{G} \geq d(n')^{G})$$

Cette relation exprime le fait que si un noeuds est fortement connecté dans le sous-graphe alors il l'est bien aussi dans le graphe complet (et de même pour les noeuds peu connectés). En d'autres termes, hiérarchiser le sous-graphe G1 hiérarchise bien une partie de G. Lorsqu'on calcule une hiérarchie sur un sous-graphe du web, on cherche à savoir si cette hiérarchie est bien celle du web dans son entier. Malheureusement, la relation n'est pas vraie dans le cas général, elle dépend de G et de G1. Voici deux cas différents : Subgraph G1 has not the same hierarchy as G, while subgraph G2 has the same. Avec un graphe simple tel qu'on en rencontre sur de tout petits crawls, on peut cependant déterminer des localités dont la hiérarchie est représentative.

The first hierarchy represents the global graph, while the second doesn ?t. It is sometimes possible to determine representative subgraphs. Les développements récents de la théorie des réseaux à invariance d'échelle démontrent que même dans le cas d'une arborescence à invariance d'échelle, les "branches" de l'arbre vérifient cette propriété. [Towards a theory of Scale-Free Graphs : Definition, Properties, and Implications (Extended Version), L. Li, D. Alderson, R. Tanaka, J. C. Doyle, W. Willinger, 2005,]

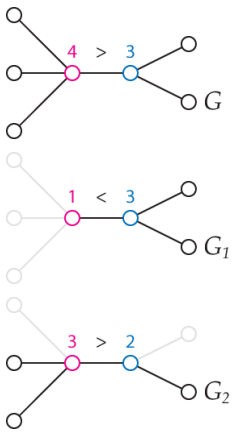

Figure 3 Subgraph G1 has not the same hierarchy as G, while subgraph G2 has the same.

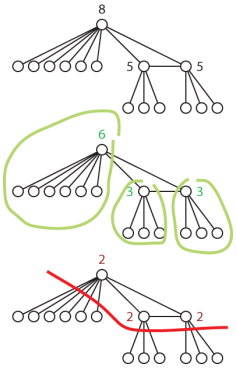

Figure 4The first hierarchy represents the whole graph, while the second doesn't. It is sometimes possible to determine subgraphs with representative hierarchies.

3.2.3. La stratégie de hiérarchisation des moteurs de recherche est empirique

D'après la théorie des scale-free networks, seuls les réseaux proches d'une arborescence à invariance d'échelle ont une hiérarchie globale de la connectivité. Autrement dit, dans un sous-graphe quelconque du web, la connectivité locale n'est pas représentative de la connectivité globale. L. Li et al. dénoncent un malentendu à propos des scale-free networks :

The Internet remains one of the most popular and highly cited application areas where power laws in network connectivity have 'emerged spontaneously', and the notion that this increasingly important information infrastructure exhibits a signature of self-organizing complex systems has generated considerable motivation and enthusiasm for scale-free networks. However, as we will show here, this basic observation is highly questionable, and at worst is the simple result of errors emanating from the misinterpretation of available measurements and/or their naive and inappropriate statistical analysis

Les algorithmes calculant le web comme un scale-free network peuvent donc s'avérer inexacts. En particulier, la croyance selon laquelle sur le web le global et le local se ressemblent grâce à des "motifs" [Network Motifs : Simple Building Blocks of Complex Networks, R. Milo, S. Shen-Orr, S. Itzkovitz, N. Kashtan, D. Chklovskii, U. Alon, Science, 2002,

http ://www.weizmann.ac.il/mcb/UriAlon/Papers/networkMotifs/networkMotifs.pdf]
est fausse. L. Li et al utilisent la fonction s(g) qui

measures the extent to which the graph g has a 'hub-like' core and is maximized when high-degree nodes are connected to other high-degree nodes

pour évaluer la self-similarité d'un graphe :

it was shown that many important technological and biological networks were self-dissimilar, in the sense coarse-grained counterparts display very different motifs at each level of abstraction. Our notion of motif self-similarity is much simpler, but consistent, in that the Internet has extremely low s(g) and thus minimally self-similar at the motif level.

Les algorithmes des moteurs calculent leurs hiérarchies sur une partie seulement du web, même si elle est de grande envergure. Mais rien ne permet de dire que ces scores correspondent à une loi globale (puisque leurs répertoires sont incomplets), ni à une loi locale (toujours à cause de la self-dissimilarité du web). On ne peut donc pas dire que les algorithmes des moteurs, comme le PageRank, se justifient par des propriétés topologiques du web. Leur stratégie d'indexation et de hiérarchisation est essentiellement empirique et vise surtout à satisfaire l'internaute.

