Les réseaux latents : un outil au service de l'intelligence économique

Eric Boutin, Pei Liu, Yi Yuan

LES RESEAUX LATENTS : UN OUTIL AU SERVICE DE L'INTELLIGENCE ECONOMIQUE

Boutin Eric (*), Liu Pei (*), Yuan Yi (**)
boutin@univ-tln.fr, liu@univ-tln.fr, yuanyixz@126.com

(*)Université du Sud Toulon Var I3M IUT TC BP 132 83957 la Garde Cedex - France
(**)Université de Mine et Technologie de Chine, Xuzhou - Chine

Mots clefs :
analyse réseau, réseau latents, intelligence économique, innovation
Keywords:
network analysis, latent network, competitive intelligence, innovation
Palabras clave :
análisis de red, red latente, inteligencia competitiva, innovación

Résumé

L'absence de quelque chose est parfois aussi, voir plus signifiante, que sa présence. Le réseau latent repose sur l'exploitation des vides. L'approche par les réseaux latents s'intéresse à l'identification de traces qui n'existent pas alors qu'elles devraient exister. Cette approche permet d'identifier des non associations remarquables. Par exemple deux auteurs n'ont jamais publié ensemble alors que leur recherche est voisine. Peut être y a t il une incompatibilité entre ces personnes qui ne peuvent travailler ensemble ? Peut être la non association préfigure-t-elle une association future, une association émergente?
Dans ce travail nous allons décrire la méthode utilisée pour révéler les réseaux latents. Cette méthodologie sera ensuite déployée de manière expérimentale sur un corpus documentaire composé des publications scientifiques des membres d'un laboratoire de recherche. On fera alors ressortir l'intérêt stratégique des résultats obtenus pour la gouvernance d'un laboratoire, pour un ministère de tutelle et plus généralement l'intérêt de l'approche des réseaux latents pour l'Intelligence Economique.

Abstract

The absence of something is sometimes as, to see more meaning, as its presence. The latent network analysis is interested in vacuums. This approach makes it possible to identify non remarkable associations. For example, two authors never published together whereas their research is close. There can be an incompatibility between them? Pehaps this non association means a new collaboration in the future? In this work, we will describe the method used to reveal latent networks. This methodology will be then deployed in an experimental way on a documentary corpus composed of the scientific publications of the members of a research laboratory. One will then emphasize the strategic interest of the results obtained for the governorship of a laboratory, a ministry and more generally the interest of this approach for Competitive Intelligence.

1 Introduction

Lorsqu'on demande au praticien de l'Intelligence Economique de donner le mot clé caractérisant toute démarche d'IE, le terme réseau revient invariablement. L'analyse réseau [6] propose une méthodologie de représentation et de caractérisation des interactions entre objets étudiés : les outils de l'analyse réseau sont d'une part des graphes permettant de visualiser ces interactions et d'autre part des indicateurs permettant de caractériser la topographie du réseau et d'identifier des nœuds aux propriétés particulières. Etre capable de représenter les réseaux, c'est donc avoir un avantage décisif en IE. Pour cette raison, l'analyse réseau connaît des développements considérables dans ce domaine [1].
Nous proposons dans ce travail l'introduction du concept de « réseau latent » et de montrer le potentiel de ce concept pour la démarche d'IE.
Lorsque la personne en charge de l'IE effectue une analyse technique d'un corpus documentaire, elle cherche à faire revivre les données qu'elle observe de l'extérieur et à comprendre le processus humain qui a été à l'origine de leur création. Le processus de production de connaissances laisse des traces. Lorsque on travaille sur un corpus issu d'une base de données bibliographiques, on peut travailler sur une trace appelée co-signature. On peut inférer de l'étude de ces co-signatures des comportements de collaboration entre auteurs. On fait donc l'hypothèse que les co-signatures sont représentatives du processus coopératif réel et on reconstitue a posteriori le réseau des acteurs. On est donc dans un processus d'analyse rétrospective qui reconstitue un réseau passé pour avoir les clés de l'avenir. L'analyse réseau repose donc fondamentalement sur l'étude des traces.
L'absence de quelque chose est parfois aussi, voir plus signifiante, que sa présence. Le réseau latent repose sur l'exploitation des vides. L'approche par les réseaux latents s'intéresse à l'identification de traces qui n'existent pas alors qu'elles devraient exister. Cette approche permet d'identifier des non associations remarquables. Par exemple deux auteurs n'ont jamais publié ensemble alors que leur recherche est voisine. Peut être y a t il une incompatibilité entre ces personnes qui ne peuvent travailler ensemble ? Peut être la non association révélée par l'analyse des réseaux latents préfigure-t-elle une association future, une association émergente?
Dans ce travail nous allons décrire la méthode utilisée pour révéler les réseaux latents. Cette méthodologie sera ensuite déployée de manière expérimentale sur un corpus documentaire composé de 300 publications scientifiques des membres du laboratoire de recherche I3M des trois dernières années. Faisant partie de ce laboratoire et le connaissant de l'intérieur, nous serons à même de valider la pertinence des résultats observés. On fera alors ressortir l'intérêt stratégique des résultats obtenus pour la gouvernance d'un laboratoire ou pour un ministère de tutelle et plus généralement l'intérêt de l'approche des réseaux latents pour l'IE.

