4e Tic & Territoire : quels développements ?
île Rousse 2005
Journée sur les systèmes d'information élaborée

CONSTRUCTION DU RESEAU D'INTERACTION ENTRE SITES WEB : TEST DE ROBUSTESSE DE LA METHODE A PARTIR DE PLUSIEURS SOURCES D'INFORMATION

Eric Boutin(*), Guillaume Perrin(**)
boutin@univ-tln.fr, guillaumeperrin@yahoo.com

(*) : Maître de Conférences laboratoire I3M, IUT de Toulon BP 132 83957 La Garde Cedex
(**) : Doctorant laboratoire I3M, Université du Sud Toulon Var BP 132 83957 La Garde Cedex

---

Mots clés : cybermétrie, analyse relationnelle, intelligence territoriale, liens hypertextes

Résumé

Lorsqu'on construit des indicateurs cybermétriques, on utilise parfois les fonctionnalités avancées des moteurs de recherche en faisant faire à ces derniers des opérations pour lesquelles ils n'ont pas été préparés. Ainsi, nous nous sommes intéressés à la représentation des interactions entre les acteurs du web public en région paca. Ce travail nous servira de fil rouge et de validation expérimentale dans cette communication. Pour construire le réseau des liens hypertextes entre les sites à étudier, on a mobilisé deux méthodes alternatives :

- Dans un premier temps, on a utilisé le moteur de recherche Google et sa commande avancée link:. La commande link permet de connaître les liens entrants sur une page ou un site web donné. On obtient à l'issue d'un processus de collecte, de traitement et de cartographie le réseau des interactions entre les sites publics de la région.
- Pour mesurer la qualité de ce résultat, nous l'avons confronté à une autre méthode de collecte de données. Il s'agissait alors d'explorer les diverses pages d'un site web à la recherche des liens que ce site a vers des sites extérieurs. Cette nouvelle collecte d'information s'intéresse aux liens sortants et non plus aux liens entrants et s'affranchit du moteur de recherche Google. Cette méthode s'est révélée à l'usage beaucoup plus riche, l'utilisation de la commande link du moteur de recherche Google rendant compte de façon très incomplète de la réalité.

L'objectif de cette communication est de confronter les résultats de ces deux méthodes. Cette communication débouche sur l'explicitation des hypothèses sous jacentes du moteur Google. Elle montre aussi tout le risque qu'il y a, pour une personne non spécialiste, à utiliser les moteurs de recherche en les détournant de leur fonction initiale. Elle renforce également l'asymétrie du contexte, l'utilisateur lambda n'ayant accès qu'à une information biaisée et limitée qu'il ne pourra exploiter que de façon très prudente dans une perspective d'analyse cybermétrique.

1. Présentation du contexte

1.1. Les analyses cybermétriques

Les analyses cybermétriques ou wébométriques ont pour objectif, à partir d'un corpus web, de représenter des indicateurs qualifiant ce corpus. Ces analyses sont transposées des domaines bibliométriques et scientométriques (Rostaing et al., 1999). Pour ce faire ces analyses mobilisent des outils de collecte, de validation et de traitement qui constituent la chaîne de traitement de l'information. Lorsqu'on réalise un travail cybermétrique, il s'agit de collecter en amont une information massive sans restriction, la plus exhaustive possible. Les techniques du datamining (Fayyad et al., 1996) permettront alors de réduire cette complexité en faisant apparaître plusieurs facettes de la réalité correspondant à autant d'axes qui sont privilégiés. Grâce aux techniques de datamining, l'utilisateur devient acteur des filtres qu'il va pouvoir activer pour révéler tel ou tel aspect qu'il cherche à décrire.
On apprend dans les livres que la valeur de l'information obtenue à l'issue d'un processus de traitement est celle du maillon le plus faible de la chaîne. Toutefois, mobilisons nous toujours notre esprit critique pour valider la pertinence de l'information collectée à chacune des étapes de la chaîne ? Cette communication a pour objet de réaliser un focus sur une source d'information privilégiée dans des processus de veille sectorielle et considérée comme facilement disponible et la plus exhaustive sur un sujet. Nous nous attacherons à montrer les biais intrinsèques d'une telle information et à voir en quoi sa prise en compte sans réserve est de nature à altérer profondément la pertinence des résultats de l'analyse.