3.2.4. Sur le web, les hiérarchies de la connectivité dépendent de localités

En l'absence de tout cadre permettant d'évaluer la valeur globale d'un critère hiérarchique dans un sous-graphe du web, nous cherchons des critères locaux permettant une validation locale. Dans cette optique, nous faisons varier les frontières de localités d'après plusieurs paramètres, et nous dégageons une liste de ressources classées par pertinence. Comme nous l'avons vu dans l'exemple ci-dessus (expérimentation), le choix d'une frontière influe très fortement sur la hiérarchie de ressources. Non seulement les algorithmes calculant la connectivité sont sensibles à ces manipulations, mais la notion même de pertinence peut se reconfigurer lorsque l'on change la définition d'une localité. La variabilité conjointe des algorithmes et de l'analyse manuelle nous incite à penser que le principe même de hiérarchie ne s'applique que localement. Des hiérarchies locales existent et les hiérarchies globales des moteurs ne les décrivent pas. Les grandes différences observées entre différents sous-graphes du web, extraits de la même façon mais à partir de ressources différentes, renforce l'idée qu'il n'existe pas de critère générique pour hiérarchiser les pages web. Cela étant, des hiérarchies locales peuvent émerger du web. Dans le domaine francophone de la culture scientifique, technique et industrielle, les quatre grandes institutions nationales jouent un rôle central sur le web comme dans la réalité. Non seulement elles disposent de sites riches en contenu, mais elles sont également d'un degré de généralité supérieur aux autres sites, et elles ont les plus forts scores d'autorité dans le corpus connexe des ressources sélectionnées manuellement. Nos expérimentations nous ont montré que dans un kMotifs/networkMotifs.pdf] plusieurs critères convergent vers la circonscription d'un domaine. Nous retrouvons ainsi le cas de figure de la théorie des agrégats [R. A. Botafogo, B. Shneiderman, Identifying Aggregates in Hypertext Structures, 1991] où une thématique et la connectivité convergent, même si nous préférons parler de "localité" plutôt que d'"agrégat" tant sont grandes les différences entre les cas observés. Ces hiérarchies locales ne sont pas restituées par les moteurs de recherche, le plus souvent à cause du fait qu'il n'existe pas de requête simple pour définir un domaine. Dans notre exemple, les grandes institutions ne sont pas porteuses du label "CSTI", ce qui ne les fait pas apparaître lorsqu'on interroge un moteur avec "CSTI" ou "Culture Scientifique Technique Industrielle". De manière générale, des termes différents décrivent un domaine dans sa généralité et dans ses spécialités, quand bien même sur le web les ressources sont très liées et traitent de la même thématique. Les localités thématiques imposent leurs propres hiérarchies, qui sont incompatibles avec une hiérarchie globale.

4. le contexte

Toute recherche d'informations de qualité nécessite de pouvoir vérifier la pertinence des ressources que l'on mobilise. Ce travail va d'ordinaire de soi avec la connaissance du contexte des informations. Si il est relativement facile de trouver le contexte des ressources dans les médias traditionnels, c'est beaucoup plus difficile sur le web. Les raisons sont profondes et commence avec avec la définition de ce qu'est le contexte d'une ressource sur le web.

4.1. visions documentaire et vision thématique

Il existe deux visions possibles de la question du contexte sur le web qui corresponde à deux niveaux de granularités : un local et un général. Le niveau local est le niveau du document lui même, de la ressource proposée. Dans le cas des moteurs de recherche, c'est en règle générale une page internet. Cette page existe dans un espace, en relation avec d'autres pages qui partagent entre elles des liens hypertextuels. Ces liens participent pleinement du document comme des pages d'un même livre. Le problème est alors de trouver où le document s'arrête car on peut suivre les liens indéfiniment. La question de la clôture du document est cruciale pour arrêter un jugement[1]. considérons l'exemple du site [2] qui au premier abord si l'on ne considère que sa page d'accueil est un site neutre en ce qui concerne l'avortement. Par contre si l'on considère l'ensemble de ses liens et en particulier leur sponsor ou leur mission, alors il devient un site prolife ce qui change la façon dont le lecteur va lire les pages. Une façon de traiter le problème de la clôture peut être de considérer le site comme un document au lieu de la page mais rien ne garantie que cela suffise. (lien avec modèle globale page et local site -¿ paradoxe conclusion) L'un des seuls moyen est une analyse topologique et sémantique des