2 L'Analyse des réseaux latents : Esprit, méthodes, outils

Il y a un paradoxe à vouloir chercher l'information qui manque alors qu'on est dans un contexte de surcharge informationnelle. Pourquoi ne pas se satisfaire de l'information pléthorique que l'on a à traiter ? La réponse à cette question est à rechercher dans notre volonté d'identifier des associations qui n'existent pas encore.
Il y a un autre paradoxe à déployer des méthodologies infométriques qui balaient des corpus documentaires lourds pour identifier une information qui n'y est précisément pas. La encore la méthode que nous utilisons mobilise une approche transitive appliquée à des corpus documentaires stabilisés permettant ainsi de faire émerger des éléments potentiellement nouveaux.
Notre approche s'intéresse aux méthodes de génération de l'émergent ou de l'innovant. Ces approches ont été développées surtout dans le domaine biomédical [5], [4], [2], [7]. Dans ce travail, nous changeons de registre applicatif et nous nous intéressons non pas à l'identification d'éléments innovants mais d'associations innovantes.

Dans ce travail, nous allons privilégier une expérimentation dans un contexte scientométrique de traitement de corpus bibliographique. En effet, de part la nature structurée des données, on peut présenter ainsi une méthode de façon plus pédagogique. Toutefois la méthode s'applique à d'autres contextes.

2.1 L'esprit de la méthode

Lorsqu'on construit un réseau d'auteurs à partir d'un corpus bibliographique, on retient le champ auteur de chaque référence bibliographique et on créé des associations entre les paires d'auteurs lorsque les auteurs correspondant sont associés au sein d'une même référence. Supposons deux auteurs A et B qui ont écrit ensemble un article. Cette co-signature se traduit par la présence d'un lien entre A et B sur le futur réseau. De la même manière, lorsqu'on s'intéresse à la représentation de réseaux de concepts, le réseau est construit à partir d'une agrégation de données élémentaires correspondant chacune à la présence conjointe de deux mots clés dans une même référence. Si les mots clés 1 et 2 sont présents dans une même référence, le réseau restituera cette coprésence par un lien entre 1 et 2. Dans les deux cas, on retient au sein de chaque notice bibliographique un champ unique (auteur, mot clé, inventeur…) et c'est la co-présence de plusieurs modalités de ce champ au sein d'une même référence qui va créer l'interaction. On parlera alors de traitement intra-référence pour désigner le fait que le lien entre deux modalités du champ observé est défini par l'information contenue dans une référence bibliographique.
Dans l'approche des réseaux latents, on retient non plus un champ mais au moins deux champs par référence. Par exemple, on retient pour un article, le champ auteur et le champ mots clés descripteurs. La création du lien ne se fait plus selon une logique intra-référence mais inter-référence. Deux auteurs ne sont plus liés parce qu'il sont présents ensemble au sein d'une même référence (analyse intraréférence). **Deux auteurs sont liés car, sans avoir jamais publié ensemble, ils ont publié l'un sans l'autre un article au moins qui est décrit par au moins un mot clé identique**. On parle alors d'analyse inter-référence. Illustrons le processus de génération des réseaux latents par un exemple simple.
Supposons deux articles pour lesquels nous disposons du champ auteur et du champ mot clé. Pour simplifier les auteurs sont des lettres et les mots clés des numéros. Le tableau 1 présente ce corpus.

