METHODOLOGIE RELATIONNELLE D'EXTRACTION DE CONNAISSANCES A PARTIR DE DONNEES PROVENANT D'UN FORUM DE DISCUSSION

Eric Boutin,
Maître de Conférences en Sciences de l'information et de la communication
laboratoire Le pont
boutin@univ-tln.fr , + 33 (0)4 94 14 23 56

Adresse professionnelle
Université de Toulon-Var ★ BP 132 ★ F-83957 La Garde Cedex

Résumé : L'information contenue dans un forum de discussion est la meilleure et la pire des choses : meilleure en ce qu'elle peut porter le germe d'informations qui se situent encore au niveau des signaux faibles. Pire en ce qu'elle consiste bien souvent en des informations dont il est difficile de mesurer le degré de fiabilité et de validité.
Actuellement, l'offre logicielle en matière d'exploitation de l'information des fora de discussion se limite souvent à des approches essentiellement américaines universitaires à l'état de prototype.
Après avoir décrit la structure des données brutes d'un forum de discussion et l'état de l'art dans le domaine des outils d'analyse des forums de discussion, cet article propose la mise en place d'une chaîne de traitement, aujourd'hui semi automatisée, débouchant sur des cartographies relationnelles. Nous montrerons l'intérêt que présente l'analyse relationnelle dans la mesure de la centralité des acteurs du forum et la fiabilisation du corpus de données.

Summary :.
Webgroups information is as helpful as it can be harmful: important information may be present at a preliminary state, but most of the information is hardly measurable in terms of reliability and truthworthiness.
At the moment, webgroups data analysing softwares are prototypes supplied by American Universities.

In this article we describe the primary information structure of webgroups, introduce current webgroups data analysing softwares and propose a semi-automated data processing chain generating relational maps. We highlight the importance of relational analysis which clearly displays key-actors and the reliability of the information.

Mots clés :.forum de discussion, analyse relationnelle, information stratégique, état de l'art

Keywords :.groups, relational analysis, strategic information

Méthodologie relationnelle d'extraction de connaissances à partir de données provenant d'un forum de discussion

Les fora de discussion sont le lieu d'échange d'informations informelles et émergentes par une communauté de colistiers qui partagent des centres d'intérêt. Les problématiques couvertes touchent aux domaines scientifiques, techniques mais aussi économiques et sociétaux et peuvent alimenter un processus de veille.
L'information contenue dans un forum de discussion est un peu, comme la langue d'Esope, la meilleure et la pire des choses : meilleure en ce qu'elle peut porter le germe d'informations émergentes. Pire en ce qu'elle consiste bien souvent en des informations dont il est difficile de mesurer le degré de fiabilité et de validité. Savoir exploiter l'information d'un forum de discussion c'est pouvoir discerner l'information pertinente de celle qui ne l'est pas, c'est saisir des opportunités d'informations qui se situent encore au niveau des signaux faibles mais c'est aussi éviter la menace que pourrait provoquer une rumeur en la désamorçant dès son apparition.
Actuellement, l'offre logicielle en matière d'exploitation de l'information des forums de discussion se limite à des approches essentiellement américaines universitaires à l'état de prototype. En France, des outils tels Wordmapper[1] proposent une chaîne de traitement reposant sur l'analyse de contenu qui peut s'adapter au traitement d'un ensemble de fils de discussion d'un forum de discussion.

Après avoir décrit la structure des données brutes d'un forum de discussion et l'état de l'art dans le domaine des outils d'analyse des forums de discussion, cet article propose la mise en place d'une chaîne de traitement, aujourd'hui semi automatisée, débouchant sur des cartographies relationnelles. Nous montrerons l'intérêt que présente l'analyse relationnelle dans la mesure de la centralité des acteurs du forum et la fiabilisation du corpus de données.

Le travail qui est ici présenté a été appliqué à divers fora de discussion. Nous avons volontairement souhaité rester au niveau méthodologique pour deux raisons :

- Des raisons liées à la confidentialité des données nous interdisent de produire des cartes ou figureraient des noms en clair.
- Dans l'étalonnage concurrentiel que nous avons mené, il n'a pas été toujours possible de confronter le même jeu de données à la batterie des analyses possibles.

