LES RESEAUX COMME OUTILS D'ANALYSE EN BIBLIOMETRIE
Un cas d'application : les réseaux d'auteurs

E. BOUTIN, P. DUMAS*
H. ROSTAING, L. QUONIAM**

INTRODUCTION

L'objectif de cet article est de présenter un outil qui permet de visualiser automatiquement un réseau de relations. La notion de réseaux est bien connue des techniciens de la bibliométrie. L'utilité de ce concept, qui met en évidence les structures des relations existant entre différentes entités, a déjà été présentée dans de nombreux travaux. Ces réseaux peuvent tour à tour représenter l'activité de la co-citation (SMALL, 1973), les réseaux socio-techniques (CALLON, 1993), la structure thématique d'un domaine (DOU et alii, 1989) ou les collaborations scientifiques (PETERS et VAN RAAN, 1991) ...

Le point de départ de ces analyses bibliométriques est un ensemble de notices bibliographiques télédéchargées à la suite de l'interrogation d'une banque de données. Cette masse d'information possède la caractéristique d'être structurée en champs homogènes.

Un expert effectuant une lecture de cet ensemble de notices peut dégager les relations entre unités bibliographiques ce qui lui permet d'identifier la structure de l'ensemble étudié. Il prend pour point d'ancrage un ou plusieurs champs descripteurs de l'ensemble des notices considérées. Il constitue des dyades, des associations transitives pour arriver finalement à un réseau de relations.

Toutefois, quand le nombre de notices à examiner augmente, la lecture séquentielle de cette information s'avère fastidieuse et l'expert a du mal à se forger une idée objective et complète de la réalité qu'il cherche à appréhender. La construction automatique du tissu de relations entre les différentes unités bibliographiques devient alors indispensable à l'analyse d'une grande masse de données.

La construction automatique d'un réseau de relations considère un champ qui servira de base à la construction du réseau de relations. Cette construction du réseau prend la forme d'une cartographie. Sur cette cartographie peuvent être greffés des renseignements supplémentaires tirés des références bibliographiques (noms des laboratoires, mots-clés, date de publications, revues ...). On obtient ainsi une grille de lecture des documents primaires.

Après une présentation des principaux algorithmes utilisés pour obtenir cette cartographie, cet article présente une application de cette approche à la construction de réseaux d'auteurs.

1. Algorithmes utilisés pour construire la cartographie

Le point de départ de la méthode : les données traitées.

* Laboratoire Lepont, IUT de Toulon
BP 132 83957 LA GARDE CEDEX FRANCE
email Boutin@rhodes.univ-tln.fr
** CRRM, Centre de Recherche Rétrospective de Marseille
Centre Scientifique de St Jérôme,
13397 MARSEILLE CEDEX 20 FRANCE
email crrm@crrm.univ-mrs.fr

Le point de départ de la méthode consiste à extraire un champ unique pour l'ensemble des références à analyser. Ce champ servira de base à la construction du réseau. Il peut s'agir du champ auteur, code, mot-clé et plus généralement de tout champ composé potentiellement de plusieurs éléments bibliographiques non mutuellement exclusifs.

Dans le cas d'une analyse du champ auteur (AU), une notice bibliographique peut être symbolisée ainsi :

TI : ....
AU : A, B, C
AF : ....
....

Cette notice signifie ici une copublication entre trois auteurs A, B, C. Dans cet exemple, nous sommes en présence de trois relations : A est lié à B, A est lié à C et B est lié à C.

Pour construire le réseau des auteur de l'ensemble des notices bibliographi ques, il faut dans un premier temp faire l'inventaire de toutes les relation entre les auteurs présents dans ces no tices. Une fois cet inventaire effectué toutes ces relations entre auteurs son disposées sous forme de matrice carré symétrique. Dans cette matrice, les li bellés des lignes et des colonnes repré sentent les différentes formes constitu tives du champ étudié et le coeur de l matrice, les fréquences d'occurrence d chaque paire d'auteurs. Ces deux étape sont réalisées avec le logiciel DATAVIEW développé au CRRM (ROSTAING, 1993).

Le point d'arrivée de la méthode : l réseau de relations.

La méthode débouche sur un posi tionnement des auteurs reliés entre eu par des arcs non valués comme le mon tre la figure 1.

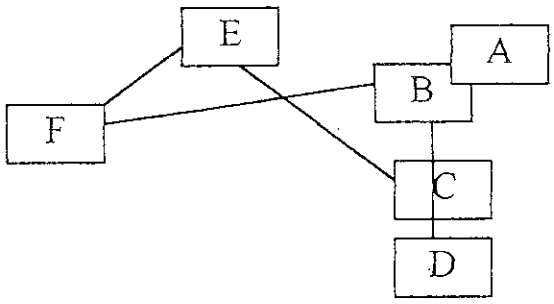

Figure 1 : Représentation des relations de copublication entre 6 auteurs.

Un arc entre deux auteurs signifie une présence des deux auteurs dans une même notice bibliographique. Dans l'exemple du réseau de la figure 1, l'arc entre A et B signifie une copublication entre A et B.

Le réseau final doit respecter certains critères d'esthétisme. On définira qu'un réseau est esthétique pour une personne qui va l'analyser si sa lisibilité correspond aux capacités cognitives de cet analyseur. Il est certain que tous les sujets n'ont pas la même capacité. Aussi avons nous laissé la possibilité à l'utili sateur de personnaliser son réseau, une fois la construction automatique réa lisée. L'esthétisme d'un réseau est ap précié à travers une fonction d'éva luation. On considèrera qu'un réseau est d'autant plus esthétique que :

◊ le nombre de chevauchements entre les étiquettes des auteurs est nul.
◊ le nombre de chevauchements entre

les arcs et les étiquettes des auteurs est faible.

◊ le nombre d'intersections entre les arcs est faible.