Tableau 1 : Corpus de départ

Article α :
Auteurs : A, B, D
Mots clé : 1,2,3
Article β :
Auteurs :D, C
Mots clés : 1,5,7

L'auteur A a publié un article décrit par le mot clé 1 (*Article □)*.
L'auteur C a publié un autre article décrit par le mot clé 1 (*Article □*.
Cette double présence du mot clé 1 dans deux références distinctes créé une interaction entre l'auteur A et l'auteur C. On a une forme de raisonnement transitif représenté par la figure 1 :

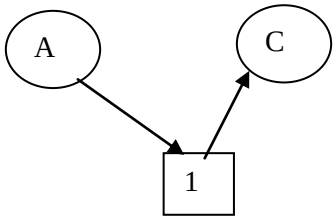

Figure 1 : la logique transitive

C'est parce que A est décrit par le mot clé 1 et que le mot clé 1 se retrouve dans une autre référence produite par B que A et B se trouvent liés.
A partir de l'exemple ci-dessus, on peut donc générer des associations entre auteurs. Il y en a 6 dans cet exemple, 6 étant le produit du nombre d'auteurs de la première référence par le nombre d'auteurs de la seconde.
Ces associations sont : AD, BD, DD, AC, BC, DC.
Parmi les 6 associations identifiés, un filtrage va maintenant permettre de retenir les associations latentes : AD n'est pas une association latente puisque A et D ont déjà produit un article en commun. Pour les mêmes raisons BD, DC ne sont pas des associations latentes. On enlève également les associations réflexives entre chaque auteur et lui-même. On obtient alors une liste d'associations appelées associations latentes. Dans notre exemple la liste des associations latentes est AC, BC. Ces associations entre auteurs ne sont pas observées actuellement et ne correspondent pas à des co-signatures. Par contre, il y a une légitimité à penser que ces auteurs pourraient être conduits à collaborer ensemble puisqu'il produisent séparément des articles décrits par au moins un mot clé commun.

2.2 Prise en compte de corpus plus réalistes :

Dans la réalité, les réseaux observés sont éloignés de cet exemple simplificateur pour des raisons liées au fait que les corpus documentaires sont plus volumineux et d'autre part que deux références peuvent partager deux ou plusieurs mots clés. La question qui se pose alors est celle de la mesure de l'intensité d'une association latente. Supposons par exemple que deux auteurs A et D aient plusieurs mots clés en commun dans plusieurs références. Comment allons nous matérialiser l'intensité de cette relation entre A et D? Nous proposons de mesurer dans un premier temps le poids d'une association latente par le nombre d'associations latentes élémentaires. Considérons l'exemple tableau 2 pour illustrer la démarche. Cet exemple reprend l'exemple précédant et rajoute simplement une référence.