Le corps du texte est organisé autour de trois parties.
La première décrit l'information brute contenue dans un forum de discussion.
La seconde partie présente un état de l'art des outils d'analyse d'un forum de discussion.
La troisième partie illustre et positionne la méthodologie d'analyse relationnelle que nous mettons en œuvre.

L'INFORMATION BRUTE ISSUE D'UN FORUM DE DISCUSSION

L'information élémentaire d'un forum de discussion est le *message.* Il est soit envoyé spontanément par son émetteur soit émis en réponse à un message préalablement existant.
Ces informations élémentaires sont regroupées en *fils de discussion* (threads). Un fil de discussion va ainsi correspondre à l'ensemble des messages suscités par un message de départ. Les fils de discussion peuvent être longs, un message suscitant une ou plusieurs réponses, chacune d'elles pouvant à son tour entraîner des réactions.

Sur la forme, l'information contenue dans un forum de discussion est structurée autour d'un certain nombre de champs. Nous avons extrait *Figure 1*, à titre d'illustration, un échange simple de type question-réponse. La question, posée par un acteur du forum, est structurée autour :

- du numéro de la question
- de l'identification de l'émetteur
- de la date d'émission de la question
- du titre de la question à travers le champ objet (pas très significatif dans ce cas là)
- du contenu de l'intervention

La réponse à la question est formatée autour des mêmes champs.

[1] développé et commercialisé par la société Grimmer logiciels

Pour pouvoir mener à bien une analyse relationnelle, il est essentiel de pouvoir avoir une traçabilité parfaite des messages : il est important de faire la distinction entre une intervention nouvelle et un message de réponse. Dans ce cas, il est nécessaire de pouvoir identifier précisément à quelle question l'intervention se rattache.

Pour automatiser ce processus de reformatage des données, nous avons considéré qu'une intervention est une réponse à une intervention précédente lorsque :

Le titre de l'intervention comporte Re : suivi du titre de l'intervention d'origine Ou Le contenu du message comporte le contenu du message initial quelque part dans le message précédé ou pas du signe « > » en début de chaque ligne.

Le processus automatisé nécessite encore pourtant l'intervention humaine dans certains cas ambigus pour reconstituer le fil de discussion.

Figure 1 : exemple de structuration de l'information d'un forum de discussion

ETAT DE L'ART DANS LE DOMAINE DES OUTILS D'ANALYSE DES FORA DE DISCUSSION

Un recherche bibliographique sur le sujet et une exploration du web nous ont permis d'identifier un certain nombre d'outils offrant une vision synthétique et/ou cartographique de l'activité d'un forum de discussion.

Ces outils ont souvent une finalité de représentation macroscopique de l'activité du forum à visée classificatoire. Nous avons choisi de distinguer ces outils selon les technologies qu'ils utilisent pour opérer leur classification. Deux technologies polaires peuvent alors être observées même si certains outils les combinent.

- Les technologies d'analyse de contenu exploitent le texte présent dans les fils de discussion. Ces analyses débouchent sur l'identification des différentes thématiques du corpus.
- Les technologies d'analyse relationnelle reconstituent l'interaction entre les acteurs du forum. Cette famille d'approche est centrée sur la reconstitution puis l'analyse du réseau constitué par l'ensemble des interventions sur le forum.

Nous allons dans les lignes qui suivent illustrer chaque technologie à l'aide de deux exemples. Il est à noter que nous n'avons pas pu appliquer un jeu de données homogène à ces divers outils. En effet, ces outils sont souvent à l'état de prototype et dans le meilleur des cas, on peut travailler sur un jeu d'essai qu'il n'est pas possible de personnaliser. Le lecteur intéressé par d'autres outils d'analyse pourra se renvoyer aux travaux de Smith M. [Smith,1997], Donath Judith S. [Donath.1995].

Les technologies d'analyse de contenu

Nous allons donner deux exemples d'outils utilisant cette technologie.

