HAL
archives-ouvertes.fr

Le Web et ses outils d'orientation. Comment mieux appréhender l'information disponible sur l'Internet par l'analyse des citations ?

Hervé Rostaing

▶ To cite this version:

Hervé Rostaing. Le Web et ses outils d'orientation. Comment mieux appréhender l'information disponible sur l'Internet par l'analyse des citations ?. Bulletin des bibliothèques de France, Ecole Nationale Supérieure des Sciences de l'Information et des Bibliothèques (ENSSIB), 2001, Les topographies du savoir, pp.68-77. sic_00000116v2

HAL Id: sic_00000116
https://archivesic.ccsd.cnrs.fr/sic_00000116v2

Submitted on 8 Nov 2006

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Le Web et ses outils d'orientation

Comment mieux appréhender l'information disponible sur Internet par l'analyse des citations ?

Je n'étonnerai personne en évoquant ma confusion devant l'évolution galopante d'Internet et plus particulièrement du World Wide Web. Je me souviens des premières années où mon laboratoire s'était connecté à ce nouveau média. Dès 1992, le directeur du Centre de Recherche Rétrospective de Marseille (CRRM) avait négocié avec le centre de calcul de notre faculté pour bénéficier d'une connexion au réseau de l'enseignement supérieur Renater avant la majeure partie des autres laboratoires.

Hervé Rostaing

CRRM, Centre Scientifique
de Saint Jérôme,
Université Aix-Marseille III
rostaing@crrm.u-3mrs.fr

À cette époque, la recherche des sites Web scientifiques français ou étrangers n'était pas aussi complexe que de nos jours. Le site de l'Urec[1] jouait le rôle de répertoire français de tous les sites Web et identifiait tous les autres répertoires nationaux. Une simple consultation de ces répertoires, lus par ordre alphabétique ou par regroupement géographique, était largement suffisante pour trouver son bonheur.

Nous sommes bien loin de l'époque où le Web gardait cette dimension intimiste. Qui n'a pas perdu plusieurs dizaines de minutes, voire plusieurs heures, à tenter de trouver l'information convoitée, sans l'assurance du moindre succès ? Les outils actuellement proposés pour le repérage de l'information sur Internet arrivent à leur apogée. Certains experts prédisent que dans moins de cinq ans, la recherche d'information sur Internet par l'emploi de mots-clés ne sera qu'un vague souvenir pour les scientifiques (1). De nombreuses équipes de recherche réfléchissent actuellement aux outils et systèmes de recherche de demain. Les réflexions s'orientent principalement vers les portails personnalisés. Nous évoquerons ici des approches complémentaires inspirées de recherches antérieures en Sciences de l'Information. Jusqu'alors, ces recherches ont plus particulièrement été appliquées en bibliométrie et scientométrie (2).

Internet et médiation par les professionnels

Les professionnels du document et de l'information, bibliothécaires et documentalistes, ne peuvent plus ignorer Internet. Ce média est devenu

1. Unité Réseau du CNRS, l'Urec est un laboratoire qui participe à la gestion et au développement du réseau RENATER. http://www.urec.fr

Hervé Rostaing *est maître de conférences au Centre de recherche rétrospective de Marseille (CRRM), et co-dirige la Maîtrise « Nouvelles technologies de l'Information pour le développement des Entreprises ». Docteur en Sciences de l'Information et de la Communication pour sa thèse* Veille Technologique et Bibliométrie : concepts, outils, applications, *soutenue en 1993 à l'Université d'Aix-Marseille III, il a notamment publié* La bibliométrie et ses techniques *(Sciences de la société), ainsi que de nombreux articles.*

un haut lieu de communication entre les chercheurs, scientifiques et autres « sachants ». De multiples centres d'information se développent, à la fois centres de diffusion entre ces « sachants » et centres de connaissances pour les étudiants ou usagers d'Internet. Le rôle des bibliothèques est primordial pour une meilleure appropriation de ce nouveau média, du fait de sa complexité et de son évolution constante. Celles-ci ne peuvent se contenter du simple rôle de fournisseur d'accès, par la mise à disposition du public de parcs d'ordinateurs connectés. L'introduction d'Internet dans les bibliothèques doit réactualiser leurs missions de soutien et de formation à la recherche d'information. Le rôle de médiateur entre l'utilisateur de l'information et les sources sur Internet n'a jamais été aussi important, puisque les systèmes d'aide à la recherche disponibles sur Internet sont largement déficients et fortement irrationnels, comme nous le mentionnerons plus loin. Offrir les clés de maîtrise et de compréhension de ces nouveaux centres de ressources informationnelles entre totalement dans la mission d'intérêt public des bibliothèques.

Bon nombre de bibliothèques ont créé leur propre site Web. L'internaute peut non seulement y trouver des catalogues en ligne, mais aussi des guides thématiques répertoriant et classant un ensemble de sites et de ressources sélectionnés. Ce nouveau service met en œuvre une démarche très similaire à celle de l'analyse documentaire employée lors du catalogage ou de l'indexation documentaire. Ce style de répertoire est même considéré, par certain, comme devant devenir l'une des missions principales du bibliothécaire (3). L'objet analysé n'est plus un document papier mais une ressource Internet (page Web, site Web, répertoire FTP, forum de discussion…). Plus que lors de l'analyse documentaire, ce travail fait réellement appel à une phase d'évaluation. Contrairement aux documents publiés qui sont soumis à de nombreuses étapes de validation (comité de lecture, éditeur) et de mise en conformité (présentation du travail selon des plans de construction peu souples, mise en forme respectant certaines normes ou conventions), les ressources Internet ne sont soumises à aucune contrainte ou vérification. La personne qui évalue un site doit non seulement faire appel à ses compétences documentaires et d'expert (extraction de connaissances dans un domaine plus ou moins pointu), mais aussi à ses compétences d'internaute. Ces compétences d'utilisateur d'Internet doivent lui permettre d'introduire d'autres éléments caractérisant la ressource comme le degré de véracité des informations, l'ergonomie de l'interface et de la navigation, les vertus pédagogiques, le sérieux porté aux renvois vers d'autres sites, etc. Tous ces éléments d'évaluation ont autant, sinon plus d'importance que le premier travail descriptif et analytique, pour bien mesurer la qualité d'un site Internet.