Tableau 2 : enrichissement du corpus de départ

Article α :
Auteurs : A, B, D
Mots clé : 1,2,3
Article β :
Auteurs :D, C
Mots clés : 1,5,7
Article Δ :
Auteurs :D, E
Mots clés : 1,2,7

La liste des associations générées par le processus transitif est la suivante :

Entre *Article* α $\varepsilon\tau$ β : AD, BD, DD, AC, BC, CD
Entre *Article* α $\varepsilon\tau$ Δ : AD, BD, DD, AE, BE, DE (mot clé partagé 1) AD, BD, DD, AE, BE, DE (mot clé partagé 2)
Entre *Article* β $\varepsilon\tau$ Δ : DD, CD, DE, CE (mot clé partagé 1) DD, CD, DE, CE (mot clé partagé 7)
Soit au total AC, 3AD, 2AE, BC, 3BD, 2BE, 2CE, 3CD, 5DD, 4DE
Si on filtre cette liste en supprimant les associations qui existent déjà et les associations entre chaque auteur et lui-même, on obtient :
AC, 2AE, BC, 2BE, 2CE.
AE, BE et CE on un poids double que celui des autres. Ces associations latentes suggèrent des associations futures.

La prise en compte de réseaux réels nous conduit à un nouveau perfectionnement de la méthode. En effet, les mots clés n'ont pas tous le même poids pour caractériser un document. Certains mots clés sont des mots qualifiés de triviaux. Ils apparaissent dans de très nombreuses références. Le poids de l'association latente entre deux auteurs partageant un mot clé trivial sera moindre que celui d'auteurs partageant un mot clé plus rare. Nous proposons de caractériser le niveau de rareté d'un mot clé dans le corpus et d'accorder un poids à une association d'autant plus élevé que le mot clé est rare.
Ce niveau de rareté du mot clé est défini après analyse de la distribution des fréquences des mots clés observés sur le corpus. Comme souvent en sciences de l'information, la distribution observée est zipfienne ce qui autorise un partage de la liste des mots clés en trois parties à partir du calcul de l'entropie de Renyii [3]. On distinguera donc des mots clés triviaux, très fréquents auxquels on donnera un poids faible (W_1), des mots clés intermédiaires auxquels on donnera un poids moyen (W_2) et des mots clés très peu présents dans le corpus auxquels on donnera un poids importants (W_3). Les mots clés retenus dans l'analyse sont ceux qui ont une présence dans le corpus supérieure à 2 car sinon ils ne peuvent pas générer d'associations latentes entre les auteurs. Le poids d'une association entre mot clés sera obtenu par la somme des associations élémentaires, chacune étant appréciée par le poids du mot clé qui permet l'association.
Si dans l'exemple qui nous intéresse on considère que le mot clé 7 est rare (poids w_3 aux terme rares) et tous les autres moyennement rares (poids W_2 aux termes moyennement rares), alors le calcul de l'intensité de la relation se déterminera comme suit.
Entre *Article* α $\varepsilon\tau$ β : W_2AD, W_2BD, W_2DD, W_2AC, W_2BC, W_2CD
Entre *Article* α $\varepsilon\tau$ Δ : W_2AD, W_2BD, W_2DD, W_2AE, W_2BE, W_2DE (mot clé partagé 1) W_2AD, W_2BD, W_2DD, W_2AE, W_2BE, W_2DE (mot clé partagé 2)
Entre *Article* β $\varepsilon\tau$ Δ : W_2DD, W_2CD, W_2DE, W_2CE (mot clé partagé 1) W_3DD, W_3CD, W_3DE, W_3CE (mot clé partagé 7)

Si on filtre cette liste en supprimant les associations qui existent déjà et les associations entre chaque auteur et lui-même, on obtient :
W_2AC, 2 W_2AE, W_2BC, 2 W_2BE, (W_2+W_3)CE.
Si on prend par exemple W2= 2 et W3= 3, on obtient
2AC, 4AE, 2BC, 4BE, 5CE. L'association latente qui a la plus forte intensité est celle qui relie C à E.
Pourquoi C ne collabore-t-il pas avec E aujourd'hui ?