1.2. L'analyse relationnelle dans le contexte de l'intelligence territoriale

La validation expérimentale qui servira de fil rouge à cette communication s'inscrit dans la problématique générale de la construction d'indicateurs relationnels permettant de décrire des corpus web. L'analyse relationnelle (Wasserman et Faust, 1994) (Degenne, Forse, 1994) est utilisée très concrètement ici dans une problématique d'intelligence territoriale (Bertacchini, 2002). L'objectif est de reconstituer le maillage territorial c'est-à-dire de révéler les interconnexions existant entre des acteurs d'un territoire, chacun des acteurs ayant ses ressorts territoriaux ou thématiques. Ce projet de recherche porte plus précisément sur la mesure du web public régional en région Paca. Nous avons pour cela identifié dans un premier temps 440 sites web publics régionaux en région Paca. Chacun de ces sites web appartient à un échelon territorial spécifique (web communal, intercommunal, départemental ou régional). Cet ensemble de 440 sites web étant défini, nous avons cherché à identifier les interactions existantes entre ces sites web. Par interactions, nous entendons l'identification des liens hypertextuels existants entre ces sites. Ces liens hypertextuels ont une signification bien particulière dans le contexte non marchand que nous étudions ici. Cette signification des liens hypertextes est directement transposée de l'analyse de la citation (Egghe, 2000). Ils signifient la recommandation, la légitimation d'un site par un autre dans l'esprit des débuts du web. A travers les interactions entre ces sites web, il s'agit de cerner les systèmes de reconnaissance et de légitimité implicites existant entre des sites web d'un territoire donné.

Pour intéressante qu'elle soit, cette problématique n'en est pas moins délicate à mettre en œuvre. En effet, de par son caractère plutôt émergent, ce type d'analyse ne peut pas être réalisée en utilisant un outil intégré du marché. Sauf à considérer des logiciels coûteux à la portée seulement de grands comptes, il n'existe pas de logiciels de type boite noire qui puisse reconstituer le réseau des interactions entre sites d'un domaine particulier. Pour cette raison, le chercheur dans ce domaine est obligé de juxtaposer des ressources disponibles sur internet (moteurs de recherche, reformateur, logiciel de traitement) sans disposer toujours des clés permettant de valider une telle information.

2. L'étape de collecte d'information

2.1. Les deux moyens de collecter l'information

Avant de calculer des indicateurs relationnels sur ce corpus, l'étape préalable consiste à identifier, de façon la plus exhaustive possible les liens hypertextuels entre les couples de sites étudiés dans ce travail. Cette étape de constitution du corpus peut être envisagée de deux manières principales complètement équivalentes :

