CONSTRUCTION AUTOMATIQUE DE RESEAUX: UN OUTIL POUR MIEUX APPREHENDER L'INFORMATION PROVENANT DE L'INTERNET

Eric Boutin
Université de Toulon et du Var IUT TC
Laboratoire Le Pont BP 132
83957 LA GARDE CEDEX
E-mail : boutin@univ-tln.fr
Bruno Mannina, Hervé Rostaing, Luc Quoniam
CRRM Faculté de Saint Jérôme
13397 MARSEILLE CEDEX 20
E-mail : crrm@crrm.univ-mrs.fr

Résumé:
The purpose of this paper is to present tools and methodologies that can be used in order to extract strategic information from Internet on a specific subject. The result of the analysis is presented through networks maps and visualize interactions of the main WEB sites on the analysed subject. The paper focuzes on a specific case study that is representing interactions between the main Web sites of the french administration that were present on the site of the « Documentation Française » in September 1996.

Les entreprises sont à la recherche d'informations stratégiques. Le WEB apparaît, dans ces conditions, comme une source d'information à explorer [Dousset et alii]. Le problème qui va nous préoccuper ici est celui de la discrimination entre l'information stratégique et l'information qui ne l'est pas. L'utilisation des moteurs de recherche renvoie bien souvent à un nombre important de sites. L'utilisation de l'indicateur de pertinence, défini par certains moteurs de recherche, est un premier pas vers l'identification des sites pertinents sur un thème donné.[Ducloy et alii] Toutefois, ces indicateurs sont soumis à caution dans la mesure où l'indicateur de pertinence ne correspond pas à la valeur intrinsèque d'un site mais plutôt à l'habileté avec laquelle le concepteur a bâti son site pour faire en sorte qu'il soit référencé en bonne position sur un thème particulier.
Nous proposons d'introduire ici une dimension complémentaire pour juger de l'importance d'un site par rapport à un autre. Nous considérerons qu'un site A est important s'il existe un nombre jugé significatif d'autres sites qui possèdent un lien hypertexte en direction du site A. Une comparaison peut être fait pour rendre compte de ce nouveau mode d'évaluation de la pertinence d'un site. Il est possible de définir l'importance d'un article scientifique par le nombre d'autres articles qui le citent. Dans le domaine des publications scientifiques, le délai avant qu'un site soit cité peut être de plusieurs années. Sur Internet, l'échelle du temps est bousculée.
Nous avons choisi d'utiliser l'outil réseau pour fournir une cartographie des sites qui font partie d'un domaine donné. L'outil réseau permet de visualiser le caractère central ou périphérique d'un site par rapport aux autres.
Cette démarche originale a été appliquée à l'ensemble des sites publics français présents sur le WEB. 109 sites ont été recensé sur le serveur de la « Documentation française » en Septembre 1996. Lorsqu'on s'intéresse aux liens hypertextes qui partent de ces 109 sites, on obtient une information qui peut être retranscrite sous forme de cartographies, l'outil réseau proposant une grille d'interprétation intéressante.

1- Des données brutes de départ à une première démarche d'analyse:

1.1-Les données brutes.

La collecte d'information a été réalisée en utilisant l'agent intelligent Auresys développé dans le cadre d'un travail doctoral au CRRM [Mannina et alii]. Cet outil considère une liste de sites de départ (liste des 109 sites) et explore successivement les différents sites qui lui sont reliés par liens hypertextes : ces sites constituent l'ensemble d'arrivée. Pour chacun de ces 109 sites, nous avons établi la liste des sites qu'il était possible d'atteindre en moins de 3 clics de souris. Cette contrainte de profondeur de 3 s'explique par des raisons liées au temps de constitution de la base de travail par le moteur de recherche. En effet, reprenons le mécanisme par lequel est construite l'information brute qui associe à chacun des 109 sites de la liste ceux qui lui sont liés. Considérons un site lambda. Le moteur va simuler la connexion à la page d'accueil de ce site et se connecter successivement aux différents renvois hypertextes présents sur cette page. Chacun de ces renvois hypertextes correspond à une page qui sera examinée de la même manière.
Si on considère pour simplifier qu'en moyenne un site est relié à k autres, raisonner sur une profondeur de n au lieu d'une profondeur de n-1 prend k fois plus de temps.
Il est possible, grâce au logiciel Dataview, développé au CRRM,[Rostaing] de traduire cette information sous forme matricielle comme le montre le Tableau 1.

ensemble d'arrivée

	X	Y	Z	A
A	1	0	1	0
B	1	0	0	1

Il est possible de passer du site A au site Z en 3 clic de souris

ensemble de départ

Tableau 1: Un exemple de matrice non carrée et non symétrique

La relation « Il est possible de passer du site X au site Y en n clic de souris » est de type non symétrique: le fait que le site X possède un lien hypertexte vers le site Y ne préfigure en rien l'existence d'un lien hypertexte entre Y et X.