3 Validation expérimentale sur un corpus scientométrique

3.1 Les données du problème :

Nous sommes partis d'un corpus de 316 références bibliographiques correspondant aux publications scientifiques du laboratoire I3M durant les 3 dernières années. Le tableau 3 présente les 5 premières références de ce corpus :

Tableau 3 : 5 premières références du corpus

REF1
Auteur(s) : ALEXIS Henri, BATAZZI Claudine
Titre : « Une approche des TIC dans l'organisation par la notion de confiance ».
Colloque : « Pratiques et usages organisationnels des sciences et technologies de l'information et de la communication, Cersic Erellif, Rennes, 7, 8 et 9 septembre.
Lieu de publication : Actes
Date de publication : 2006

REF3
Auteur(s) : ARASZKIEWIEZ Jacques
Titre : Dispositif, intention perception
Colloque : Penser les images : intentionnalités, enjeux et médiations, Université Paris XIII
Date de publication : 2006

REF4
Auteur(s) : BATAZZI Claudine
Titre : «(Ré) introduire la notion de confiance » dans les organisations par les TIC »,
Colloque : « Culture et confiance pour une économie de la connaissance », IR2I, 29 et 30 juin 2006 Saint Denis La Plaine
Lieu de publication : Actes de colloque
Date de publication : 2006

REF5
Auteur(s) : BERTACCHINI Yann, Herbaux Philippe
Titre : L'intelligence territoriale, entre rupture et anticipation
Colloque : 15°congrés de la SFIC du 10 au 12 mai 2006 - Questionner les pratiques d'information et de communication ; Agir professionnel et agir social
Date de publication : 12 mai 2006

REF6
Auteur(s) : JOACHIM Joelle, KISTER Jackie, BERTACCHINI Yann, DOU Henri
Titre : « Intelligence économique et système d'information »
Revue : Information, Savoirs, Decisions et Mediations (ISDM)
Num. de la revue : 24
Date de publication : 1e trimestre 2006

3.2 Présentation de la méthode :

Pour chaque référence, nous avons retenu le champ auteur et le champ titre.
Ne disposant pas de champ mot clé, nous avons entrepris un retraitement du champ titre qui a consisté en trois opérations :

- Reconstitution de groupes de mots à partir du jargon du métier : *intelligence économique devient intelligence_économique*
- Lemmatisation des termes par mise au singulier et à l'infinitif
- Elimination des mots vides à forte occurrence mais sans pouvoir discriminant par rapport au domaine étudié. Cette étape consiste non seulement à supprimer les mots vides de la langue française (de, en par, et…) mais également les termes qui n'ont pas de signification particulière dans le

domaine étudié (enjeux, projets, concept…). Deux auteurs qui emploient un ou plusieurs de ces mots dans plusieurs références n'ont par de raison d'être regroupés. Cette extraction a été réalisée manuellement et permet de se constituer un dictionnaire du domaine. Le tableau 4 présente les termes les plus occurents de ce dictionnaire:

Tableau 4 : mots clés du domaine les plus occurents

terme	fréquence
communication	44
culture	18
tic	16
territoire	15
influence	14
sic	14
web	14
intelligence_territoriale	13
internet	13
organisation	13
éthique	10
information	10
distic	9

A partir de la statistique d'apparition des mots clés identifiés dans le corpus, on supprime les mots clés qui apparaissent une seule fois car ces mots clés ne pourront pas faire l'objet d'association transitive.

On génère, après cette étape de traitement trois listes :

- Une liste référence auteur qui identifie les auteurs de chaque référence. Cette liste est constituée comme illustré tableau 5.

Tableau 5 : liste référence – auteurs

numéro de référence	auteur
1	BATAZZI Claudine

1	ALEXIS Henri
	ARASZKIEWIEZ
3	Jacques
4	BATAZZI Claudine
5	Herbaux Philippe
5	BERTACCHINI Yann
6	KISTER Jackie
6	JOACHIM Joelle
6	DOU Henri
6	BERTACCHINI Yann

- Une liste référence mot clé qui fait correspondre à chaque référence les mots clés qui en caractérisent le titre. Un exemple de cette liste est fourni tableau 6