Quels sont les outils actuellement disponibles pour aider les internautes et les professionnels de l'information dans l'identification des sources et dans leurs évaluations ? Hormis la grille d'évaluation qui constitue le support technique de la phase finale de l'évaluation d'un site (4), nous allons d'abord évoquer et commenter rapidement les différents outils conçus pour aider à repérer l'information sur Internet, puis mentionner ceux qui ont pour objet de proposer des indicateurs d'évaluation.

> Le rôle de médiateur entre l'utilisateur de l'information et les sources sur Internet n'a jamais été aussi important, puisque les systèmes d'aide à la recherche disponibles sur Internet sont largement déficients et fortement irrationnels

Localiser l'information sur Internet

Les dernières estimations sur la taille du Web se chiffrent à plus d'un milliard de pages (5). Les outils actuellement proposés pour rechercher une information ne suivent plus la croissance galopante du Web.

Les annuaires

Les annuaires ou répertoires, résultat d'un traitement et d'une expertise humaine, ne sont plus, depuis très longtemps, représentatifs de l'information accessible sur Internet. Ils ne référencent principalement que des sites (page d'accueil) et non l'ensemble des pages présentes dans ces sites (enchaînement des pages liées à la page d'accueil). En revanche, ils ont l'énorme avantage d'être le résultat d'un travail de qualité offrant une sélection, un classement hiérarchique et une description analytique des sites. Pour des raisons évidentes de course à l'exhaustivité, la qualité de ce travail humain se dégrade en proportion de l'accroissement du Web. Ces annuaires sont des points d'entrée privilégiés pour les internautes inexpérimentés.

Les moteurs de recherche

Les moteurs de recherche, pour leur part, indexent automatiquement les pages des sites Web. Un robot est chargé de repérer et de collecter les nouvelles pages en parcourant le Web selon les hyperliens. Une page collectée est alors caractérisée par extraction automatique de mots présents dans cette dernière. Ces mots sont alors introduits dans l'index du moteur, qui sera par la suite consulté par les utilisateurs. Ce système de qualification de l'information étant relativement fruste et totalement automatisé, on aimerait s'attendre à une couverture presque exhaustive du Web, quitte à négliger la qualité de l'indexation. Ce n'est malheureusement pas le cas : une récente étude du NEC Research (6) montre qu'aucun des moteurs de recherche ne couvre plus de 16 % du Web. La réunion des six plus grands moteurs de recherche ne couvrirait que 60 % du Web. Ce dernier chiffre nous indique non seulement l'impossibilité d'une couverture parfaite, mais surtout le très faible taux de recouvrement entre les moteurs !

De plus, de nombreuses erreurs et instabilités persistent dans l'index d'un moteur : des pages supprimées dans les sites mais maintenues dans l'index, des pages modifiées dans les sites et toujours caractérisées par les mots de l'ancienne version dans l'index, des pages de grandes tailles indexées uniquement avec un ensemble restreint de premiers mots, la disparition de pages de l'index alors qu'elles sont toujours présentes dans les sites, la disparition de mots caractérisant une page sans que la page ait été modifiée (7), etc.

D'autres déficiences, moins inhérentes à la faiblesse des algorithmes informatiques qu'à la nature changeante du Web, viennent s'ajouter à celles précédemment citées. Les moteurs ne savent pas traiter les textes conçus dans des formats plus riches que le HTML, tels que les formats PDF[2] ou PS[3]. Or, une grande partie des textes scientifiques est disponible sous cette forme sur les sites Web, pour exploiter pleinement toute la richesse des fonctions des traitements de texte. Les moteurs sont par ailleurs incapables de référencer les pages appartenant à ce qu'on appelle le « Web invisible ». Le Web invisible est constitué de pages générées dynamiquement lors du parcours de l'utilisateur dans un site. Les moteurs sont dans l'incapacité de référencer ces pages qui n'ont qu'une durée de vie éphémère et non un caractère statique. Or, là encore, un nombre grandissant de sites est conçu ainsi, pour proposer un environnement et des services adaptés au profil du visiteur, et toutes les bases de données consultables par Internet en font partie.

La réunion des six plus grands moteurs de recherche ne couvrirait que 60 % du Web

Toutes ces lacunes désorientent profondément les utilisateurs des moteurs de recherche et ce ne sont pas les méta-moteurs[4] ni les agents de recherche[5] qui peuvent les combler, puisqu'ils exploitent tous ces moteurs comme point de départ de leur traitement. La plupart d'entre eux permettent tout de même d'éliminer les erreurs les plus grossières, comme la suppression des liens vers des pages inexistantes et vers des pages ne contenant plus les termes de la requête, offrant ainsi un gain de temps inestimable pour l'utilisateur. Certains améliorent le taux de rappel en travaillant sur la reformulation de la requête de recherche[6] par l'emploi de traitements lexicaux, d'un dictionnaire de synonymes, voire de traitements linguistiques. D'autres optent pour la propagation[7] par les hyperliens des pages pertinentes[8], estimant que ces liens ont une probabilité assez forte d'aboutir sur des pages elles-mêmes pertinentes.