1.2 Méthode d'analyse :

Nous proposons de représenter cette réalité en utilisant le logiciel Matrisme développé au sein du laboratoire Le Pont dans le cadre d'une travail de doctorat [Boutin et alii 1995]. Cet outil prend pour point de départ une matrice carrée symétrique.
Le problème consiste donc mathématiquement à transformer une matrice non carrée et non symétrique en matrice carrée symétrique. Il existe plusieurs façons complémentaires de résoudre ce problème. Nous allons les présenter successivement. Un réseau étant composé de sommets et de liens, il s'agira donc successivement de définir quel sera l'ensemble des sommets étudiés et quel est le sens de la relation entre ces sommets:
- L'ensemble des sommets peut correspondre selon le cas:
    - à l'ensemble des 109 sites de l'ensemble de départ.
    - à une sélection de certains sites possédant certaines propriétés particulières.

- à l'ensemble des sites cités par les sites initiaux .

- Un lien entre deux sommets X et Y correspondra à une relation symétrique entre les deux sommets. On peut concevoir plusieurs types de relations symétriques entre X et Y:

- Il existe un lien hypertexte de niveau 3 entre X et Y ou entre Y et X.
- Il existe un lien hypertexte de niveau 3 entre X et Y et entre Y et X.
- Le lien entre X et Y peut représenter le nombre de sites identiques auxquels X et Y sont reliés.
- Le lien entre X et Y peut représenter le nombre de sites qui citent X et Y.

Le croisement de ces différentes analyses illustre la richesse des développements qui peuvent être conduits. Nous nous proposant de présenter successivement ces différentes approches en allant de la plus simple à la plus élaborée.

2. Présentation de quelques résultats.

2.1 Un réseau de départ inextricable

Le premier réseau qui vient à l'esprit consiste à représenter les relations entre les 109 sites Web retenus par le serveur de la « Documentation française ». Dans ce réseau, nous allons considérer que deux sites X et Y sont reliés par un lien s'il est possible de passer du site X au site Y ou du site Y au site X en moins de 3 clic de souris.
Le résultat, obtenu automatiquement sous le logiciel Matrisme est présenté figure 1.

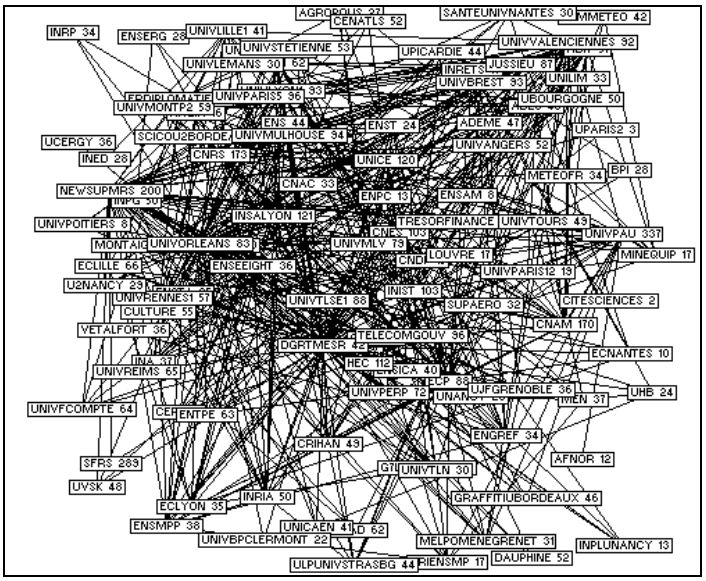
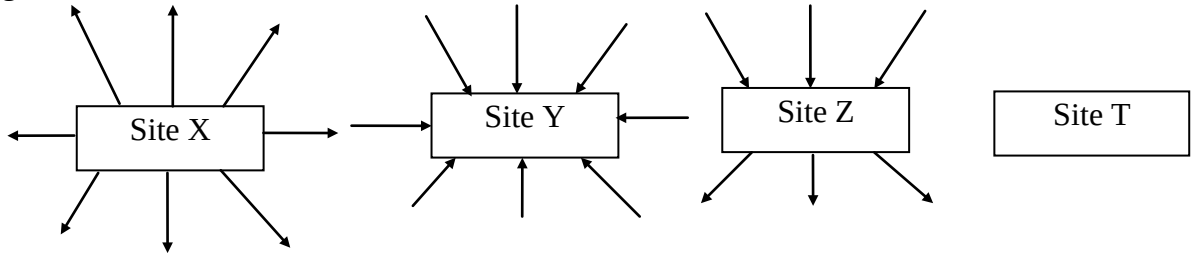