Tableau 6 : liste référence – mots clés

Numéro de référence	Mot clé du titre
1	tic
1	organisation
1	confiance
3	perception
3	dispositif
4	tic
4	organisation
4	confiance
5	rupture
5	intelligence_territoriale
5	anticipation
6	intelligence_économique

- Une liste qui fait correspondre à chaque mot clé sa valeur basée sur un critère de rareté. On définit le poids d'un mot clé comme inversement proportionnel à sa fréquence d'apparition dans le corpus. Le caractère zipfien de la distribution statistique des fréquences d'apparition des mots clés dans le corpus fournit une clé de ventilation des mots clés entre mots clés très occurents (valeur 1), mots clés moyennement occurents (valeur 2) et mot clés faiblement occurents (valeur 3). Le tableau 7 fournit un échantillon de ce tableau avec précision de la fréquence d'apparition du mot clé dans le corpus et sa valeur.

Tableau 7 : liste référence – auteurs

Mot clé descripteur du titre	fréquence d'apparition dans le corpus	valeur du mot clé
communication	44	1
culture	18	1
tic	16	1
territoire	15	1
influence	14	1
…		
éthique	10	2
information	10	2
distic	9	2
entreprise	9	2
médias	9	2
…		
serendipity	2	3
swanson	2	3
télévision	2	3
théâtre	2	3
wiki	2	3
…		

On recherche alors chaque mot clé dans chaque paire de référence et si on trouve ce mot clé on croise deux à deux les auteurs de ces deux publications en leur affectant le poids du mot clé. La liste des paires d'auteurs pour lesquels la valeur de l'association est supérieure à 0 est appelée liste d'associations potentiellement latentes. Elle est composé de 1106 associations dans le corpus (il existe un maximum de 3403 associations latentes entre les 83 auteurs des articles du corpus). Nous avons extrait dans le tableau 8 les 10 associations latentes potentielles qui ont la valeur la plus forte :

Tableau 8 : quelques associations latentes potentielles

Auteur	Auteur	Valeur de l'association
BERTACCHINI Yann	Herbaux Philippe	139

Auteur	Auteur	Valeur de l'association
BOUTIN Eric	COURBET Didier	75
BOUTIN Eric	RASSE Paul	70
BERTACCHINI Yann	RASSE Paul	70
RASSE Paul	BOUTIN Eric	70
BERTACCHINI Yann	BOUTIN Eric	65
BERTACCHINI Yann	DUMAS Philippe	64
COURBET Didier	FOURQUET-COURBET Marie-Pierre	62
BOUTIN Eric	GALLEZOT Gabriel	62
VANHUELE Marc	COURBET Didier	57
…		

Parmi ces associations, certaines ont déjà donné lieu à des articles co-signés. Pour obtenir la liste des associations latentes, il faut donc supprimer de cette liste les associations entre auteurs correspondant aux publications antérieures. Le tableau 9 correspond aux associations latentes ayant le plus forte poids.

Tableau 9 : associations latentes les plus fortes

associations latentes	poids
BOUTIN Eric COURBET Didier	75
BOUTIN Eric RASSE Paul	70
BERTACCHINI Yann RASSE Paul	70
BERTACCHINI Yann BOUTIN Eric	65
DUMAS Philippe RASSE Paul	49

Le tableau 10 fournit, à titre d'illustration les mots clés qui permettent aux deux premiers auteurs d'être reliés

Tableau 10 : mots clés partagés par deux auteurs

Mots clés à l'origine de la relation latente entre Boutin et Courbet			
mot clé	**Auteur1**	**Auteur 2**	**SommeDeVALEUR**
web	BOUTIN Eric	COURBET Didier	27
internet	BOUTIN Eric	COURBET Didier	20
communication	BOUTIN Eric	COURBET Didier	12
modèle	BOUTIN Eric	COURBET Didier	8
sic	BOUTIN Eric	COURBET Didier	3
indicateurs	BOUTIN Eric	COURBET Didier	3
image	BOUTIN Eric	COURBET Didier	2

On représente dans le réseau de la figure 2 les interactions latentes entre les 13 associations latentes qui ont le poids le plus fort et qui correspondent à elles seules à 10% du poids total des 1200 associations latentes.