Websom est un outil développé par une équipe finlandaise du Neural networks Research Centre (NNRC) [Teuvo Kohonen et alii, 2000]

Cet outil débouche sur une cartographie qui positionne en regard les diverses thématiques d'un forum de discussion. La figure 2 illustre le type de résultat obtenu par Websom. Les labels voisins dans une zone lumineuse correspondent à une thématique proche. Les zones sombres sont des trous noirs entre deux espaces lumineux. Cette application rend possible une navigation dans un corpus de données de type forum à partir d'une première vision macroscopique qui peut être affinée en utilisant une fonction de zoomage pour accéder à un niveau de détail plus fin.

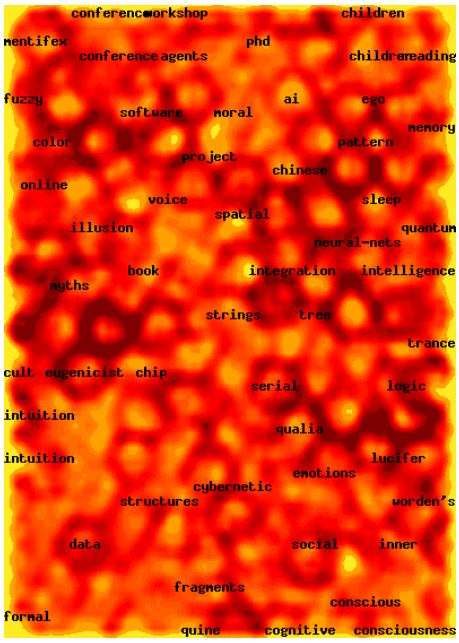

Figure 2 : résultat obtenu sous Websom

L'outil Wordmapper, développé dans une logique commerciale, est un autre exemple de logiciel d'analyse d'information provenant d'un forum qui utilise une technologie d'analyse de contenu. L'outil Wordmapper est un logiciel d'analyse et de représentation d'information complexe. Il n'est pas uniquement destiné à exploiter l'information provenant d'un forum de discussion, le traitement de l'information issue des fora de discussion faisant l'objet d'un patch spécifique.
Utilisant l'analyse des mots associés, cet outil est capable d'identifier les différentes thématiques d'un corpus de texte : dans le cas de la figure 3, l'analyse débouche sur des cartes qui permettent d'identifier les diverses composantes d'un forum.

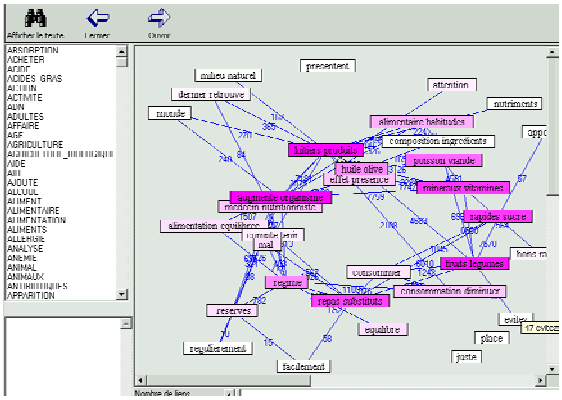

Figure 3 : résultat obtenu sous Wordmapper

Les technologies d'analyses relationnelles

Les outils qui utilisent cette technologie exploitent l'interaction existant entre les acteurs du forum. L'objectif de ces outils n'est pas de donner une vision d'ensemble des thématiques du forum mais de regrouper les acteurs qui interviennent sur les mêmes fils de discussion. Ces associations peuvent correspondre à des regroupements thématiques mais ils ne sont pas construits en analysant le texte des interventions mais le réseau construit à partir du jeu des questions réponses des acteurs.

Là encore, nous allons illustrer ces technologies par deux exemples issus de notre étalonnage concurrentiel.

Pdgarden est un outil développé par Rebecca Xiong, du Media Lab du MIT [Xiong, 99a, Xiong, 99b]. L'objectif de cet outil graphique métaphorique, dont une capture d'écran est donnée Figure 4, est de permettre à un nouvel arrivant sur un forum de discussion de se faire une idée des interactions sur ce forum à travers une représentation graphique des divers acteurs qui y sont intervenus. Chaque co-listier du forum est représenté par une fleur, l'ensemble des fleurs constituant un jardin. La position des pétales permet de dresser une typologie des acteurs du forum. Ce type d'application permet de connaître le nom des acteurs impliqués dans le forum ainsi que ses experts.