Discrimination de l'information sur Internet

Obtenir une liste, la plus exhaustive possible, des sources répondant à une requête de recherche est un début nécessaire, mais insuffisant dès lors que le nombre de réponses dépasse la cinquantaine. Il devient important de pouvoir discriminer, ordonner et évaluer tous ces résultats. L'internaute a besoin d'un ordre de lecture de toutes ces pages. Mais il peut aussi éprouver l'envie d'avoir une vue globale des résultats, pour l'aider à mieux appréhender l'intégralité de l'information obtenue.

Le principal outil d'aide à la lecture proposé par les systèmes de recherche d'information sur Internet est un simple classement, selon un indicateur souvent nommé « indice de pertinence ». Cette mesure est fondée à la fois sur la fréquence d'apparition des termes de la requête dans la

2. *Portable Document Format :* format de fichier développé par Adobe. http://www.adobe.com
3. Postscript : langage de codage de la mise en forme de l'impression pour imprimante compatible Postscript.
4. Systèmes qui interrogent en frontal plusieurs moteurs de recherche puis fusionnent les différents résultats obtenus en un seul :
MetaCrawler http://www.metacrawler.com
SavySearch http://www.search.com
ProFusion http://www.profusion.com
Beaucoup! http://www.beaucoup.com, etc.
5. Logiciels qui jouent le rôle de méta-moteurs et qui apportent des fonctionnalités supplémentaires pour aider l'utilisateur dans sa recherche et dans l'exploitation des résultats. Voir le site de Cybion pour son labo sur les agents intelligents : http://www.veille.com/labo_agents/labo_agents.htm
6. StrategicFinder http://www.strategicfinder.com, DigOut4U http://www.arisem.com.
7. Websleuth http://www.promptsoft,ware.com, DigOut4U http://www.arisem.com
8. Ici, le terme pertinent est utilisé pour qualifier toute page contenant les termes de la requête selon la logique booléenne employée.

page et sur leurs localisations [9]. Cet indicateur est utilisé systématiquement par tous les moteurs de recherche, de façon à classer le résultat d'une recherche par ordre d'intérêt décroissant selon cette mesure. Tous les internautes ont pu vérifier, par expériences, du peu d'intérêt qu'a ce classement. Il n'est pas rare de retrouver, en tête de liste, des pages Web qui ne sont pas du tout en adéquation avec la requête.

L'effet Saint Matthieu

Quelques moteurs de recherche, dont le plus connu est Google [10], ont pris le parti d'employer un autre mode de classement des résultats. Les pages Web sont ordonnées selon leur *notoriété*. Ce principe est directement inspiré de recherches antérieures en scientométrie et principalement des travaux de Price (8) et Garfield (9) sur la pratique de la citation entre les articles scientifiques. Cette théorie veut qu'un article scientifique très fréquemment cité par les autres scientifiques fasse partie du cœur de la littérature scientifique. Selon cette théorie, ce cœur constitue le creuset de référence scientifique appartenant au fonds commun de connaissances, utile à tout nouveau développement. La citation serait comme une mesure du pouvoir d'utilité d'un article et par là même une certaine marque de qualité.

Appliquant cette théorie à l'espace hypertextuel du Web, une page qui est la cible d'un très grand nombre de liens est probablement non seulement une page validée [11] mais aussi une page détenant un contenu utile à un grand nombre. Elle a donc un degré d'utilité assez fort pour la communauté des internautes.

Bien évidemment, cet indice a des failles, tout comme la citation en scientométrie. L'autocitation [12], dans le monde du Web, correspond aux liens pointant sur une page alors qu'ils proviennent d'une page du même site. Ces liens sont probablement des liens de structuration du site lui-même, liens essentiellement utiles au parcours du site. Il paraît juste de ne pas les prendre en considération lors des comptages. Le phénomène de « l'effet Saint Matthieu [13] », qui symbolise le fait qu'on prête plus facilement aux riches, se vérifie là encore comme un invariant universel. Plus une page ou un site sera pointé par un grand nombre de liens, plus la probabilité d'y accéder sera grande, plus forte sera la probabilité qu'elle soit de nouveau la cible de prochains liens. Le taux de citations reçues par une page Web ne présume donc pas forcément de la qualité de son contenu, mais tout au moins de sa notoriété, de sa popularité, voire seulement de sa visibilité.

Le taux de citations reçues par une page Web ne présume pas forcément de la qualité de son contenu

Le pouvoir rayonnant

Un second indicateur peut être élaboré à partir de ce phénomène de référencement entre pages Web : le *pouvoir rayonnant*. Plus une page Web contient de liens vers d'autres sites Web plus son pouvoir rayonnant est important. Le projet Clever d'IBM (10) a même perfectionné cette mesure en donnant un poids plus fort aux liens pointant des pages à très forte notoriété. Plus une page fait référence à de nombreuses pages fortement citées, plus son pouvoir rayonnant s'amplifie. Les pages Web ayant un fort pouvoir rayonnant sont nommées *sites pivot* au sein du projet Clever. Les chercheurs impliqués dans ce projet ont très rapidement constaté que le système de classement qu'offrent les moteurs de recherche, fondé sur le calcul des occurrences/localisations des termes de la requête, n'était pas assez significatif. Ils ont donc cherché à améliorer la qualité de ce classement en appliquant la théorie de la citation et plus particulièrement les travaux sur la mesure du *facteur d'impact* mis au point par Garfield. L'objectif est de détecter puis de classer les pages appartenant aux deux catégories *page populaire* (appelée « page de référence ») et *page rayonnante* (appelée « page pivot »). Les estimations du degré de popularité et du degré de rayonnement des pages sont évaluées selon une logique circulaire : une page est d'autant plus populaire qu'elle est citée par des pages rayonnantes et réciproquement, une page est d'autant plus rayonnante qu'elle cite des pages populaires. Appliqué sur un ensemble de pages obtenues à la suite d'une requête sur un moteur de recherche, le logiciel Clever met en œuvre une heuristique itérative faisant converger ces estimations vers des valeurs finales permettant d'identifier et de classer les pages, soit par ordre de popularité, soit par ordre de rayonnement.