2.1.1. Identifier les liens avals :

L'ensemble des interactions peut être construit en privilégiant l'étude des liens avals. Cette démarche vient la plus naturellement à l'esprit même si ce n'est pas la plus facile à mettre en œuvre sur le web. Il s'agit dans ce cas de considérer successivement les 440 sites web à analyser. Pour tous les sites web considérés successivement, on identifie toutes les pages du site à la recherche de liens hypertextes vers les 439 autres. Bien souvent, il faut disposer de crawler pour obtenir ce type d'information. Toutefois, il existe quelques logiciels libres qui, détournés de leurs fonctions primitives permettent de générer la liste des liens sortants d'un site donné. Ainsi le logiciel Xenu, connu des webmestres comme outil d'identification de liens cassés, peut être utilisé pour identifier les liens sortants d'un site. Ce logiciel fonctionne de façon assez simple. On fournit à Xenu en input une liste de sites web dans un fichier texte et on indique un niveau de profondeur. Le niveau de profondeur correspond au nombre de pages successives à partir de la page d'accueil, par lesquelles il faut passer pour arriver à une page déterminée. Ainsi le niveau de profondeur est égal au nombre de liens internes d'une page à une autre. Le logiciel explore alors les pages du site une à une à la recherche de liens sortants en se limitant au niveau de profondeur spécifié. Dans l'étude que nous avons menée, nous avons injecté dans Xenu, en plusieurs passes successives, la liste des pages d'accueil des 440 sites web à étudier. Pour éviter de tomber dans les puits que constituent certains sites contenant des forums très riches, nous nous sommes limités à une profondeur de 17 pour cette étude. Un reformatage des données dans un gestionnaire de base de données nous a permis de restreindre le corpus aux relations existantes entre les sites du corpus pris deux à deux. Xenu est donc un mini crawler paramétrable permettant de parcourir le web et de récupérer une information relationnelle pré-formatée.

2.1.2. Identifier les liens amonts

La seconde manière de constituer le corpus consiste à privilégier les liens amonts. Cette démarche n'est pas la plus naturelle mais c'est elle qui viendra la première à l'esprit de l'internaute averti. En effet, les moteurs de recherche majeurs disposent d'une commande avancée (link chez google ou chez Yahoo) permettant de connaître la liste des pages pointant sur une page ou un site donné. En systématisant l'utilisation de la commande link pour les 440 sites du corpus, il est donc possible de reconstituer rapidement les interactions entre les sites de notre corpus. L'information est ensuite reformatée selon des processus semi automatiques puis analysée en utilisant un outillage spécifique. Cette pratique revient à détourner les moteurs de recherche de leur fonction principale. En utilisant la commande avancée d'un moteur de recherche majeur, on accepte les limitations d'un tel moteur et en particulier la limite de son index. Toutefois, les progrès des moteurs dans leur couverture du web, leur plus grand focus sur l'identification sur les sites pertinents conduit bien souvent à se contenter de l'utilisation de la commande link d'un Google avec la meilleure conscience professionnelle du monde. De même, la rapidité des moteurs de recherche joue un grand rôle dans le choix de leur utilisation.

2.2. Validation expérimentale

La validation expérimentale que nous avons conduite mesure l'écart de résultat qu'il existe entre un corpus primaire constitué à partir de Xenu (identification des liens avals) et un corpus primaire constitué à partir de la commande link de Google (identification de liens amonts).
En réalisant ce travail, on pensait qu'il y aurait une différence peu significative entre les résultats des deux méthodes. Les différences de résultat se sont révélées très importantes :

- Xenu permet d'identifier 1128 liens hypertextes entre les 440 sites web étudiés
- la commande link de Google permet d'identifier 308 liens entre les 440 sites web
- 155 liens hypertextes sont communs aux deux méthodes
- 153 liens ne se retrouvent que dans le link de google
- 973 liens ne se retrouvent que dans la méthode de collecte des liens avals (Xenu).

La Figure 1 représente cette différence.