pages. Si la page est trop éloigné tant topologiquement que thématiquement de la page initial alors on peut considérer que l'on atteint une frontière du document. En procédant ainsi de toutes parts il est alors possible de clôturer le document pour en faire l'analyse. Toutefois ce problème local du document se retrouve au niveau globale dans les ensembles de documents. Le contexte devient ce dans quoi la ressource web s'inscrit en terme thématique (sémantique) mais aussi topologique. Il est important de savoir dans quoi s'inscrit un document pour comprendre les intentions initiales de l'auteur. Une démarche classique herméneutique d'analyse du sens d'un document, et c'est ce qui nous préoccupe car c'est là le critère numéro 1 de pertinence fait appel aux écrits contemporain et dans le même domaine. Il en est de même sur le web mais si l'on pourrait penser que ce travail est d'autant plus facile que les documents sont liés entre eux, l'absence de datation des ressources qui est indispensable à ce travail herméneutique rend impossible la compréhension de comment ces liens sont apparus. L'archéologie du web est impossible compte tenu de sa plasticité et dynamicité. Dès lors le contexte devient une notion très flou d'entourage à la fois sémantique et topologique qui délimite une localité organisé par des acteurs/ auteurs. Ces deux définitions renvoient à deux apports de de la connaissance du contexte sur le web.

4.2. Apport de la connaissance du contexte sur le web

Non content d'aider à la navigation [3], le contexte est surtout important pour aider le lecteur à déterminer la pertinence d'une ressource et de l'information qu'elle véhicule ou d'un ensemble de ressources. Connaître le contexte d'une ressource sur le web se rapproche de la démarche herméneutique d'analyse du document avec les limites et les spécificités de cette question sur le web. Le document est-il un hub ou une autorité, un pont entre deux communautés, vers qui pointe-t-il, etc. Dans le cas de l'avortement par exemple le site [2] dont nous avons parlé plus haut, il est intéressant de remarquer qu'il possède des liens bidirectionnel avec des sites prolife et aucun lien vers des sites prochoice. De même au niveau globale connaître l'organisation thématique dans laquelle la ressource s'intègre pour en déduire les faits remarquables. Dans cet autre cas on veut avoir une vue d'ensemble d'un sujet précis pour en tirer des informations. Ce qui nous intéresse alors n'est pas tant d'avoir des résultats précis mais de savoir ce qui est saillant ou remarquable dans la thématique ou l'ensemble de sites/pages proposées. Trouver un 10 sites prolife contre 1 seul prochoice amène à repenser l'importance de ce dernier. Rendre ces saillances perceptibles permet au lecteur ou naviguant de prendre en compte le contexte et donc augmenté la pertinence et l'efficacité de cette navigation.

4.3. La liste et ses alternatives

Les moteurs de recherche rendent traditionnellement leurs résultats sous forme de liste. La liste est une structure graphique qui présente sur une même espace un ensemble de ressources. Or une présentation graphique dans un même espace est une mise en contexte de facto. Les ressources co-présentes dans cet espace sont un pool de ressources de ce que le moteur de recherche a jugé les plus pertinentes en réponse à une requête donnée. La mise en contexte résulte alors de la recherche du commun et de la nécessité de faire sens avec cet espace. En effet, si le moteur de recherche était sûr de pouvoir répondre à la requête de manière pertinente, il ne proposerait pas une liste de résultats et ne préciserait pas non plus qu'il a trouvé plusieurs milliers de résultats correspondant à la requête. Au lieu de ça, c'est au lecteur de trouver ce qui lui l'intéresse dans le paquet de résultats classé de manière explicite avec des numéros [5] ou non et ce parmi un nombre très important de pages toujours signalé dans la page. Il doit donc faire émerger des saillances de cette liste et en contextualiser les résultats ce qui est tout sauf évident dans le mesure où ces listes ne sont pas conçues à cette fin. leur lecture analytique (une ressource après l'autre) ne permet pas de voir comment les pages s'articulent. La présentation des catégories des pages dans les listes de résultat est un pas dans ce sens mais le manque de vue d'ensemble réduit considérablement l'efficacité de la mise en contexte. un autre problème est l'affichage page par page. Les résultats sont présentés 10 par 10 et forme ainsi des groupes et pas une seule liste ordonnée alors que c'est ce que le moteur de recherche renvoie. Ce mode de présentation peut être dépassé en se dotant de dispositif adéquat synoptique qui donne une vue d'ensemble du sujet comme c'est le cas avec une carte par exemple que l'on a pris soin de calculer pour qu'elle situe les ressources les unes par rapport aux autres. Dans le cas de la carte des blogopoles par exemple (insertion de la figure !), on distingue tout de suite les différents groupes politiques et leur importance relative sur le web ainsi que les ressources les plus importantes. L'information est contextualisée. Cela met aussi l'accent sur des portions d"espace précise, des localités thématique et leur organisation là ou la liste était globalisante et ne permettait de rendre compte des localités et de leur organisation.