Cartographie relationnelle de l'information sur Internet

Les techniques de discrimination de l'information Web offrent une aide à la lecture en proposant les pages triées selon un indice d'évaluation. L'internaute se voit soutenu dans son investigation par cet ordonnancement

9. Poids plus grand pour les occurrences de termes dans le titre, les métadonnées et le début de la page que dans le reste de la page.
10. http://www.google.com
11. Cette page a été parcourue par un grand nombre de lecteurs, qui ont jugé bon de la citer en référence.

12. Le fait qu'un auteur cite ses propres travaux.
13. Ou encore appelé « avantage cumulé » en scientométrie.

des informations. En revanche, il lui est très difficile de se faire une idée générale du contenu de l'ensemble de ces pages. Des approches par cartographie relationnelle peuvent lui offrir cette vision globale des données, l'aidant dans sa démarche de synthèse de l'information.

Les techniques d'analyse relationnelle ont largement été employées en bibliométrie et scientométrie (11), depuis les premières cartes construites en 1973 par la méthode de l'analyse de la co-citation (12). Deux tendances majeures se confrontent en biblio-scientométrie pour la construction de ces cartes relationnelles. La première, la plus ancienne, se fonde sur l'analyse des relations entretenues entre les travaux scientifiques par le phénomène de la citation. Deux articles citant les mêmes travaux sont considérés comme très proches dans leur contenu et entretiennent donc des liens très forts, par le biais des travaux de référence communs. La seconde, plus récente, estime qu'il est préférable de mesurer les relations entre articles en évaluant directement la ressemblance des idées abordées dans les textes.Cette méthode repose sur l'exploitation du résultat de l'analyse documentaire effectuée lors de l'indexation d'une base de données bibliographique : les mots-clés ou les codes de classification. Ces descripteurs sont alors considérés comme les meilleurs représentants des concepts abordés dans les articles et sont donc utilisés pour comparer les profils de concepts des travaux à mettre en relation.

À l'instar de la biblio-scientométrie, l'espace du Web peut-être analysé selon ces deux tendances. Les concepteurs actuels de logiciels d'analyse du Web privilégient la cartographie des concepts[14]. Plus que le mode de représentation graphique de l'espace des relations des concepts, la méthode d'extraction des concepts des pages Web est la phase critique du traitement. L'extraction des concepts constitue en effet une phase très sensible, car elle met en œuvre des techniques de traitement du texte intégral. Pour que ces traitements fournissent un résultat pertinent, ils doivent respecter assez finement les règles linguistiques et les subtilités sémantiques du sujet étudié. De par la dimension internationale et multidisciplinaire d'Internet, il est encore peu envisageable de pouvoir parfaitement cartographier le Web par cette approche, sans un très lourd investissement en temps pour l'affinage de l'étape d'extraction.

> Soutenir l'internaute dans sa recherche, en lui donnant une image de l'organisation de l'information et des parcours possibles

La carte d'orientation du Web par l'analyse réseau des citations

L'expérience évoquée ci-dessous privilégie l'analyse de l'espace du Web par *la cartographie de la structure relationnelle des hyperliens entre les pages.* Cette approche s'affranchit des contraintes linguistiques et met en valeur la structure même de l'organisation de l'information sur Internet. Sans toutefois pouvoir se représenter le contenu même des pages Web,l'internaute obtiendra un soutien dans sa lecture, en disposant d'une image de l'organisation de l'information et des parcours possibles dans sa quête de savoir.

Cette étude a été réalisée en juin 1999 et présenté à la conférence Cybermetics'99 (13). L'objet de l'étude était de fournir une image claire de la présence de la communauté des biblio-scientométriciens sur Internet.Contrairement au projet Clever, l'objet d'évaluation dans cette étude n'est pas la page, mais le site Web[15]. Ce niveau de « granularité » semble plus approprié à la construction d'une cartographie. Cette cartographie a été élaborée en suivant la démarche suivante : collecte des pages Web contenant les termes *bibliométrie* ou *scientométrie,* extraction du nom des sites à partir des URL mentionnées dans les hyperliens, répartition des sites en quatre catégories (site référence, site rayonnant, site passerelle et site trou noir), représentation graphique du réseau des citations entre sites avec mise en évidence de la catégorie pour chaque site. La cartographie obtenue est présentée en figure 3 (page 77).

Les deux premières phases ont été obtenues grâce à l'agent de recherche Auresys développé au CRRM[16]. Ce robot est unique en son genre car il possède des fonctionnalités qui ne sont jamais réunies en un seul outil du commerce. Une première fonction rarement proposée par les agents du commerce[17] nous sera utile dans cette étude pour parfaire la collecte des données. Auresys propage sa recherche par navigation dans la toile des hyperliens pour identifier de nouvelles pages comme le font les robots de collecte des moteurs de recherche.Ainsi, à partir