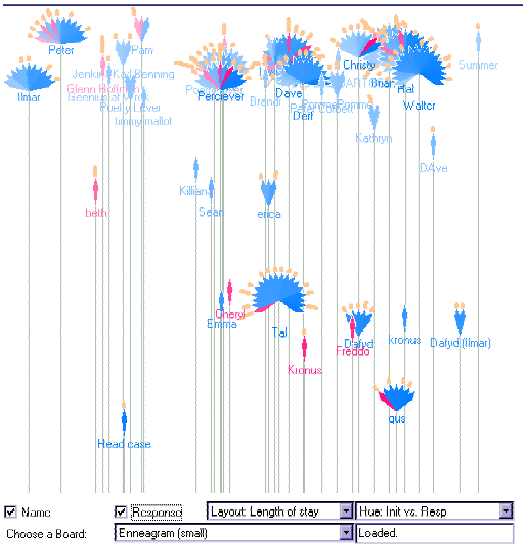

Figure 4 : résultat obtenu sous pdgarden

Webfan est un outil également développé par Rebecca Xiong du Media Lab du MIT. L'objectif de cet outil graphique est de représenter une synthèse des fils de discussion intervenus sur un forum. Il est alors possible d'identifier les centres d'intérêt de chaque co-listier et de connaître les problématiques brulantes. Le graphe se présente sous forme d'un demi cercle du centre duquel partent des fils de discussion représentés par des arbres. Chaque acteur du forum est représenté par une couleur différente.

Il existe un site web accessible à l'adresse suivante :

http://smg.media.mit.edu/%7Ebecca/webfan/. Une simulation peut être faite à partir d'un exemple précalculé. Nous en avons issu la capture présentée figure 5.

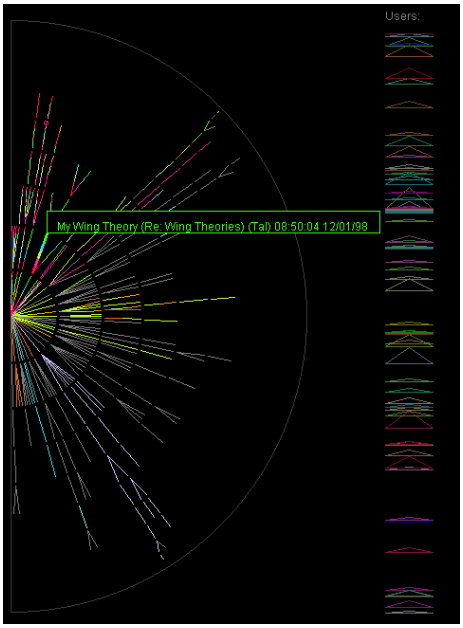

Figure 5 : résultat obtenu sous webfan

Les solutions mixtes

Certains outils développés utilisent l'approche relationnelle et de contenu et présentent ainsi une vision enrichie de l'activité d'un forum de discussion.

On peut illustrer cette démarche par l'exemple de *conversation map* développé par Warren Sack, du Social Technologies Group SIMS UC Berkeley [Warren Sack, 2001].
Conversation map est un navigateur de forum de discussion. Lorsqu'on se connecte au forum, l'outil analyse les interactions entre les divers messages échangés et affiche une interface graphique. Cette interface permet de comprendre les relations sémantiques et sociales entre les divers fils de discussion. L'outil se compose de 4 parties que l'on peut visualiser sur l'exemple présenté figure 6 : en haut à gauche se trouve le réseau des interactions entre co-listiers. En haut à droite se trouve le réseau sémantique des thématiques du forum. Au centre du document la partie Thèmes définit les principaux centres d'activité du forum. Dans la partie sud de la représentation, on retrouve les différents fils de discussion.