Figure 1: Le réseau brut de départ

Ce résultat correspond à un réseau dont la seule interprétation possible est très générale. Ce réseau exprime une forte inter-relation entre les différents sites le constituant. Il n'est pas possible de mener une analyse au niveau des sommets tant l'imbrication est forte. L'approche réseau n'apporte pas un gain en lisibilité significatif. Pour analyser ce réseau, il faut réaliser des opérations de filtrage qui ont pour objectif de supprimer certains liens et/ou certains sommets.

2.2 Le découpage des sites selon leur nature.

Outre son enchevêtrement, le réseau présenté précédemment est difficilement interprétable pour une autre raison. En effet, un arc relie le site X au site Y s'il existe un lien hypertexte entre X et Y **ou** entre Y et X. Il est donc impossible de connaître le site de départ et le site d'arrivée. De façon plus globale lorsqu'un site possède plusieurs arcs, ceci peut signifier indifféremment qu'il est particulièrement « ouvert » sur les autres sites ou particulièrement « référence » par les autres ou une combinaison des deux.
Cette limite doit nous conduire à distinguer les sites en fonction de leur degré d'ouverture et leur caractère de site de référence. Le degré d'ouverture d'un site est d'autant plus élevé que le site en question renvoie à de nombreux autres sites. Un site de référence correspond, pour sa part, à un site qui est souvent cité par les autres sites analysés.
Ce découpage permet de dresser une typologie des sites autour de quatre catégories illustrées Figure 2.



(1) X: un site ouvert
(2) Y: un site de référence
(3) Z: un site relais
(4) T: un site fermé

Figure 2: Typologie des sites

- L'ouverture signifie la non fermeture sur soi et la prise en compte de l'esprit du net qui repose sur l'interaction.
- Etre site de référence signifie avoir une reconnaissance de la qualité de son site par d'autres qui s'y réfèrent.

Nous avons mesuré le degré d'ouverture d'un site par le nombre de liens vers des sites externes au départ de ce site. Le caractère de site de référence est attribué à un site en fonction du nombre de liens externes qui permettent de se connecter à ce site. Le problème principal est celui de la fixation du seuil permettant de répartir les sites suivant qu'ils soient purs (sites ouverts ou sites de référence) ou impurs (sites relais).
Le tableau 2 fournit quelques exemples de sites caractéristiques:

SITE	NOMBRE DE FOIS OU CE SITE EST REFERENCE PAR LES AUTRES	NOMBRE D'AUTRES SITES AUQUEL CE SITE SE REFERE	NATURE DU SITE
www.afnor.fr	0	0	Site fermé
www.abes.fr	0	8	Site ouvert
newsup.univ-mrs	3	43	Site ouvert
www.cnrs.fr	27	1	Site de référence
www.ens.fr	9	4	Site de référence
www.enst.fr	6	4	Site relais

Tableau 2: Exemple de découpage des sites

Ce découpage est fécond dans la mesure où il permet de rendre compte de la centralité dont fait l'objet un sommet du réseau [Degrenne et alii]. La nature du site (référence, fermé,

ouvert, relais) peut servir de clé de filtre des données pour obtenir une meilleure visualisation du réseau.

2.3 La prise en compte de la référence mutuelle de deux sites.

Il existe une autre façon de lever l'ambiguité associée à la faible lisibilité des arcs du réseau de la Figure 1. Le réseau présenté figure 3 a été obtenu en transformant le sens donné à l'arc dans un réseau: au lieu de considérer qu'un arc entre deux sommets X et Y signifie l'existence d'un renvoi de X à Y ou de Y à X, nous allons considérer qu'un lien sur le réseau entre X et Y correspond au fait que le site X renvoie au site Y **et** que le site Y renvoie au site X. Ce type de relation correspond à une relation symétrique qui peut donner lieu à une interprétation simple. Un lien entre deux sites exprime la reconnaissance mutuelle des deux sites l'un pour l'autre. Le réseau obtenu est très lisible et très peu fourni. Les sommets qui le constituent sont surtout des instituts de formation.