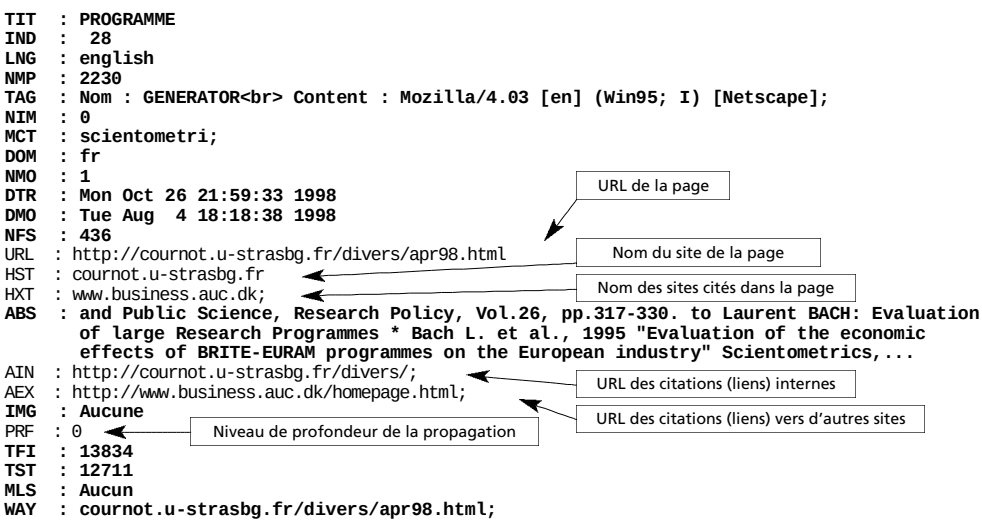
14. WebSom http://websom.hut.fi/websom/, SemioMap http://www.semio.com, Umap http://www.trivium.fr, CISpider et MetaSipder http://ai.bpa.arizona.edu/go/downloads.html, WordMapper http://www.grimmersoft.com.

15. Le terme site est utilisé tout au long de cet article, alors que l'étude porte plus exactement sur les noms de serveurs de sites Web. Un serveur pouvant héberger plusieurs sites Web, cet abus de langage équivaut à une légère approximation.

16. Auresys a été développé par Bruno Mannina au cours de sa thèse. Pour des compléments d'information consulter son site http://ms161u06.u-3mrs.fr

17. DigOut4U http://www.arisem.fr et Websleuth http://www.promptsoftware.com propose cette fonction.

Figure 1. **Fiche bibliographique* d'une page Web générée par Auresys**



* La signification des champs est accessible en ligne à l'adresse http://ms161u06.u-3mrs.fr/fichebiblio.html

de la requête *scientometri* or biblio - metri*[18] or scientometry or biblio - metry* soumise à Altavista, Auresys rapatrie les 1010 premières pages proposées par Altavista sur les 3518 répertoriées[19]. Il conserve uniquement les pages répondant réellement à la requête[20], et enrichit sa collecte en parcourant les pages pointées par ce premier ensemble de pages. Les nouvelles pages obtenues sont testées pour vérifier si elles correspondent bien à la requête. Dans le cas positif, elles sont conservées et deviennent le point de départ d'une nouvelle progression du parcours de la toile des liens. Ce principe de propagation tentaculaire (14) est réalisé tant que de nouvelles pages pertinentes sont localisées et autant de fois que l'utilisateur l'a précisé.

Dans notre étude, la profondeur de propagation a été paramétrée à 3. Le tableau 1 recense quelques données sur cette progression. La ligne *profondeur 0* récapitule le résultat obtenu à la suite de la requête soumise à Altavista. Sur les 1010 pages proposées par Altavista, seules 421 répondent réellement à la requête. La proportion de pages erronées[21] pointées par Altavista est de 58 % pour cette requête. La fonction de collecte par propagation d'Auresys, configuré à profondeur 3, va bonifier ce résultat par l'apport de 362 (783-421) nouvelles pages soit une augmentation de 86 %. Cette amélioration est au sacrifice d'une performance temps non négligeable, puisqu'il a fallu visiter 36519 pages pour ne découvrir que 362 nouvelles pages pertinentes. En terme de sites Web (colonne 4 du tableau 1), Auresys n'a permis une amélioration quantitative que de l'ordre de 11 %. On peut dire que le principe de collecte par propagation améliore nettement l'identification des pages pertinentes par un parcours en profondeur dans les sites proposés par le moteur de re-

Tableau 1. **Collecte par propagation des pages Web en biblio-scientométrie avec Auresys**

Profondeur	*Pages visitées*	*Pages retenues*	*Sites trouvés (ensemble A)*	*Sites cités*	*Sites cités inclus dans l'ensemble A*
0	1010	421	299	1189	64
1	4029	501	315	1249	67
2	12612	597	321	1367	83
3	37529	783	331	1785	97

18. Cette requête identifie toutes les pages contenant soit *bibliometry* soit *scientometry* soit un mot commençant par les racines *bibliometri* ou *scientometri*. Le symbole « * » représente la troncature large pour Altavista.

19. Altavista n'offre à l'utilisateur que les 1010 premières réponses de son index.

20. Nous avons vu plus haut que les moteurs sont dans l'impossibilité de mettre à jour leur index dès que des modifications sont apportées à des pages Web. Il n'est donc pas rare que les pages pointées par l'index d'un moteur ne contiennent plus les termes de la requête.

21. Les raisons de cette divergence sont : serveurs Web inaccessibles, pages inexistantes, pages ne contenant pas les termes de la requête.

cherche. En revanche, il n'améliore que modestement l'identification de nouveaux sites. L'emploi de plusieurs moteurs de recherche comme point de départ de la propagation compléterait astucieusement cette collecte. La cinquième colonne du tableau 1, indique l'accroissement des sites (et non des pages Web) référencés par les pages collectées par le robot. L'amélioration de 50 % de ces sites cités se réduit à 27 % (colonne 6 du tableau 1) dès lors que tous les sites référencés par les pages Web ne sont pas considérés.