A titre d'illustration, le réseau présenté figure 1 ne satisfait pas pleinement les trois critères retenus : l'étiquette de l'auteur F chevauche celle de l'auteur C, les arcs reliant A à C et B à E créent une intersection, l'arc entre C et D recouvre l'étiquette de l'auteur E.

Pour un réseau donné, il est possible d'associer une valeur correspondant à une combinaison linéaire de ces trois critères. La pondération affectée à chacun d'entre eux sera fonction des préférences de l'analyseur. L'objectif est de minimiser cette valeur. Pour aboutir à ce résultat, le choix s'est porté sur des algorithmes itératifs qui font converger le réseau vers une situation jugée esthétiquement satisfaisante.

Les algorithmes utilisés :

L'objectif de la méthode étant d'exprimer des relations sous forme d'un réseau, plusieurs familles d'algorithmes sont potentiellement utilisables. On peut recourir aux algorithmes de calcul matriciel dans la mesure où le point de départ de la méthode est une matrice (BOUTIN et alii, 1995b). On peut utiliser également des résultats de la théorie des graphes (AHO et alii, 1987) dans la mesure où le point d'arrivée de la méthode est un graphe particulier appelé réseau. On peut également recourir à des algorithmes permettant de résoudre des problèmes n*p complexes : algorithmes génétiques (GROVES et MICHALEWICZ, 1990) et recuits simulés (DAVIDSON et HAREL, 1989) par exemple.

L'évaluation disjointe de ces différents algorithmes met en évidence, pour chacun d'eux une loi des rendements décroissants. Si on indique en abscisse le temps de traitement et en ordonnée la mesure de l'esthétisme du graphe, on obtient une courbe dont l'allure générale est représentée figure 2.

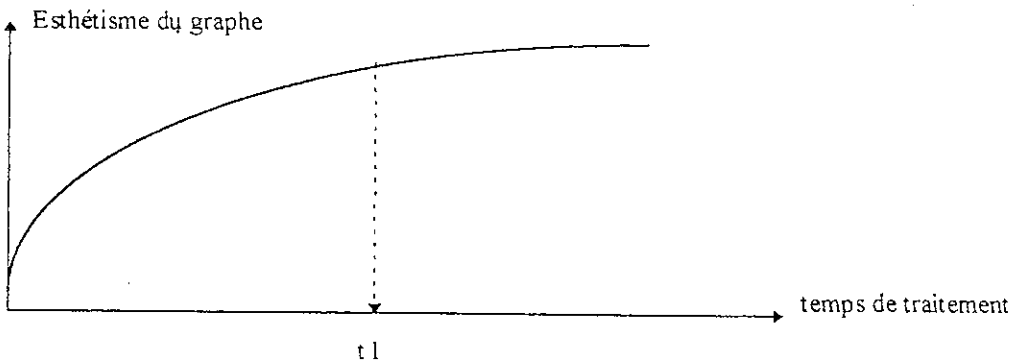

Figure 2 : Schéma général de comportement des algorithmes utilisés.

On voit sur cette courbe qu'à partir de t 1, un gain marginal d'esthétisme se paie au prix d'un temps de traitement rédhibitoire. Par ailleurs, l'enchaînement de ces algorithmes met en lumière des effets de synergie intéressants. Considérons deux algorithmes A et B que l'on exécute sur une durée t 1 (figure 3).

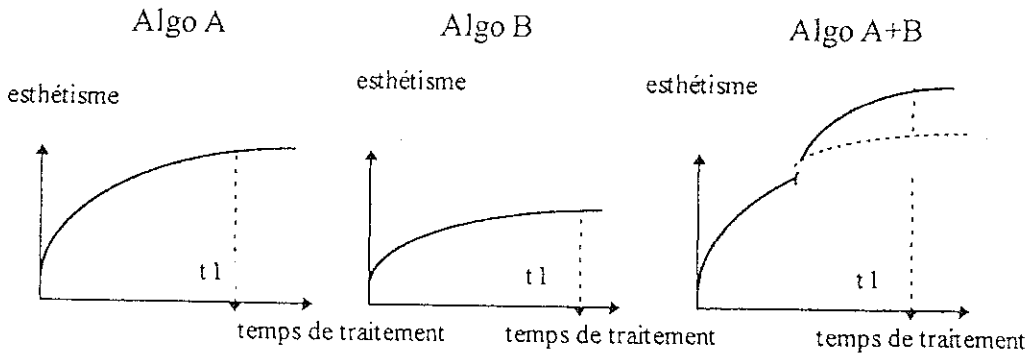

Figure 3 : Mise en évidence de l'effet de synergie.

L'enchaînement des deux algorithmes sur une durée t 1 donne un résultat esthétiquement supérieur à ce que rendrait une évaluation disjointe sur l'un ou l'autre algorithme. En d'autres termes, l'utilisation successive de A et de B permet de repousser la limite des rendements décroissants (BOUTIN et alii, 1995a). Un tel résultat ne se manifeste pas quel que soit l'enchaînement des algorithmes A et B. Il existe entre les deux séquences d'enchaînement A puis B ou B puis A une séquence esthétiquement supérieure.

Le travail présenté ici n'a pas pour objectif de se focaliser sur un algorithme particulier. Il s'agit plutôt d'arriver rapidement à un graphe esthétique en enchaînant plusieurs algorithmes.

Le premier algorithme appartient à la famille des algorithmes génétiques. Cet algorithme prend pour point de départ un positionnement aléatoire des sommets du réseau sur le plan. Cette situation initiale est appelée génération 1. Chaque sommet de cette génération 1 est évalué au travers de la fonction d'esthétisme présentée précédemment : plus sa valeur est forte, plus le point est "responsable" par son positionnement d'un nombre élevé de chevauche ments. La méthode consiste à conserve d'une génération à l'autre les sommet qui ont les valeurs les plus faibles et les combiner entre eux pour générer d nouveaux points qui remplaceront le sommets dont les valeurs sont les plu fortes. On aboutira ainsi à la génératio suivante. Par un processus itératif, o converge vers une solution esthétique ment supérieure.