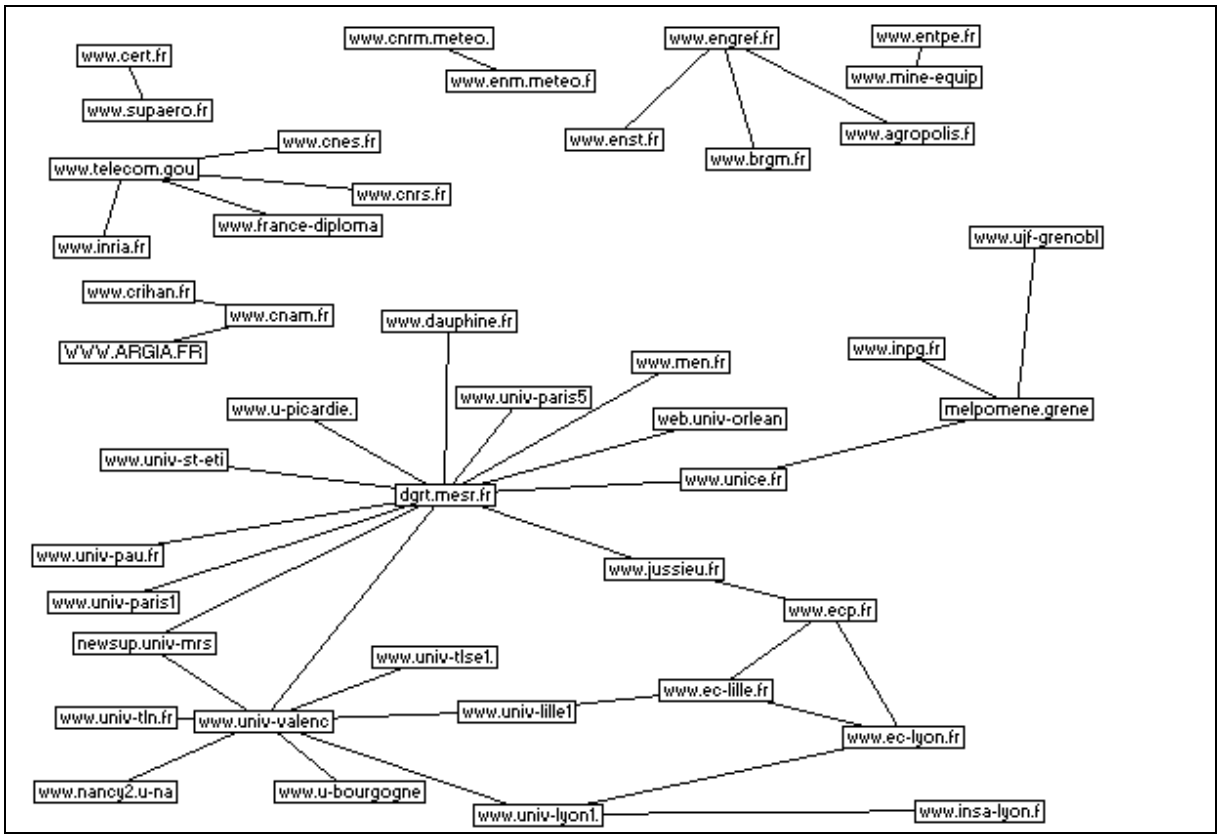


Figure 3: Réseau exprimant les relation symétriques entre les sites

2.4 Le produit matriciel :

Il est possible de faire subir au tableau 1 des manipulations mathématiques simples permettant de construire de l'information pertinente. Nous allons raisonner sur la matrice brute présentée Tableau 3. Elle associe à chacun des 109 sites analysés l'ensemble des sites auxquels il font référence.

Liste des sites auxquels se référent les 109 sites.

extrait de la liste des 109 sites à analyser

	WWW.UREC.FR	WWW.SFRS.FR	WWW.CNRS.FR	WWW.UNIV-PAU.FR	WWW.CERT.FR
www.univ-pau.fr	1	0	1	1	0
www.sfrs.fr	0	1	0	0	0
newsup.univ-mrs.fr	1	1	1	1	0
www.cnam.fr	1	0	0	0	0
www.insa-lyon.fr	1	0	0	0	0
www.hec.fr	0	0	0	0	0

Tableau 3: Extrait de la matrice de départ.