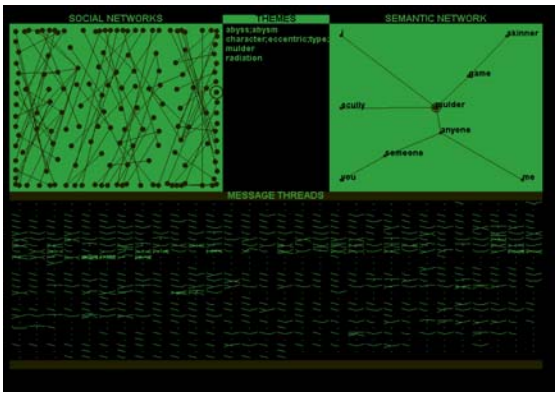

Figure 6 : résultat obtenu sous conversation map

VERS UNE ANALYSE RELATIONNELLE RENOUVELEE DES FORA DE DISCUSSION

L'analyse que nous appliquons au forum de discussion est issue de recherches réalisées dans le domaine de l'analyse des réseaux sociaux [Boutin, 1999]. Elle repose sur une analyse relationnelle spécifique qui mérite d'être positionnée par rapport aux exemples que nous avons présentés précédemment.

L'analyse relationnelle que nous envisageons débouche sur deux familles de cartographies

De la cartographie de départ à l'application de filtres

Le premier réseau qui vient à l'esprit est un réseau dans lequel chaque sommet correspond à un acteur du forum : un lien orienté entre A et B signifie que A a répondu à une question émise par B. Le réseau s'alimente à chaque nouvelle contribution sur le forum.

Considérons l'exemple pédagogique ci-dessous dans lequel on s'intéresse à 4 fils de discussion.

B pose une question à laquelle A, D et E répondent, C répondant lui même à A

F pose une question à laquelle D répond

A pose une question à laquelle répondent C et D

Les interactions entre les acteurs de ce forum peuvent être représentées par le réseau de la figure 7.

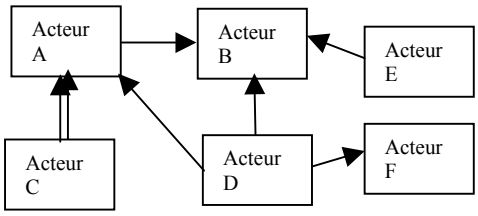

Figure 7 : exemple de réseau entre acteurs d'un forum

A partir du moment ou on s'intéresse à un volume d'échanges conséquent, le problème de la lisibilité des données se pose : la représentation des interactions entre les acteurs conduit à des graphes inextricables impossibles à exploiter. Le réseau présenté figure 8 est un extrait d'un exemple réel dans lequel nous avons étudié les interactions entre 178 intervenants d'un forum de discussion. Les noms des acteurs ont été anonymés. Le graphe de la Figure 8, réalisé à partir du logiciel matrisme [Boutin, 1999], illustre bien cette complexité. Chaque sommet du réseau est un acteur du forum. L'acteur est identifié par son numéro et décrit par le nombre d'interventions qu'il a réalisé sur la période considérée. L'acteur 73 présent en haut à gauche du réseau est par exemple intervenu 5 fois.

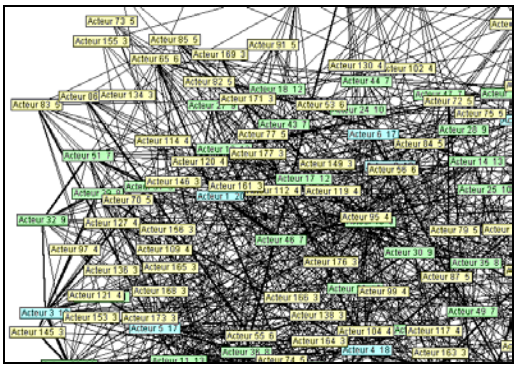

Figure 8 : exemple d'interactions entre les acteurs d'un forum de discussion

Ce type de représentation est inexploitable si elle n'est pas accompagnée de filtres puissants qui vont permettre d'apporter différents éclairages à la réalité complexe que nous souhaitons décrire.
Nous pouvons, par exemple, opter pour un filtre qui privilégie la redondance. On observe, dans le graphe figure 7, qu'à deux reprises C répond à A. Cette redondance peut être soulignée et servir de clé de filtre. Dans ce cas, seuls les interactions réalisées deux fois au moins seront retenues ce qui conduit à une simplification du graphe.
Nous avons choisi d'appliquer ce type de filtre aux données de l'exemple réel anonymé. Si on représente uniquement les liens entre les acteurs du réseau lorsque ces liens sont présents au moins deux fois et si on se restreint aux acteurs intervenus au minimum 5 fois dans le forum, on obtient le réseau simplifié présenté figure 9.