Lorsque la phase des rendements dé croissants se manifeste, on recourt alor à un algorithme "tectonique". L'algo rithme tectonique prend pour géné ration initiale la génération finale d l'algorithme génétique. Chaque poin est alors évalué pour juger de s contribution à l'esthétisme du graphe L'algorithme a pour objectif de mo difier la position du point le moins bie placé eu égard à la fonction d'esthé tisme. Le nouveau positionnement d ce point définit une nouvelle généra tion qui sera analysée de la même ma nière.

Le choix de l'enchaînement géné tique puis tectonique se justifie pleine ment par le fait que l'algorithme tec tonique n'opère que de façon très marginale sur les points. Il ne peut

intervenir efficacement qu'une fois un prépositionnement opéré. L'objectif de l'algorithme génétique est d'effectuer ce prépositionnement.

2. Application de l'approche réseau à un corpus bibliographique

L'exemple présenté ci-dessous se propose de visualiser le réseau des auteurs travaillant dans le domaine de la bibliométrie au niveau mondial, et ceci après la consultation de la banque de données Pascal.(1)

La bibliométrie est un domaine de recherche qui a pour objectif de développer des outils et des méthodes visant à dégager d'un ensemble de notices bibliographiques un certain nombre de caractéristiques synthétiques destinées à fournir une grille de lecture des documents primaires. Pour en savoir plus sur cette discipline, nous recommandons la lecture d'un article rédigé par WHITE et MC CAIN car les auteurs ont effectué une bonne synthèse des recherches menées en bibliométrie et scientométrie (WHITE et MC CAIN, 1989).

La collecte des références bibliographiques a été effectuée à partir du CD-ROM Pascal 1984-1994 avec comme équation logique : "bibliometr? or scientometr? or informetr?". Le résultat de cette collecte est la constitution d'un corpus bibliographique de 1191 références.

Le premier travail a consisté à isoler le champ auteur de chaque référence bibliographique, puis à faire l'inventaire des relations entre auteurs pour finalement présenter ces relations sous forme matricielle.

(1) Banque de données produite par l'INSIT. Elle a l'avantage d'être une source d'information multidisciplinaire et internationale pour les domaines des sciences exactes.

L'application, sur cette matrice, des algorithmes présentés en première partie permet de construire plusieurs types de réseaux qui sont présentés de façon successive.

Sur le réseau de départ de la figure 4, ne sont représentés, par souci de clarté que les auteurs qui ont une fréquence de publication supérieure ou égale à 3. Chaque auteur est symbolisé par un sommet du réseau. Chaque sommet est représenté par une étiquette qui contient le nom de l'auteur et le nombre d'articles qu'il a publié. Un lien n'est visualisé qu'à partir du moment où les deux auteurs correspondant ont publié ensemble au moins deux articles.

Ce réseau fait apparaître un certain nombre d'auteurs isolés compte tenu des restrictions émises précédemment. Toutefois, on peut identifier un certain nombre de groupes correspondant aux différentes équipes de recherche dans le domaine. L'ensemble est appelé métastructure du réseau. Il faut noter que la proximité entre groupes n'a aucune signification ni justification, si ce n'est l'esthétisme de la représentation globale. En effet, à aucun moment dans l'analyse n'intervient de notion de distance entre les auteurs. Pour une matrice de relations, il existe une quasi-infinité de réseaux esthétiquement équivalents qui permettent de la représenter graphiquement. Le choix de l'un ou l'autre s'effectuera de façon arbitraire.

L'analyse de ce réseau peut s'effectuer en répondant à quelques questions (DEGRENNE et FORSE, 1994) :

◊ Au sein de chaque groupe, l'organisation entre auteurs est elle une organisation plutôt centralisée ou partagée ? Ce type de réflexion permet de rendre compte de la politique des laboratoires en matière de publications scientifiques.

◊ Les groupes formés correspondent-ils à des centres de recherche identifiables ? Ce type d'interrogation permet de mettre en évidence des collèges invisibles d'auteurs qui, sans faire partie du même laboratoire, n'en publient pas moins ensemble.

Toutefois, ce réseau très général ne représente pas la réalité dans toute sa complexité. Celle-ci ne peut être appréhendée que lorsque l'on effectue un zoom sur une certaine portion du réseau. C'est ce qui est suggéré par l figure 5. Ce zoom s'effectue en ajoutar certains critères pendant la constructio du réseau de relations. Ce réseau v sualise en effet sans restriction de fr quences les relations de copublicatio entre les auteurs qui ont publié e langue française. Sur ce réseau, le auteurs qui ont plus d'une publicatio sont représentés en utilisant une taill de caractères plus grande. On retrouv la structure par groupe de la figure mais celle-ci est moins apparente.