On peut faire subir à cette matrice deux manipulations mathématiques mettent en oeuvre les techniques du produit matriciel. On parlera de problème primal ou dual.

2.4.1 Résolution du problème primal.

Lorsqu'on multiplie la transposée de cette matrice par la matrice elle même, on obtient une matrice dont le tableau 4 dessous fournit une illustration:

	WWW.UREC	WWW.SFRS.	WWW.CNRS	WWW.UNIV-I	WWW.CERT.
WWW.UREC.FR	49	1	22	4	2
WWW.SFRS.FR	1	2	1	1	0
WWW.CNRS.FR	22	1	28	3	1
WWW.UNIV-PAU.FR	4	1	3	4	0
WWW.CERT.FR	2	0	1	0	5

Deux sites parmi les 109 renvoient conjointement aux sites de l'UREC et du CERT

Tableau 4: Matrice réalisée par produit matriciel.

Cette matrice comporte la liste des sites référencés par les sites de la liste initiale. Désormais, le lien entre deux sites correspond au nombre de fois où ces deux sites sont référencés ensemble. Le réseau global obtenu est inextricable. Par contre le filtrage de ce réseau permet d'obtenir des informations significatives. Nous avons retenu dans le réseau l'existence d'une relation entre deux sites à partir du moment où 15 des 109 sites y renvoyaient conjointement Le réseau obtenu est présenté figure 4:

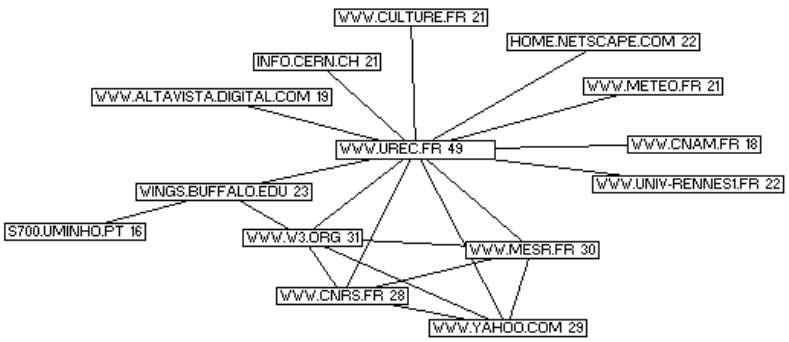


Figure 4 Réseau obtenu en conservant les paires supérieure ou égale à 15.

Dans ce réseau un lien entre deux sites correspond au nombre de sites de l'ensemble de départ que ces deux sites ont en commun. La relation entre deux sites n'est plus une relation binaire mais une relation valuée.
On arrive donc à exprimer la proximité des sites de référence pris deux à deux par le nombre de sites qui les ont co-cités. Cette analyse n'est pas sans rappeler l'analyse bibliométrique dite de la co-citation [Small].

2.4.2 La résolution du problème dual.

Le paragraphe précédant repose sur une propriété du produit matriciel qui permet, lorsqu'on multiplie la transposée d'une matrice par la matrice elle même d'obtenir une matrice carrée symétrique qui peut faire l'objet d'une transcription en réseau. Il est possible, de façon symétrique, de faire le produit de la matrice par sa transposée.
On obtient alors une matrice carrée symétrique dite duale de la première.
Dans l'exemple considéré, lorsqu'on multiplie cette matrice par sa transposée, on obtient une matrice carrée symétrique dans laquelle figure les sites de la liste initiale comme l'illustre le tableau 5.

	WWW.UREC.FR	WWW.SFRS.FR	WWW.CNRS.FR	WWW.UNIV-PAU.FR	WWW.CERT.FR
www.univ-pau	1	0	1	1	0
www.sfrs.fr	0	1	0	0	0
newsup.univ-r	1	1	1	1	0
www.cnam.fr	0	0	0	0	0

X

	www.univ-pau.fr	www.sfrs.fr	newsup.univ-mrs.fr	www.cnam.fr
WWW.UREC.FR	1	0	1	0
WWW.SFRS.FR	0	1	1	0
WWW.CNRS.FR	1	0	1	0
WWW.UNIV-PAU.FR	1	0	1	0
WWW.CERT.FR	0	0	0	0

=

	www.univ-pau	www.sfrs.fr	newsup.univ-r	www.cnam.fr
www.univ-pau	31	1	29	14
www.sfrs.fr	1	2	2	1
newsup.univ-r	29	2	39	15
www.cnam.fr	14	1	15	16

Tableau 5: L'application du produit matriciel.