L'étude n'a traité que les sites cités appartenant à notre corpus de collecte,c'est-à-dire uniquement les sites qui contiennent d'une façon certaine des pages répondant à notre requête, et abordant donc des aspects concernant le sujet de notre étude : le phénomène de citation entre les sites de la communauté biblio-scientométrie. Il est à noter que la pratique de la citation Web entre les membres de cette communauté n'est pas très forte, puisque sur les 1 785 sites cités, seulement 97 (5 %) sont des sites de la communauté [22]. Seuls ces 97 sites de la communauté biblio-scientométrie ont été retenus pour la cartographie finale.C'est-à-dire des sites étant à la fois en position de citant et de cité au sein du corpus.

Ces traitements de comptage ,de filtrage et tous les traitements présentés ultérieurement sont facilités par une fonction unique à Auresys. En plus du rapatriement des pages retenues et de leur organisation selon un mode de classement favorisant une navigation plus cohérente entre les pages, le robot établit aussi une fiche descriptive structurée pour chaque page. Cette fiche descriptive de page Web est analogue à la référence bibliographique d'un article scientifique (analyse documentaire en moins).

Une fois le noyau dur des 97 sites de la communauté établi, la seconde phase du traitement est leur classement parmi les quatre catégories précitées : *site référence, site rayonnant, site passerelle* et *site trou noir.* La règle de détermination de cette qualification de sites est bien moins complexe que celle proposée par le projet Clever. Les critères employés sont synthétisés dans la table de qualification présentée au tableau 2. Ainsi, un *site référence* est un site qui reçoit deux fois plus de citations qu'il n'en fait, un *site rayonnant* est un site qui effectue deux fois plus de citations qu'il n'en reçoit, un *site trou noir* est un site cité par la communauté, mais qui ne pratique pas la réciprocité, un *site passerelle* est presque autant cité qu'il cite lui-même. Restent alors les sites qui ne sont ni cités ni citants, des sites totalement isolés ne participant pas au principe de construction collective d'Internet.

Les sites sont catégorisés en appliquant ces critères à un tableau de citations croisées [23] construit par le logiciel bibliométrique Dataview (15) à partir des notices descriptives des pages Web. Ce tableau de 97 lignes par 97 colonnes représente en ligne les sites en position de citant et en colonne en position de cités (tableau *X* sur la figure 2). Ce tableau synthétise l'ensemble des citations entretenues entre les 97 sites. La somme d'une ligne du tableau correspond au nombre de citations effectuées par un site et la somme d'une colonne au nombre de citations reçues par un site.

Ce tableau a la propriété d'être carré, mais asymétrique. Le logiciel Matrisme (16), employé pour la représentation infographique du réseau des citations, ne permet pas de construire de graphes orientés, et n'accepte donc que des tableaux symétriques en entrée. Un artifice de calcul (somme matricielle du tableau *X* et de sa transposée, voir figure 2) nous permet de transformer ce tableau de citations croisées en un tableau symétrique mesurant la force d'association du phénomène de citation entre deux sites. À la suite de ce calcul, la valeur à l'intersection d'une ligne et d'une colonne représente le nombre de citations (hyperliens) existant entre deux sites, indépendamment du sens de ces citations.

Ce tableau, exploité par le logiciel Matrisme, permet de construire la carte des citations entretenues entre les sites Web de la communauté biblio-scientométrie (figure 3). Cette carte est un réseau constitué de nœuds et d'arcs les reliant. Les nœuds symbolisent le nom des sites Web participant à la structuration

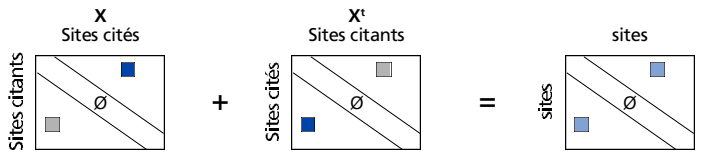
Tableau 2. **Tableau récapitulatif des règles de catégorisation des sites**

		Nombre de citations reçues par un site				
		0	1	2	3	4
Nombre de références effectuées par un site	0	Isolé	Trou noir			
	1			Référence		
	2	Rayonnant				
	3			Passerelle		
	4					

22. Il est possible que des pages citées à profondeur 3 soient pertinentes (au sens de la requête) et appartiennent à des sites encore non identifiés, puisqu'à chaque étape de la progression le robot a détecté de nouveaux sites cités (colonne 6 du tableau 1). Pour le vérifier, il faudrait consulter chacun de ces liens, ce qui correspondrait à une propagation de profondeur4.

23. Ce tableau est nommé « tableau de citations croisées » par analogie à l'analyse des citations croisées développée en scientométrie pour l'analyse de la structuration des relations entre des revues scientifiques (2).

Figure 2. **Calcul de la mesure d'association du phénomène de citation entre les sites Web**



d'Internet [24]. La couleur d'un nœud représente, parmi les quatre catégories précitées, celle à laquelle le site a été affecté (voir légendes figure 3). Chaque fois qu'une relation de citation est entretenue entre deux sites, un arc relie les deux nœuds correspondant sur le graphe. L'épaisseur de cet arc symbolise l'intensité de relation entre deux sites, mesurée en nombre de citations. Comme ces arcs ne sont pas orientés, il a fallu créer cette notion de catégorie de site pour identifier si les arcs connectant les nœuds représentent principalement des relations entrantes (citations reçues par ce site) ou des relations sortantes (citations effectuées par ce site). La couleur d'un nœud complète donc l'information sur la connectivité et la place d'un site au sein du réseau, par l'expression de l'orientation de ses connexions. Ainsi, certains nœuds sont au cœur du réseau, soit parce qu'ils sont très fréquemment mentionnés dans les pages de la communauté (référence), soit parce qu'ils sont des centres d'orientation vers les sites de la communauté (rayonnant), soit parce qu'ils jouent les deux rôles (passerelle). Ainsi, le CWTS de l'université de Leiden, qui est une référence dans le domaine de la biblioscientométrie par ses travaux scientifiques, l'est aussi dans l'espace du Web. Par ailleurs, un site comme celui du Cindoc, en Espagne, joue le rôle fédérateur de répertoire de ressources présentes sur le Web.