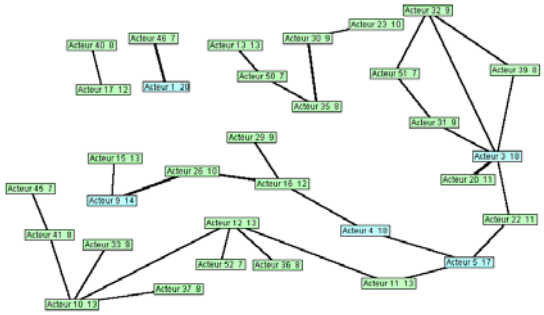

Figure 9 : Exemple de réseau après filtrage sur les paires

Le graphe présenté figure 9 restitue à lui seul 40% des échanges totaux contenus dans le forum de discussion. Il est à noter que ces échanges sont réalisés entre 23 acteurs soit 12% des acteurs du forum. On a donc ainsi accès à une information de synthèse qui doit nous orienter vers la lecture sélective de certaines interventions.

Un autre filtre pourrait consister à éliminer du graphe tous les acteurs isolés ou périphériques dont les contributions ne s'intègrent pas dans le cœur des débats du forum de discussion. Si on itère ce filtre, on obtient le noyau des intervenants. Piliers du forum, ces acteurs sont une dizaine à être intervenus pour échanger avec au moins 8 autres acteurs du noyau. La figure 10 présente le noyau du réseau au sens de la théorie des graphes.

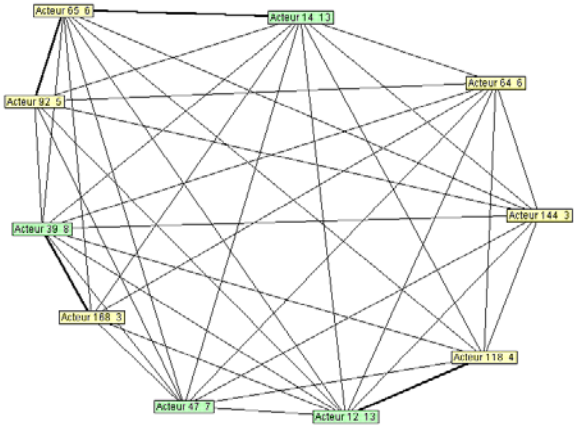

Figure 10 : le noyau dur des acteurs de ce réseau

Cartographie représentant les relations de citation entre acteurs

Il est possible, en transposant l'analyse de la citation et l'analyse du couplage bibliographique [Egghe, 90], [Garfield, 83] au domaine qui nous préoccupe de représenter des cartographies relationnelles dont la signification est différente. Ces cartographies sont construites sur le principe suivant. Supposons que A émette une question et que A, B, C y répondent. Jusqu'à présent nous avions construit une structure réticulaire au centre de laquelle se trouvait A qui recevait 3 liens de B, C, D. L'analyse de la citation consiste à construire une nouvelle matrice qui mettrait un indice de 1 à l'association B, C, D. Même si ces acteurs n'ont jamais eu de relations directes les uns les autres, le fait de répondre en parallèle à la même question permet de les associer au sein d'un collège virtuel partageant sans doute des compétences voisines. Ce faisant, on pose une hypothèse sur la signification de la citation [Liu, 93].

Un laboratoire virtuel sera composé d'acteurs du forum qui interviennent souvent ensemble pour répondre aux questions émises. Ces témoignages sur certaines questions signifient qu'ils ont des compétences ou un axe d'intérêt sur le sujet.