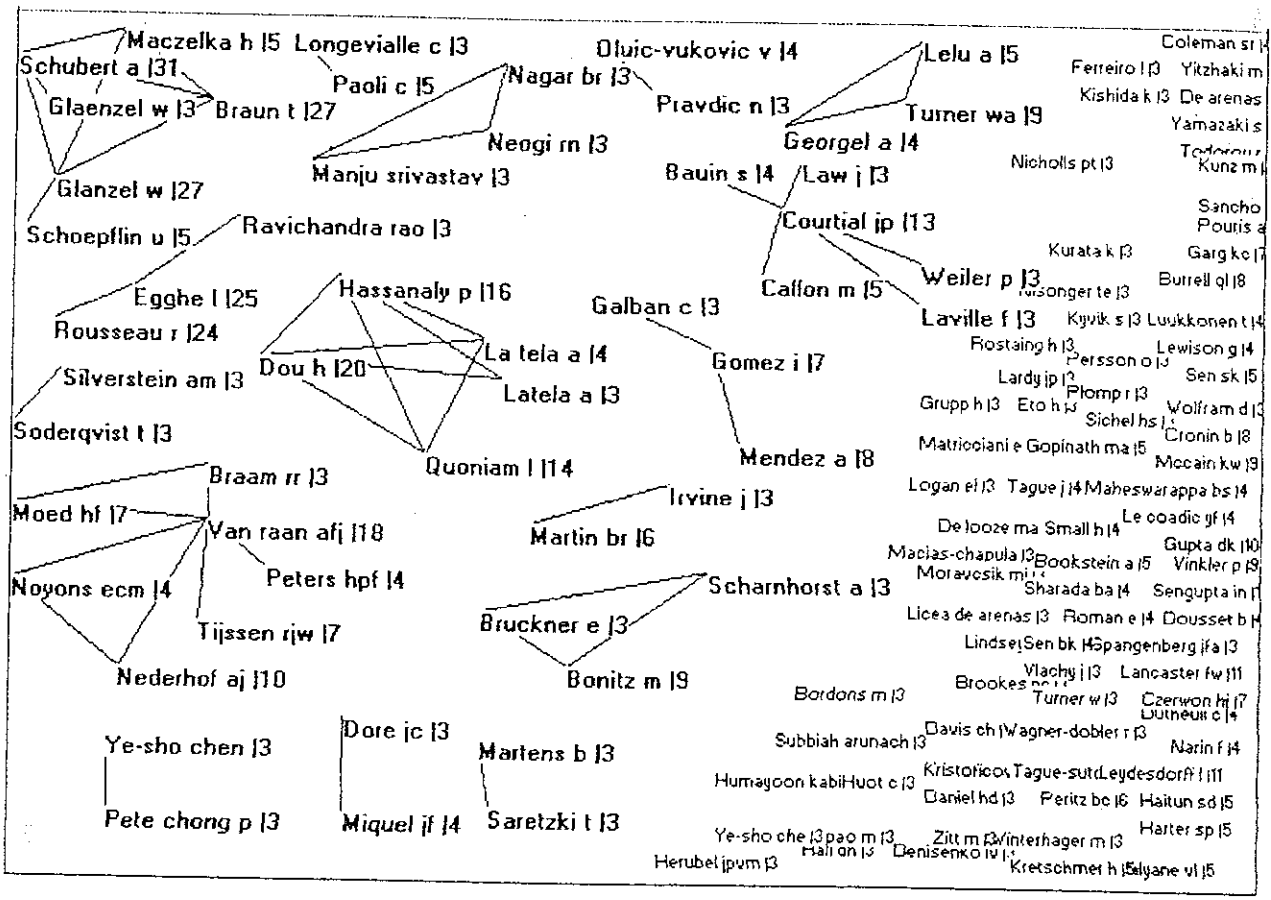

Figure 4 : Construction du réseau d'auteurs.

On voit poindre dans cette figure 4 le rôle de pivots joué par certains auteurs. Ces pivots appelés aussi isthmes sont représentés dans un rectangle encadré. Si on enlève ces auteurs, on déconnecte à chaque fois le groupe en plusieurs sous-groupes. Ces auteurs particuliers se situent à la frontière de plusieurs sous-groupes de recherche. Ils sont les points de passage obligé pour rapprocher les recherches des sous-groupes. On peut définir de la même façon la notion de x isthme. Dans la figure 5, PAOLI, DIONNE, LA TELA constituent un 3 isthme. Il faut enlever en effet simultanément ces trois auteurs pour disconnecter le groupes en deux sous-groupes.