Cette carte relationnelle livre, à un instant donné, une image assez fidèle de l'organisation de l'information

> Le rôle de médiateur et de formateur aux outils et aux démarches de recherche et de collecte de l'information n'a certainement jamais été autant d'actualité

Web pour une communauté scientifique. L'utilisateur peut très rapidement privilégier des parcours de navigation dans cet espace informationnel. S'il veut parcourir scrupuleusement l'ensemble des sites majeurs de cette communauté, il aura intérêt à consulter le site du Cindoc pour rayonner en direction des autres sites. Par contre, s'il a peu de temps, il peut privilégier l'accès direct aux sites qui font autorité dans le domaine. L'utilisateur peut ainsi élaborer une tactique d'investigation sans se sentir totalement désorienté.

Cette méthode d'analyse de l'information Web prend tout son intérêt dès lors qu'elle est intégrée à un agent de recherche. Un tel outil, pour qu'il soit adapté aux changements perpétuels d'Internet, doit être aussi réactif que possible. Pour simple exemple, prenons le cas du site de mon laboratoire. À l'époque de l'étude, notre serveur était sur le domaine Internet « univ-mrs.fr ». Notre site est donc représenté sur le graphe figure 3 par le nœud « crrm.univ-mrs.fr ». Or depuis, le service de gestion du réseau de notre faculté nous a imposé un nouveau domaine « u-3mrs.fr ». Le nouveau nom de notre site est devenu « crrm.u-3mrs.fr». Le graphe présenté n'est plus tout à fait représentatif de notre position dans le réseau des citations, car la majeure partie des liens dirigés vers notre site n'a pas été réactualisée. Nous avons perdu une partie de notre visibilité sur le Web, et donc de notre notoriété. Un outil qui, dès la collecte des pages terminée, générerait une carte d'orientation pour la navigation dans les données, aiderait immédiatement l'utilisateur à tirer un meilleur profit de l'ensemble des données. Ce produit est en cours de réalisation [25].

La fuite des technologies de l'information

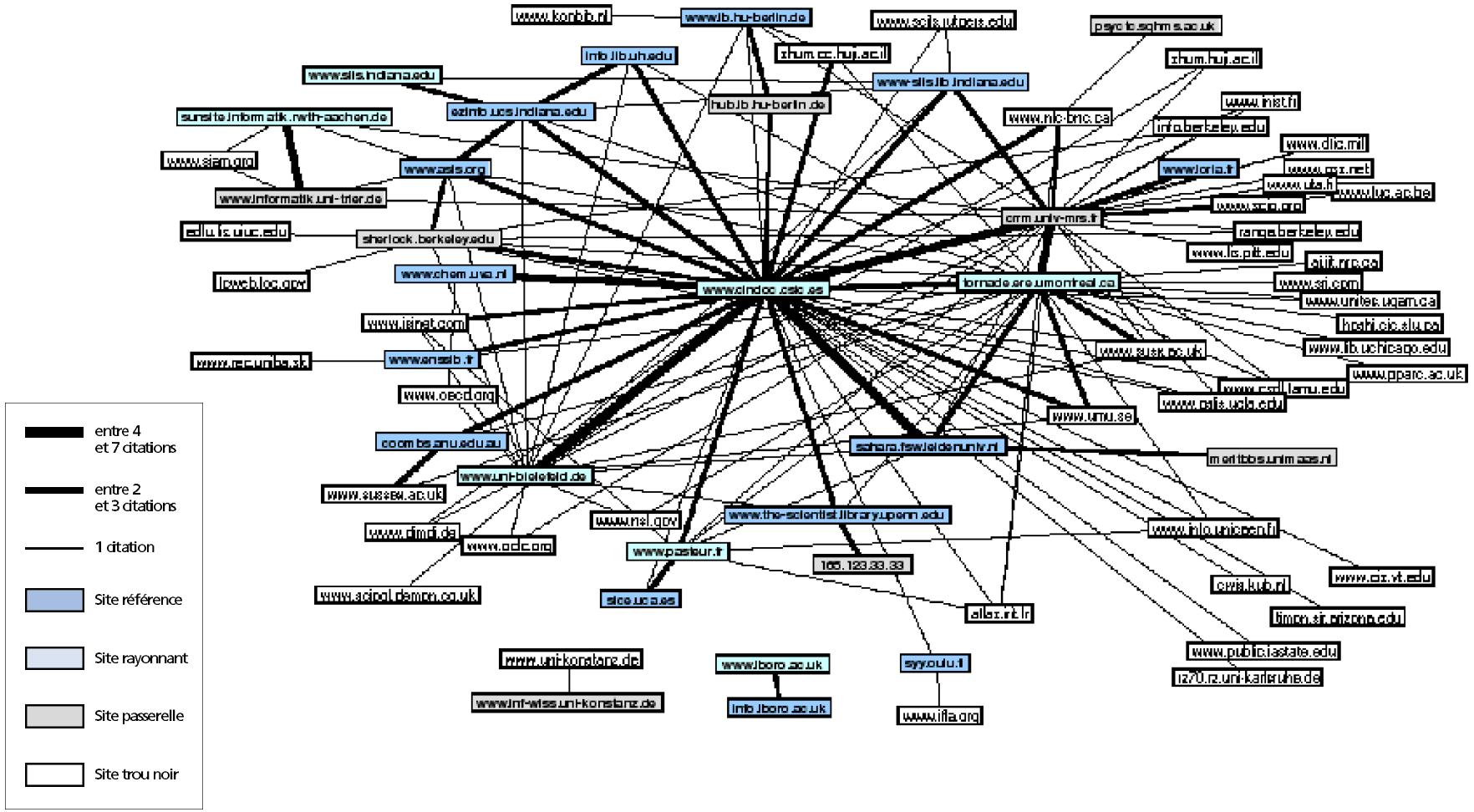
Dans le tourbillon incessant et chaotique d'Internet, il devient urgent de mettre à disposition des internautes des outils et des techniques de recherche adaptés. Les agents de recherche et leur nouvelle génération, les agents intelligents, n'en sont encore qu'à leurs balbutiements. Les professionnels de l'information, au même titre que les internautes avertis, doivent maintenir leur perspi-