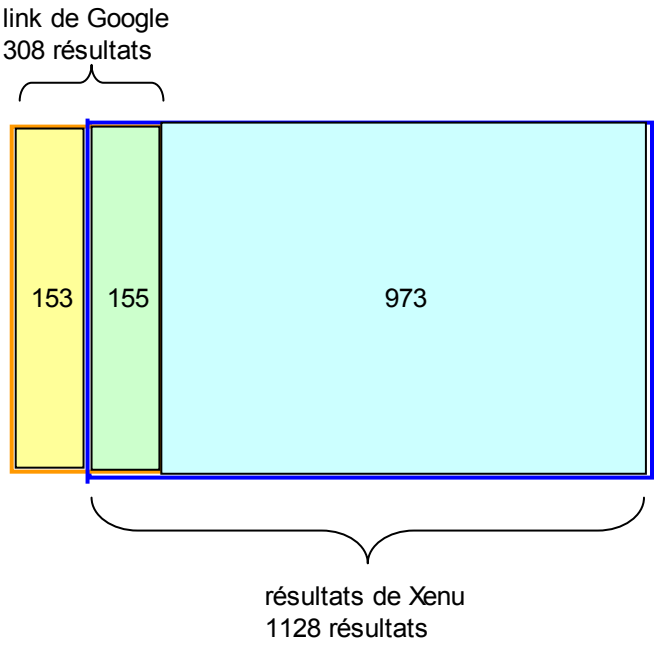

Figure 1 : Comparaisons des résultats obtenus sous Google et Xenu

Ces résultats sont troublants. Il existe un rapport de 1 à 3,66 entre le nombre de résultat de Google et celui de Xenu.

Ces chiffres sont encore plus marqués lorsqu'on s'intéresse aux interactions existantes entre les sites des communes de la région paca. Le tableau 1 montre qu'alors qu'en collectant l'information avec Xenu, on obtient 117 liens hypertextes entre les 220 communes de la région Paca qui ont un site web, on n'obtient aucun lien entre ces mêmes communes lorsqu'on privilégie la commande link de Google.

	Nbre de liens hypertextes entre les 220 communes de la région Paca
Xenu	117
Commande Link de Google	0

Tableau 1 : comparaison du nombre de liens hypertextuels entre les communes de Paca

Il est évident qu'un tel écart quantitatif entre la source de données va se traduire par des modifications très significatives dans l'interprétation des résultats. Avant d'analyser ces effets, il est intéressant de s'interroger sur les raisons d'une différence si marquante.

3. Raisons des différences observées :

3.1 La question de la taille de l'index de Google

La première raison qui vient à l'esprit est la faiblesse de l'index de google. Google est sans doute le moteur de recherche généraliste le plus complet. Toutefois, le fait d'utiliser la fonction link dans cette expérimentation revient à le comparer aux performances d'un moteur sectoriel. On découvre alors une certaine limitation, la taille de l'index de google se trouvant mise à défaut sur notre problématique. Nous avons cherché à mesurer l'effet « index ». Pour cela, nous avons fait un focus sur les 117 liens hypertextes entre les 220 communes de Paca qui avaient été identifié par Xenu et pas par Google. Dans 40% des 117 cas, on observe que la page d'accueil ou la page d'arrivée du lien n'est pas présente dans l'index de Google.
De ce point de vue, la recherche des liens hypertextes s'avère beaucoup plus exigeante que la recherche de pages web simples. En effet, il suffit que la page de départ ou que la page de destination ne figure pas dans l'index de Google pour que le lien ne soit pas trouvé. Le nombre de liens entre un ensemble de sites est fonction du carré du cardinal de cet ensemble. Cela signifie que si l'index de google est deux fois plus petit sur un sujet donné que celui d'un moteur de recherche thématique, cela se traduira par le fait que Google aura 4 fois moins de liens hypertextes dans ce domaine que l'autre moteur. La recherche des liens hypertextes joue donc le rôle d'accélérateur.
Si on analyse le rapport entre le nombre de liens entre les deux méthodes sur l'exemple retenu dans cette analyse, on a un rapport de 1 à 3,66 dans le nombre de liens hypertextes. Cela signifie que si on donne un index de 100 au moteur de Google on obtient un index de 191 au « moteur de Xenu ».