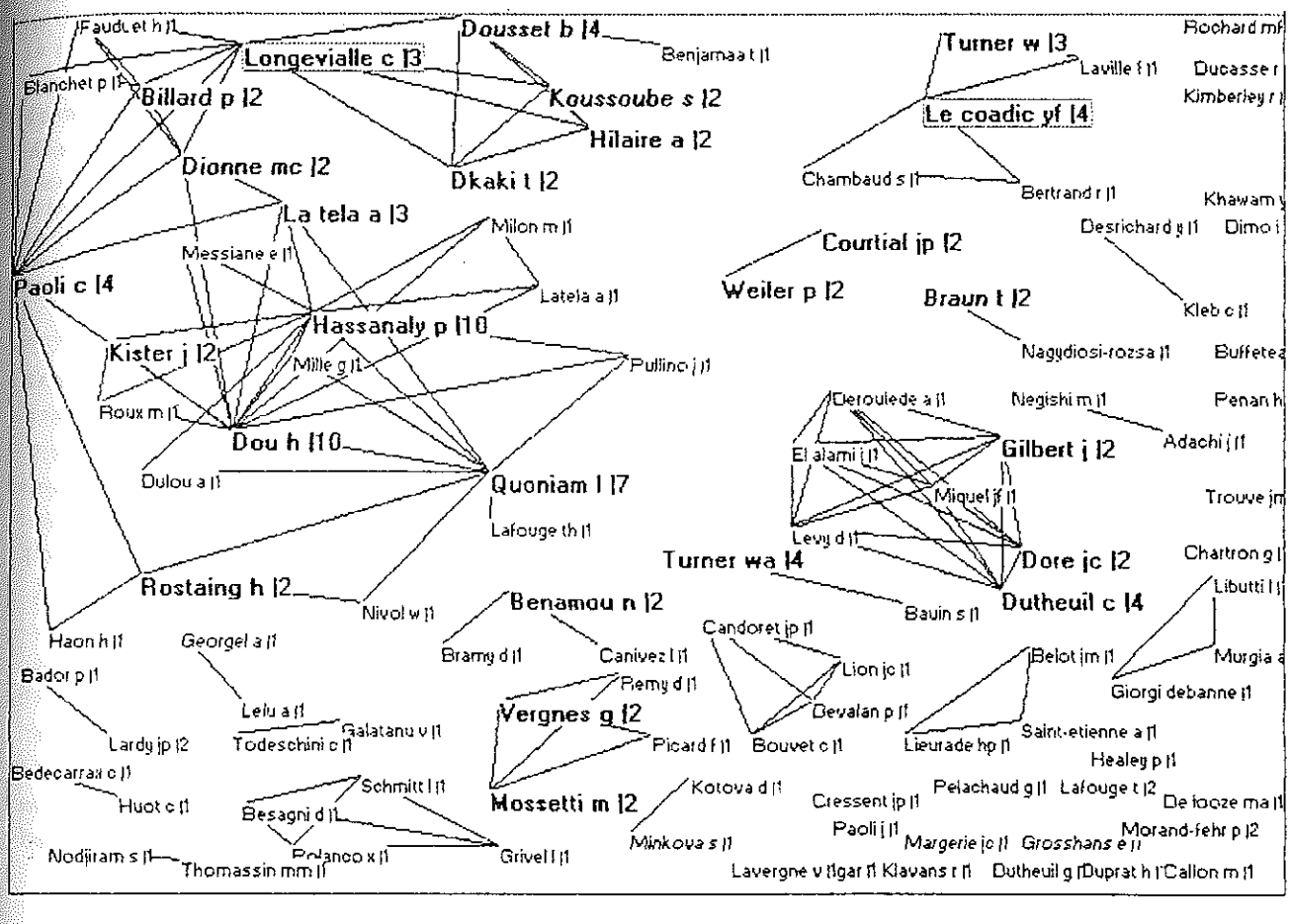

Figure 5 : Réseau des auteurs ayant publié en langue française.

Le même principe de "zoom" peut nous conduire à visualiser les relations de copublication des auteurs qui ont publié dans une revue particulière. A titre d'illustration, le réseau de la figure 6 représente les relations de copublication des auteurs dans la revue *Scientometrics* spécialisée dans le domaine de la bibliométrie et de la scientométrie. Ce réseau est intéressant en soi mais aussi en comparaison avec le réseau initial présenté figure 4.

Plusieurs questions peuvent là encore servir de piste à l'interprétation :

◊ La littérature publiée en langue française sur le sujet permet-elle de rendre compte de l'activité de recherche mondiale sur le sujet ? Pour répondre correctement à cette question, il faudrait étudier également les thèmes de recherche à travers l'analyse du champ mots-clés.

◊ Les équipes publiant en français sur le sujet ont-elles une reconnaissance internationale ?

Une dernière approche peut être présentée. Celle-ci fait du réseau de la figure 4 le point d'arrivée d'une évolution dynamique qui a conduit ce réseau à se structurer de cette façon. Pour ce faire, un découpage arbitraire en trois périodes a été obtenu à la suite d'un découpage temporel : 1984-1990, 1990-1992, 1992-1994. Ceci nous permet de représenter trois réseaux. Le réseau correspondant à la période 1984-1990, le réseau correspondant à la période 1984-1992 et le réseau correspondant à l'ensemble de la période étudiée.

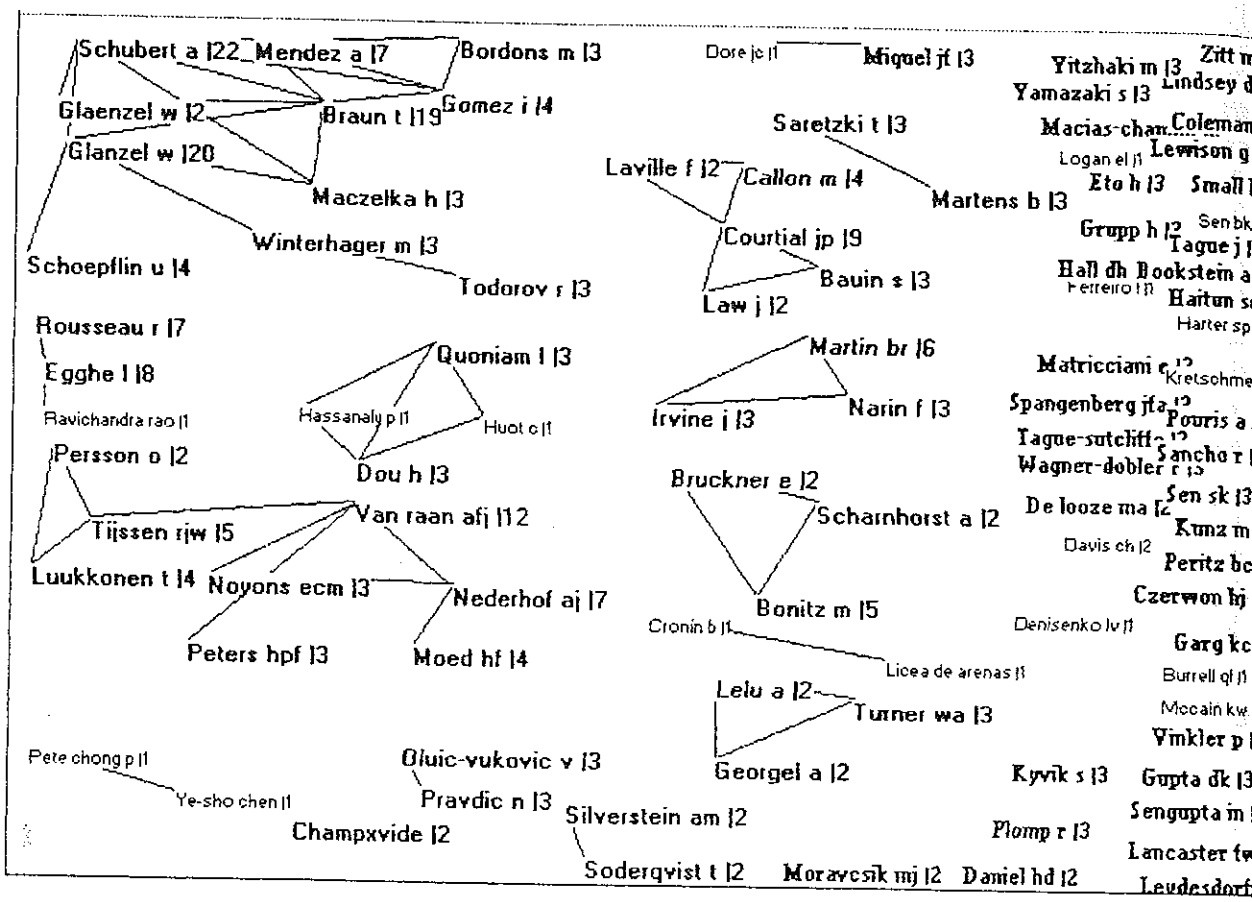

Figure 6 : Réseau des auteurs ayant publié dans la revue Scientometrics.