On peut constater par exemple que le Cnam et l'Université de Pau font référence à 15 sites en commun.
Cette matrice exprime le nombre de sites qu'ont en commun les couples de sites pris deux à deux : un arc valué d'une valeur de v entre les sommets i et j signifie que ces deux sommets renvoient à v sites en commun. Munis de cette relation on peut interpréter le résultat obtenu Figure 5 et présenter un réseau possible correspondant à l'affichage des fréquences de paires supérieures ou égales à 15.

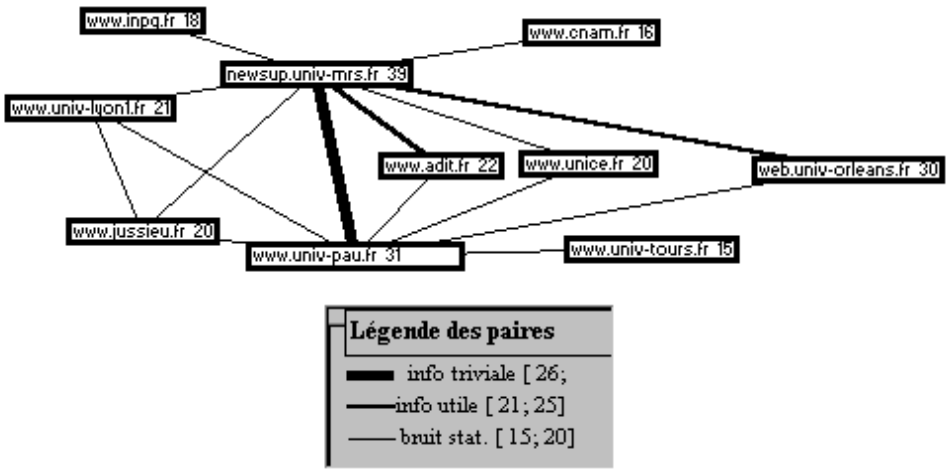


Figure 5: Réseau correspondant à la résolution du problème dual

Un lien entre deux sites signifie que ces deux sites ont la même politique de renvoi vers d'autres sites.
Plus un site renvoie à de nombreux autres, plus on risque de retrouver sur ce site de nombreux points commun avec les autres. Les sommets représentés sur ce réseau sont surtout des sites liés à l'enseignement: ils se catactérisent par des stratégies d'ouverture similaires.

Cet article propose de représenter sous forme de cartes appelées réseaux, un ensemble de sites Internet dans une perspective interactionniste. Chaque carte visualise les relations hypertextes entre des sites sur un thème donné. L'automatisation complète de la chaîne de traitement et son intégration en aval d'un moteur de recherche devraient permettre d'extraire plus rapidement les sites de références lors d'une recherche d'information sur Internet.

REFERENCES

Boutin E., Quoniam L., Rostaing H., Dou H. (1995*) A new approach to display co-authorship and Co-topicship through Network mapping* , Acte du colloque « Fifth International conference on scientometrics and informetrics », Chicago.

Degrenne A., Forse M. (1994*), Les réseaux sociaux*, éditions Armand Colin.

Dousset B., Dkaki T., Mothe J.(1997) *Veille scientifique et technique sur Internet*, actes du colloque : « Les systèmes d'information élaborées, SFBA », île Rousse, 12 au 16 Mai 1997.

Ducloy J., Nauer E., Lamirel J-C (1997) R*echerche précise d'information sur le www et veille technologique: utilisation de données structurées pour l'interrogation via les moteurs de recherche.* Actes du colloque : « Les systèmes d'information élaborées, SFBA », île Rousse, 12 au 16 Mai 1997.

Mannina B., Quoniam L., Dou H. (1997*) Auresys: un agent intelligent au service de l'information stratégique*, Actes du colloque : « Les systèmes d'information élaborées, SFBA », île Rousse, 12 au 16 Mai 1997.

Rostaing H (1993), Veille technologique et bibliométrie : concepts, outils et représentations, Thèse : Aix Marseille III, 353 p.

Small H.G (1973*), Co citation in the scientific litterature: a new measure of the relationship between two documents*, Journal of the American Society for Information Science, Vol 24, N°4, p.265-269.