3.2 Vers d'autres explications :

L'interprétation précédente est assez satisfaisante. Par contre elle ne rend compte que de 40 % de la réalité. Les 60% des cas d'études analysés sont beaucoup plus troublants.
Dans tous ces cas, on observe que les pages de départ et d'arrivée du lien figurent dans l'index de Google. A partir du moment ou :

- « une page A » est dans l'index de Google,
- une page B est présente également dans l'index de google
- la page A comporte un lien hypertexte vers la page B
- le contenu de la page A en cache dans le moteur de recherche comporte aussi un lien vers la page B

Comment se fait il que la commande Link de Google récapitulant les liens entrants sur une page ne permette pas de restituer le lien entre A et B ?
Aucune explication n'a été fournie. On aurait pu incriminer certains formats de fichiers non reconnus par Google. Ce n'est pas le cas. On a pu observer que les liens non pris en compte étaient souvent situés sur des pages comportant de nombreux liens sortants.
A défaut d'explications convaincantes, on en est réduit à émettre un doute sur la pertinence du link de google.

4. Interprétation des résultats :

Il est évident qu'une telle différence dans la source de données se traduit par des interprétations très différentes. Nous avons souhaité mettre en évidence deux types de différences liés d'une part à la caractérisation du niveau hub et authorities (Ding et al. 2001) des sites et d'autre part à la représentation relationnelle des interactions existantes entre les sites communaux en région Paca.

4.1 Identification des hubs et authorities :

Nous allons comparer, dans ce paragraphe, les résultats des deux méthodes en terme de nombre de liens entrants et sortants sur chaque site. Les tableaux 2 et 3 reportent pour un site web donné, le nombre d'autres sites parmi les 440 pointant sur ce site web. Ces sites sont classés du plus au moins

fréquents et seuls les 10 premiers sites ont été repris. On obtient donc un classement des sites les plus authorities. Ces sites se caractérisent par le fait qu'ils sont souvent cités par les autres et donc légitimés par eux. Ce tableau a été calculé à partir du corpus récupéré sous Xenu (tableau 2) puis à partir du corpus récupéré à partir du link de Google (Tableau 3).

Identification des authorities d'après Xenu	
site arrivee	Nbre de liens pointants
www.cr-paca.fr	64
www.anpe.fr	40
www.cg83.fr	33
www.cg13.fr	29
www.cg06.fr	27
www.paca.pref.gouv.fr	25
www.ac-nice.fr	24
www.marseille-provence.cci.fr	19
www.var.pref.gouv.fr	18
www.mjspaca.jeunesse-sports.gouv.fr	17

Tableau 2 : Sites web les plus authorities - corpus récupéré à partir de Xenu

Identification des authorities d'après Google	
site arrivée	Nbre de liens pointants
www.cr-paca.fr	9
www.cg13.fr	6
www.cg06.fr	6
www.cg83.fr	6
www.lepilote.com	6
www.pacac.cci.fr	6
www.visitprovence.com	5
www.paca.drire.gouv.fr	5
www.paca.equipement.gouv.fr	4
www.paca.pref.gouv.fr	4

Tableau 3 : Sites web les plus authorities - corpus récupéré à partir de la commande link de Google

On observe dans le Top 10 des changements assez significatifs. On ne peut pas considérer que les résultats de Google constituent un échantillon représentatif de la réalité puisqu'on a uniquement un site web sur deux qui est commun aux deux classements. Cette observation est intéressante. On aurait pu penser que l'index de Google était composé des pages les plus pertinentes et que le classement des pages selon leur degré d'autorité allait être plus proche de celui de Xenu.

Les résultats sont tout aussi différents lorsqu'on s'intéresse au hit parade des sites en fonction de leur nombre de liens sortants (appelés hubs). On observe encore des différences très significatives entre les deux méthodes qui ne proposent que 40% des sites en commun. Les tableaux 4 et 5 présentent les résultats.