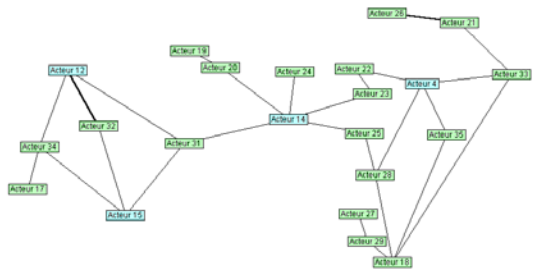

Figure 11 : réseau primal construit à partir du logiciel matrisme

Pour obtenir ce réseau, nous avons considéré tous les acteurs du forum qui étaient intervenus au moins trois fois pour répondre à une question. On a ensuite construit un réseau avec le logiciel Matrisme qui permet d'associer par un lien deux acteurs s'ils ont répondu chacun à deux questions au moins ayant le même émetteur. L'épaisseur de la relation entre deux acteurs correspond au nombre de questions en commun auxquelles les deux acteurs ont répondu.

Prenons l'exemple de l'acteur 27 et l'acteur 29 qui figurent au sud du réseau. Ces deux acteurs sont intervenus pour répondre à deux questions du même émetteur. Cela signifie qu'ils partagent probablement une zone d'intérêt et qu'ils ont peut être des compétences similaires.

Le graphe résultant permet d'observer une structure réticulaire organisée autour de 2 parties reliées par un lien entre acteur 14 et acteur 31. Ces deux acteurs correspondent, selon la théorie des graphes, à des points d'articulation : il s'agit d'intervenants sans lesquels le graphe aurait été déconnecté en plusieurs sous parties. Il s'agit sans doute d'intervenants qui ont des problématiques transversales, ce qui explique leur présence à l'interface de plusieurs sous groupes.

CONCLUSION :

A travers l'étude que nous venons de faire, les fora de discussion apparaissent comme des sources riches en données informatives qui peuvent couvrir une partie des besoins des entreprises en matière de veille marketing, de veille concurrentielle ou sociétale. Le fait d'offrir une vision d'ensemble des données d'un forum de discussion permet selon le cas :

- d'accéder de façon privilégié à l'information qui constitue le cœur de l'activité du forum
- de positionner une intervention au regard des autres interventions ce qui peut permettre d'en crédibiliser le contenu.

BIBLIOGRAPHIE

Boutin E., *Le traitement d'une information massive par l'analyse réseau : méthodes, outils et applications*, thèse de doctorat, Université d'Aix-Marseille, 1999

Donath Judith S. *Visual Who: Animating the affinities and activities of an electronic community*, ACM Multimedia 95 - Electronic Proceedings November 5-9, 1995 San Francisco, California

Egghe L., Rousseau R.: *Introduction to Informetrics, Quantitative Methods in Library, Documentation and Information Science*. Elsevier, 1990.

Garfield E.: *Citation Indexing: Its Theory and Application in Science, Technology and Humanities*. The ISI Press, 2nd ed., Philadelphia, PA, 1983.

Kohonen T., S. Kaski, K. Lagus, J. Salojärvi, J. Honkela, V. Paatero, and A. Saarela.

Self Organization of a Massive Document Collection. IEEE Transactions on Neural Networks,

Special Issue on Neural Networks for Data Mining and Knowledge Discovery, volume 11, number 3, pages 574-585. May 2000.

Liu M.: *The Complexities of Citation Practice*: a Review of Citation Studies Journal of Documentation, 49(4), 1993, 370-408

Morvan Y., Fondements d'économie industrielle, Economica, 1985

Smith, Marc. 1997. *Netscan: Measuring and Mapping the Social Structure of Usenet, Pressented at the 17th Annual International Sunbelt Social Network Conference*, Bahia Resort Hotel, Mission Bay, San Diego, California, February 13-16, 1997

Warren Sack, "*What does a very large-scale conversation look like*?" in the Electronic Arts Proceedings of ACM SIGGRAPH 2001 (Los Angeles, CA: ACM, August 2001).

Xiong Rebecca, Judith Donath ; "*PeopleGarden: creating data portraits for users*", Proceedings of the 12th annual ACM symposium on User interface software and technology, 1999 a , Asheville, North Carolina, United States

Xiong Rebecca, Judith Donath ; "*PeopleGarden: creating data portraits for users*", Proceedings of the 12th annual ACM symposium on User interface software and technology, 1999 b , Asheville, North Carolina, United States