Ces réseaux sont présentés successivement figure 7, 8, 9. Les auteurs figurant dans des étiquettes encadrées, sont des auteurs nouveaux entrants dans le domaine. Ce type d'approche permet de ne pas rester à une vision statique de la réalité mais permet d'intégrer une vision dynamique des choses.

L'analyse de cette dynamique peut être effectuée en répondant à quelques questions :

◊ Quels sont les pères fondateurs du domaine ?
◊ Y a-t-il une forte "mortalité" dans ce domaine se traduisant par la disparition de certaines équipes ? Pour répondre parfaitement à cette question, il aurait fallu procéder à des découpages de temps totalement disjoints.
◊ Y a-t-il une forte "natalité" dans c domaine se traduisant par l'arrivé de nouvelles équipes ?
◊ Y a-t-il une tendance à la concen tration des groupes ou à la dispe sion, à l'isolement ou à la collabora tion des chercheurs ?

REFLEXION SUR LA VALIDITE D L'APPROCHE RESEAU

La construction automatique de ré seaux fait apparaître un certain nombr de limites.

Si le réseau de relations est construi à partir de références télédéchargées d banques de données, deux limite apparaissent suivant que l'on se situ en amont ou en aval de l'algorithme D'une part, la qualité du réseau es tributaire de la qualité de l'informatio livrée par le producteur de la banque d

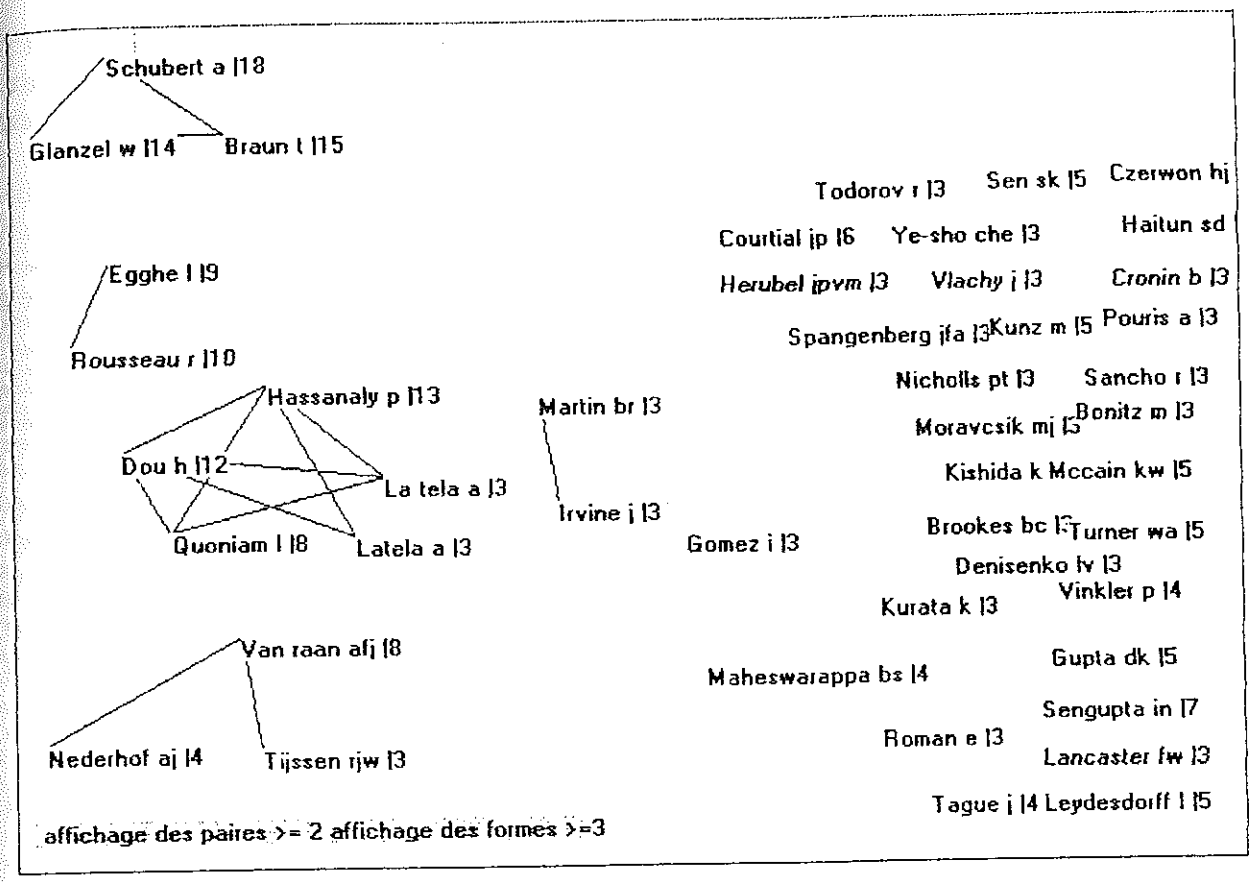

Figure 7 : Réseau d'auteurs correspondant à la période 1984-1990.