Identification des hubs d'après Xenu	
site depart	CompteDesite depart
www.mairie-marseille.fr	34
www.vaucluse.fr	33
www.cr-paca.fr	31
www.cg06.fr	28
www.alpes-de-haute-provence.pref.gouv.fr	25
www.mjspaca.jeunesse-sports.gouv.fr	24
www.cuges-les-pins.fr	24
www.mairie-le-cannet.fr	22
www.arpe-paca.org	21
www.crt-paca.fr	21

Tableau 4 : Sites web les plus hubs - corpus récupéré à partir de Xenu

Identification des hubs d'après Google	
site depart	CompteDesite depart
www.debatpublic-lgvpaca.org	38
www.crt-paca.fr	22
www.mediterranee-technologies.com	17
www.cr-paca.fr	17
www.vaucluse.fr	12
www.ac-nice.fr	8
www.paca.cnfpt.fr	8
www.agglo-paysdaix.fr	7
www.seillans-var.com	7
www.mairie-marseille.fr	6

Tableau 5 : Sites web les plus hubs - corpus récupéré à partir de la commande link de Google

4.2 Représentation des interactions entre sites communaux :

A partir du corpus constitué à partir de Xenu, il est possible de représenter la cartographie des interactions entre les 117 liens hypertextuels existants entre les 220 sites web des communes de la région Paca. Cette cartographie présentée figure 2, superposée à une carte du territoire, est assez éclairante :