24. Cette carte ne présente pas tous les sites du corpus mais que ceux qui participent à la structuration du Web. Parmi les 97 sites répertoriés dans le tableau des citations croisées, 23 n'entretenaient aucun lien avec les autres sites. Ces sites sont à la fois des sites en position de citant et de cité lorsque le référentiel est le corpus au complet. Mais dès lors que les sites n'ayant pas ces doubles attitudes sont ignorés, certains sites se retrouvent isolés dans le nouveau référentiel car ils citent et sont cités par des sites mis de côté.

25. Le concepteur du robot Auresys a fondé une entreprise avec d'autres anciens doctorants de notre laboratoire. Il intègre régulièrement les résultats de recherches menés en collaboration avec le CRRM dans le développement de son nouvel agent intelligent WebProcess. Une version restreinte de cet agent est en démonstration gratuite sur Internet à l'adresse http://www.searchprocess.fr

BBF 2001
Paris, t. 46, n° 1

Figure 3. **Réseau des citations entre les sites Web de la communauté biblio-scientométrie en juin 1999**



cacité en alerte pour ne pas se faire distancer par les évolutions galopantes des technologies de l'information, et plus encore par leur pratique. Leur rôle de médiateur et de formateur aux outils et aux démarches de recherche et de collecte de l'information n'a certainement jamais été autant d'actualité. L'information est de plus en plus massive, disponible et accessible mais encore faut-il pouvoir la localiser, l'appréhender, l'assimiler avant de la diffuser !

Septembre 2000

Bibliographie

1. Butler, Declan , « Souped-up search engines », *Nature,* 2000, vol. 405, p. 112-115.

2. Rostaing, Hervé, *La bibliométrie et ses techniques,* Sciences de la société, 1996, coll. «Outils et méthodes».

3. Le Crosnier, Hervé, « Les bibliothécaires et le réseau. Un métier qui évolue avec les technologies» *in* Rouhet, Michèle (dir.), *Les nouvelles technologies dans les bibliothèques,* Paris, Électre-Cercle de la Librairie, 1996, coll. « Bibliothèques ».

4. Basset, Hervé, *Sélection et Évaluation de Sites Web scientifiques,* Mémoire de maîtrise de Sciences de l'Information et de la Communication, CRRM, Université Aix-Marseille III, 2000.

5. Inktomi Webmap: http://www.inktomi.com/webmap

6. Lawrence, Steve ; **Gilles,** C. Lee, «Accessability of information on the web», *Nature,* 1999, vol. 400, p. 107-109.

7. Bar-Ilan, Judit, « Search engine results over time – A case study on search engine stability », *Cybermetrics,* 1999, vol 2, p. 1. http://www.cindoc.csic.es/cybermetrics/articles/v2i11.html

8. De Solla Price, Derek, *Science et Suprascience,* Trad. française de *Little Science Big Science* (par G. Lévy), Paris, Fayard, 1972.

9. Garfield, Eugène, *Citation Indexing – its Theory and Application in Science, Technology, and Humanities,* New York, John Wiley & Sons, 1979.

10. Membres du projet Clever, « Hypersearching the Web », *Scientific American.* [En ligne], 1999. http://www.sciam.com/1999/0699issue/0699raghavan.html ; voir aussi : Membres du projet Clever, « Recherche intelligente sur l'Internet », *Pour la science,* 1999, n° 262. http://www.pourlascience.com/numerso/pls262/clever.htm ; voir aussi : *Le site Internet du projet Clever* http://www.almaden.ibm.com/cs/k53/clever.html

11. Rostaing, Hervé. *Op. cit.*

12. Small, Henry G., « Co-citation in the scientific literature: a new measure of the relationship between two documents », *Journal of the American Society for Information Science,* 1973, vol. 24, n° 4, p. 265-296.

13. Rostaing, Hervé ; **Boutin,** Éric; **Mannina,** Bruno, « Evaluation of Internet resources: Bibliometric technique applications », *Cybermetrics'99,* ISSI'99 post-conference seminar, University of Colima, Mexico, 9 juillet 1999. http://www.cindoc.csic.es/cybermetrics/cybermetrics99.html

14. Mannina, Bruno ; **Quoniam,** Luc, « Le Problème lié à la limitation du nombre de résultats par les moteurs de recherche », *Les Cahiers de la documentation,* à paraître.

15. Rostaing, Hervé ; **Nivol,** William ; **Quoniam,** Luc ; **La Tela,** Albert, « Le logiciel bibliométrique Dataview et son application comme outil d'aide à l'évaluation de la concurrence », *Revue française de bibliométrie appliquée,* 1993, n° 12, p. 360-387. http://crrm.u-3mrs.fr/res-teach/staff/pubrostaing.html

16. Boutin, Éric, *Le traitement d'une information massive par l'analyse réseau : méthodes, outils et applications,* Thèse de l'Université Aix-Marseille III, 14 janvier 1999.