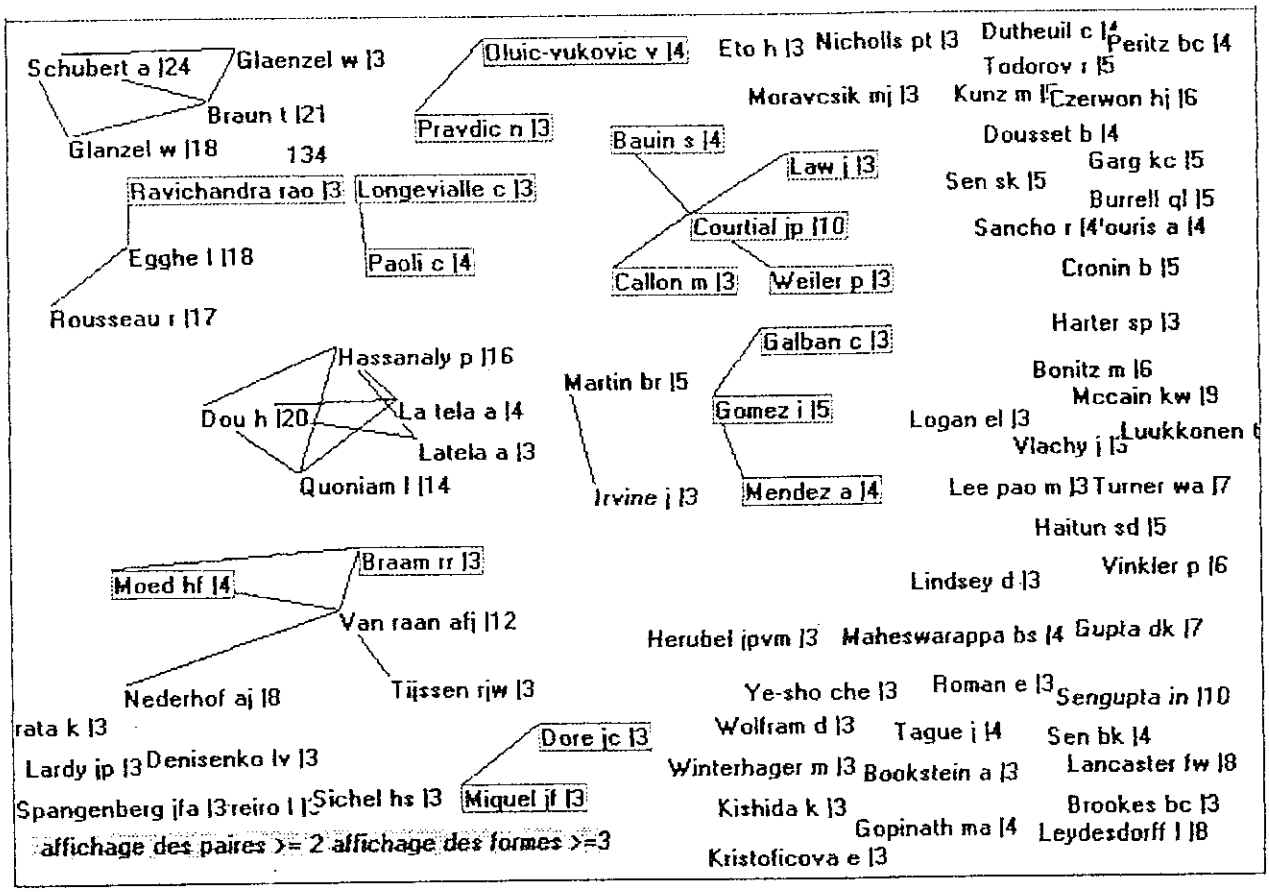

Figure 8 : Réseau d'auteurs correspondant à la période 1984-1992.

données. Le réseau de la figure 8 exprime bien ce genre de préoccupation. Les auteurs "LA TELA" et "GLANZEL" sont orthographiés de deux façons différentes dans les références d'origine. Ce problème peut être résolu en amont pa un prétraitement du champ auteu avant analyse bibliométrique : la nor malisation des noms d'auteurs par re formatage.