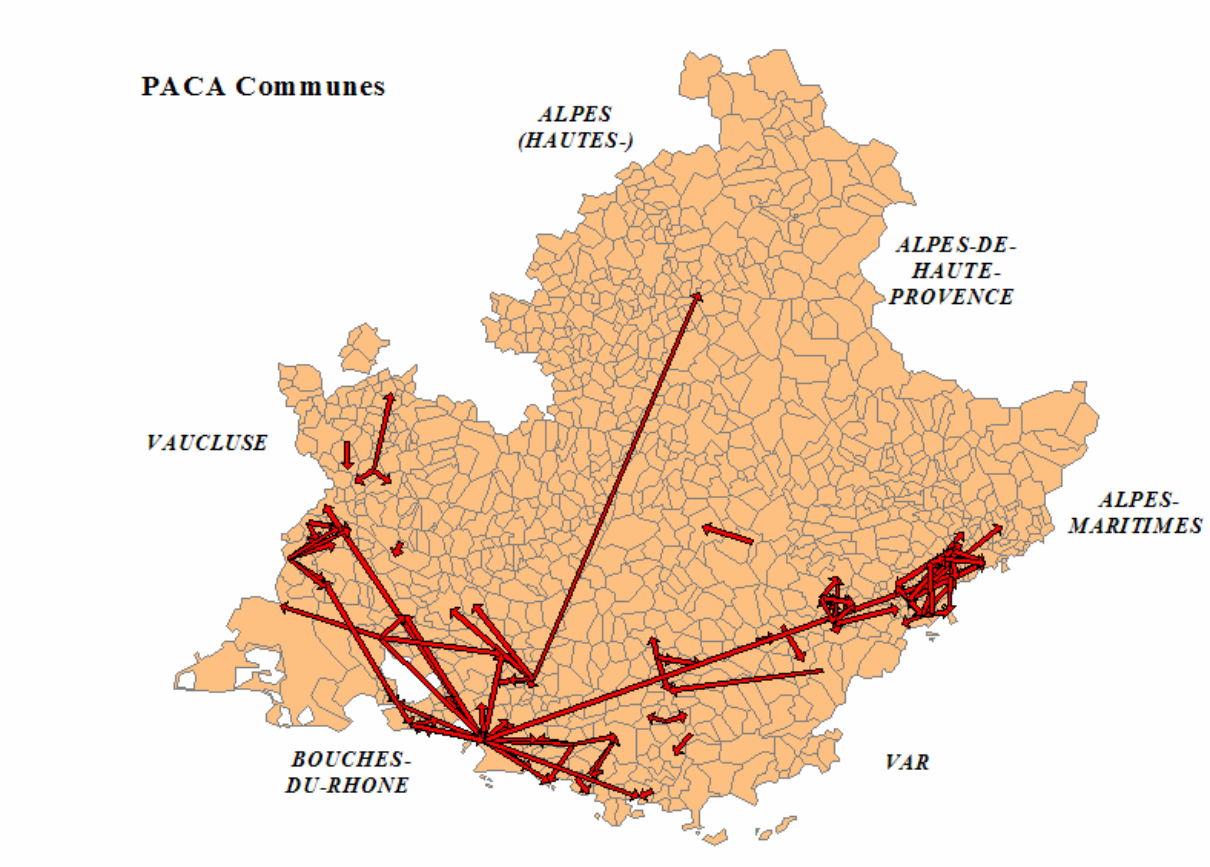

Figure 2 : Interactions entre sites web des communes de Paca – source d'information Xenu

- On observe que sur l'ensemble des liens hypertextes, plus de 95% sont le fait de relations hypertextes entre des communes du même département. La géographie physique du territoire est donc un outil adapté pour rendre compte des relations virtuelles entre sites web
- les interactions hypertextuelles entre sites web de communes concernent essentiellement des interactions des communes du littoral mais ceci est à mettre en parallèle avec la statistique des communes disposant de sites web qui sont particulièrement plus représentées dans la même frange littoral

Aucune de ces conclusions ne peut être valablement déduite du corpus résultant de la commande link de Google puisqu'il n'y a aucun lien hypertexte entre les 220 sites web communaux lorsque l'information est collectée à partir du corpus construit à partir de la commande Link de Google.

Conclusion :

Cet article a été détourné de son but premier. Il s'agissait comme l'indique le titre de tester la robustesse d'une méthode à partir de deux sources d'information différentes. Toutefois ce qui ressort de ce travail expérimental, c'est que ce n'est pas la robustesse de la méthode qui a été testée mais la robustesse des sources d'information qui a été mise en cause.
Un certain nombre de recommandations ressortent de cette communication sous forme de mise en garde :
- Les outils de recherche se situent dans une problématique volontairement opaque. En effet, la transparence de leur algorithme les rendrait beaucoup plus sensibles au spam de la part de webmestres soucieux de mettre leur site au premier plan. Dans ce contexte, il est imprudent d'utiliser ces outils en pensant savoir comment ils marchent.

- Dans cette expérimentation, nous avons mis en lumière tout le danger qu'il y avait à prendre un moteur généraliste pour en faire un outil de veille sectorielle. Les résultats issus de Xenu sont presque 4 fois plus nombreux que ceux renvoyés par la commande Google.
- De façon plus générale, il est important, à toutes les étapes de la chaîne de traitement de l'information, de verbaliser les hypothèses sous jacentes aux outils de collecte et de traitement car trop nombreux sont les systèmes boîte noire qui sortent un résultat biaisé en méconnaissant les algorithmes sous jacents.

Bibliographie :

BERTACCHINI (2002) *Territoires et Territorialités*. Vers l'intelligence territoriale, volet 1, 200 pages, Collection *les Ecrits des Technologies de l'Information et de la Communication*.

DEGENNE A. FORSE M., (1994), *Les réseaux sociaux*, Editions Armand Colin, 1994

DING C., ZHA H, HE X., HUSBANDS P. , SIMON H.(2001) "Link Analysis : hubs and authorities on the world wide web", LBNL

EGGHE L.(2000), « New informetrics aspects of the internet : some reflections, many problems », journal of information Science, 26(5) : 329-335

FAYYAD U., PIATETSKY-SHAPIRO G., SMYTH P, UTHURUSAMY R. (1996), Advances in knowledge Discovery and Data Mining, AAAI Press / Mit Press

ROSTAING H. , BOUTIN E., MANNINA B.(1999), "Evaluation of internet resources : bibliometric techniques applications". In cybermetrics 99, Colima

WASSERMAN S., FAUST K. (1994). *Social Network Analysis : Methods and Applications*. Cambridge, England, and New York : Cambridge University Press