5. conclusion

Juste pour ne pas zapper la référence : L'indexation des données repose sur un modèle du web. Dans le cas de la plupart des moteurs de recherche, les pages web sont indexées selon une certain ordre de priorité. Comme nous l'avons vu, les algorithmes des moteurs de recherche ne reposent plus sur l'exploitation des formes du web, mais sur du comportement des usagers. La raison en est que ces outils visent uniquement à restituer suffisamment de ressources pertinentes. [Crawling the Infinite Web, R. Baeza-Yates, C. Castillo, 2005, http ://citeseer.ist.psu.edu/723270.html]

les modèles et les SIWs, qu'est ce qu'un SIW, localité hierarchie locale, contexte locale, modele local contre les modèle globaux qui ne sont pas des SIW mais une source d'information

6. References

[1] J. M. Kleinberg, *"Authoritative sources in a hyperlinked environment,"* Journal of the ACM, Volume 46 Number 5, p.604-632, 1999.

[2] S. Brin, L. Page, *"The anatomy of a large-scale hypertextual Web search engine,"* In Ashman and Thistlewaite [2], p. 107-117. Brisbane, Australia, 1998.

[3] T. H. Haveliwala, S. D. Kamvar, G. Jeh, *"An Analytical Comparison of Approaches to Personalizing PageRank,"* Preprint, June, 2003.

[4] S. D. Kamvar, T. H. Haveliwala, C. D. Manning, G. H. Golub, *"Exploiting the Block Structure of the Web for Computing PageRank,"* Preprint, March, 2003, http ://www.stanford.edu/ sdkamvar/papers/blockrank.pdf

[5] G. Jeh, J. Widom, *"Scaling personalized web search,"* in Proceedings of the Twelfth International World Wide Web Conference, 2003.

bibitemHaveliwala02 T. H. Haveliwala, *"Topic-sensitive PageRank,"* in Proceedings of the Eleventh International World Wide Web Conference, 2002.

[6] F. Ghitalla, C. Lenay, *"Les Territoires de l'Information : Navigation et Construction des Espaces de Compréhension sur le Web,"* in *La Navigation, Les Cahiers du Numérique*, Paris : Hermès Editions, 2003.

[7] Dillon, Vaughan, *"It's the journey and the destination : Shape and the emergent property of genre in digital documents,"* in *New Review of Multimedia and Hypermedia*, 3, p. 91-106, 1997.

[8] visual complexity http ://www.visualcomplexity.com/vc/index.cfm ?domain=World

[9] M. Dodge, R. Kitchin, *"Atlas of Cyberspace,"* England : Addison Wesley, Harlow, 2001.

[10] S.K. Card, J.D. Mackinlay, B. Shneiderman, *"Information visualization,"* in *Readings in Information Visualization : Using Vision to Think,* San Francisco, CA : Morgan Kaufmann, pp. 1-34, 1999.

[11] J. Bertin, *"Semiology of Graphics : Diagrams, Networks, Maps,"* Madison/Wisconsin : University of Wisconsin Press, 1983.

[12] L. Li, D. Alderson, R. Tanaka, J. C. Doyle, W. Willinger, *"Towards a theory of Scale-Free Graphs : Definition, Properties, and Implications,"* (Extended Version), arXiv :cond-mat/0501169, 2005.

[13] R. Milo and al., *"Network motifs : simple buildingblocks of complex networks,"* Science 298, p.824-827, 2002.

[14] R. Botafogo, B. Shneiderman, *"Identifying aggregates in hypertext structures,"* in Proceedings of *Hypertext'91*, p. 63-74, New York :ACM, December 1991.

[15] F. Pfaënder, M. Jacomy, *"Explorer et appréhender l'internet,"* in *13èmes journées de Rochebrune : Rencontres interdisciplinaires sur les systèmes complexes naturels et artificiels,* ENST 2006 S001, 2006.

[16] P. Speroni, *"On Tag Clouds, Metric, Tag Sets and Power Laws",* May 2005, http ://blog.pietrosperoni.it/2005/05/25/tag-clouds-metric/

[17] T. Hammond, T. Hannay, B. Lund, J. Scott, *"Social Bookmarking Tools (I) A General Review,"* D-Lib Magazine, Volume 11 Number 4, April 2005.

[18] B. Bachimont, *"Arts et Sciences du numérique : ingénierie des connaissances et critique de la raison computationnelle,"* Mémoire d'Habilitation à Diriger les Recherches, Université de Technologie de Compiègne, 2004.

[19] F. Ghitalla, A. Le Berre, M. Renault,*"Des documents, des liens et des acteurs,"* in proceedings of H2PTM conference, 2005.