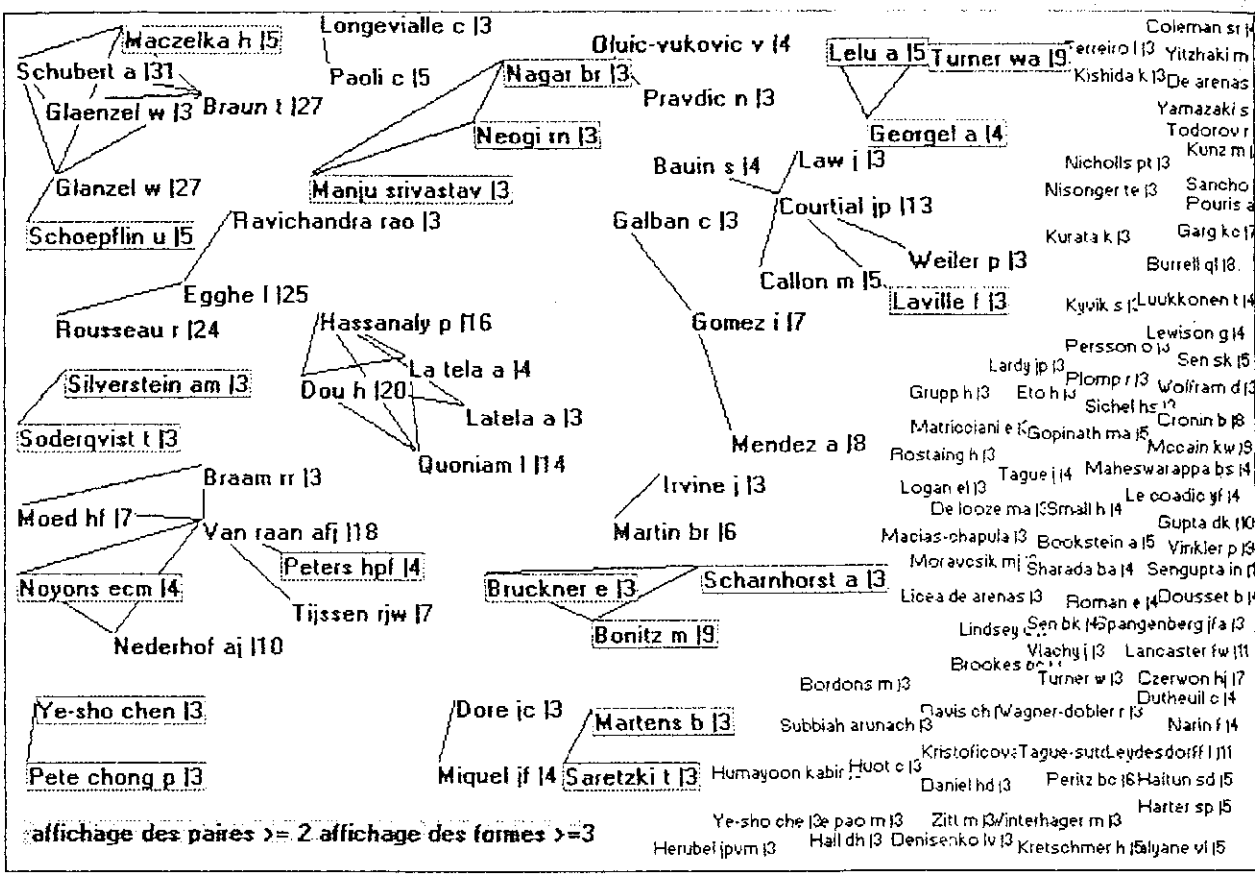

Figure 9 : Réseau d'auteurs correspondant à la période 1984-1994.

Nous n'avons volontairement pas pratiqué ce retraitement pour faire ressortir cette limite. D'autre part, dans l'interprétation des résultats, il peut y avoir confusion entre l'hypothèse et la conclusion. L'outil réseau survalorise en effet le groupe au détriment de l'individu. Si on examine la position sur la figure 9 des deux auteurs LEYDESDORFF et NEDERHOF, on remarque que NEDERHOF se trouve survalorisé par le fait qu'il appartient à un groupe. Cette limite peut être partiellement contournée si on introduit des tailles d'étiquettes proportionnelles au nombre de publications. L'information fournie par le réseau et les fréquences est plus riche que la seule fréquence qui est le seul critère actuellement reconnu dans les travau scientifiques sur le sujet.

CONCLUSION

L'article a donc présenté un out assurant la construction automatiqu de réseaux et en a montré l'intérêt travers une application parlante pou des personnes spécialistes dans domaine de la bibliométrie et de scientométrie.

Cet exposé fait ressortir un certai nombre de qualités de l'analyse.

D'une part, cet outil est applicable un grand nombre de domaines alla

d'un traitement bibliométrique à la construction de réseaux de communication interne.

D'autre part, cette approche se caractérise par son caractère réaliste. Il n'y a pas de déformation de la réalité comme c'est parfois le cas dans les méthodes d'analyse de données, ni utilisation de métriques.

Enfin, le caractère visuel des réseaux obtenus rend l'outil facilement appropriable par le décideur qui peut dans tous les cas in fine faire glisser les boîtes pour personnaliser son réseau.

BIBLIOGRAPHIE

◊ A. AHO, J. HOPCRAFT, J. ULLMAN, *Structure des données et algorithmes*, Interéditions, France, 1987.

◊ E. BOUTIN, P. DUMAS, L. QUONIAM, H. ROSTAING, H. DOU, (1995a), *Génération automatique de réseaux en bibliométrie*, Actes du colloque "les systèmes d'informations élaborées", Ile Rousse, 30 mai - 6 juin 1995.

◊ E. BOUTIN, L. QUONIAM, H. ROSTAING, H. DOU, (1995b), *A new approach to display real co-authorship and co-topicship through network mapping*, Acte du colloque "Fifth International Conference on scientometrics & infometrics", Chicago, 7 - 10 juin, 1995.

◊ M. CALLON, J.P. COURTIAL , H. PENAN, *La scientométrie*, Edition Presses universitaires de France, Paris, 126 p., 1993.

◊ R. DAVIDSON, D. HAREL, *Drawing graphs nicely using simulated annealing*, Technical Report CS89-13, Department of Applied Mathematics and Computer Science, The Weizmann Institute, Rehovot, Israel, July 1989.

◊ A. DEGRENNE , M. FORSE ,*Les réseaux sociaux*, Editions Armand Colin, 1994.

◊ H. DOU, P. HASSALANY, L. QUONIAM, *Infographics analytical tools for decision makers*, Scientometrics, Vol 17, n° 1-2, p. 61-70, 1989.

◊ L. GROVES, Z. MICHALEWICZ, P. ELIA et C. JANIKOW, *Genetic algorithms for drawing directed graphs*, Methodologies for intelligent Systems, 5, p. 268-276, 1990.

◊ H.P.J. PETERS, A.F.J. VAN RAAN, *Structuring scientific activities by co-author analysis. An exercise on a university faculty level*, Scientometrics, Vol 20, n° 1, p. 235-255, 1991.

◊ H. ROSTAING, *Veille technologique et bibliométrie : concepts, outils et applications*, Thèse : Aix-Marseille III, 353 p., 13 janvier, 1993.

◊ H.G. SMALL, *Co-citation in the scientific literature : a new measure of the relationship between two documents*, Journal of the American Society for Information Science, Vol 24, n° 4, p. 265-269, 1973.

◊ H. WHITE , K. MC CAIN, *Bibliometrics*, Annual review of information Science and Technology, Vol 24, p. 119-186, 1989.

* * *