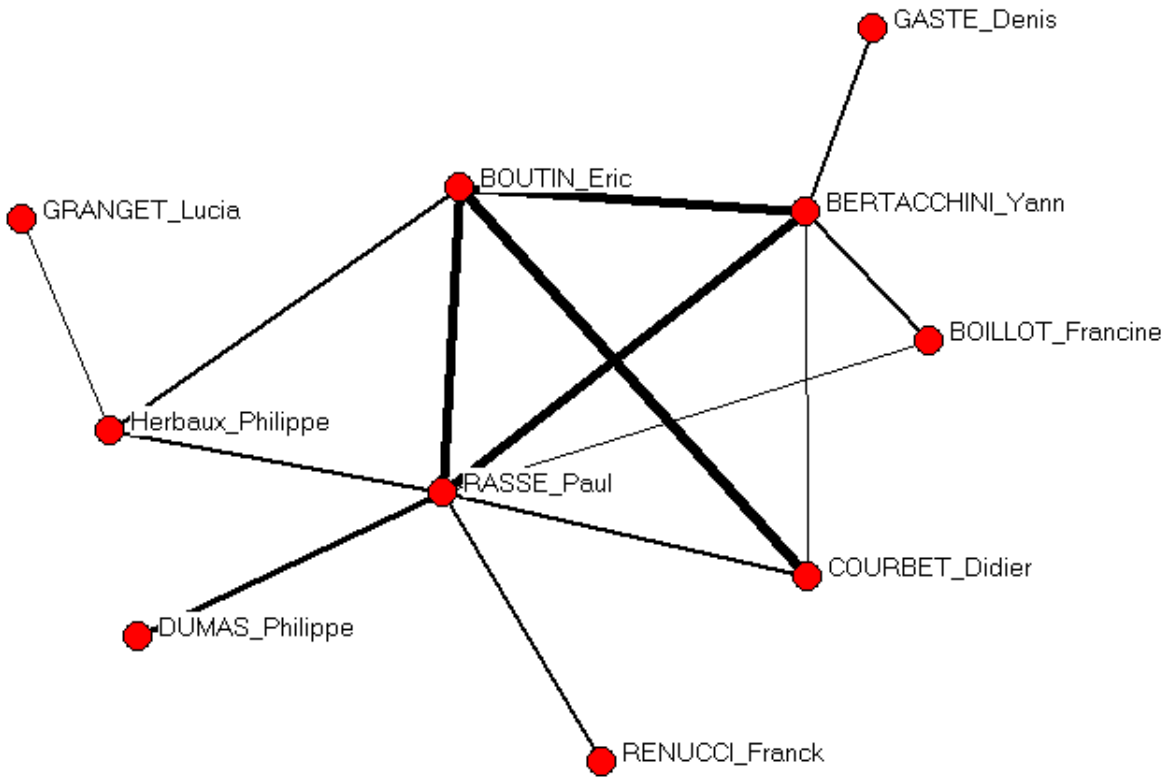

Figure 2: réseau des associations latentes les plus fortes

On observe que la méthode a tendance à survaloriser la création de relations latente entre les auteurs qui ont publié le plus. Les associations latentes découvertes correspondent à des associations qui concernent souvent de gros publieurs du laboratoire.

Il serait intéressant de disposer d'un réseau qui positionne les liens de co-publication existant entre les membres du laboratoire et les liens latents révélés par la méthode. Ce réseau est proposé figure 3. En bleu, nous avons représentés les liens latents et en rouge les liens réels. Certains liens bleus permettraient de connecter entre eux des réseaux jusqu'à présent disjoints. Au sud est du graphe, le réseau de Courbet qui était jusqu'à présent disjoint du reste du graphe se trouve désormais connecté à la grosse composante fortement connexe du fait de la relation avec Boutin.

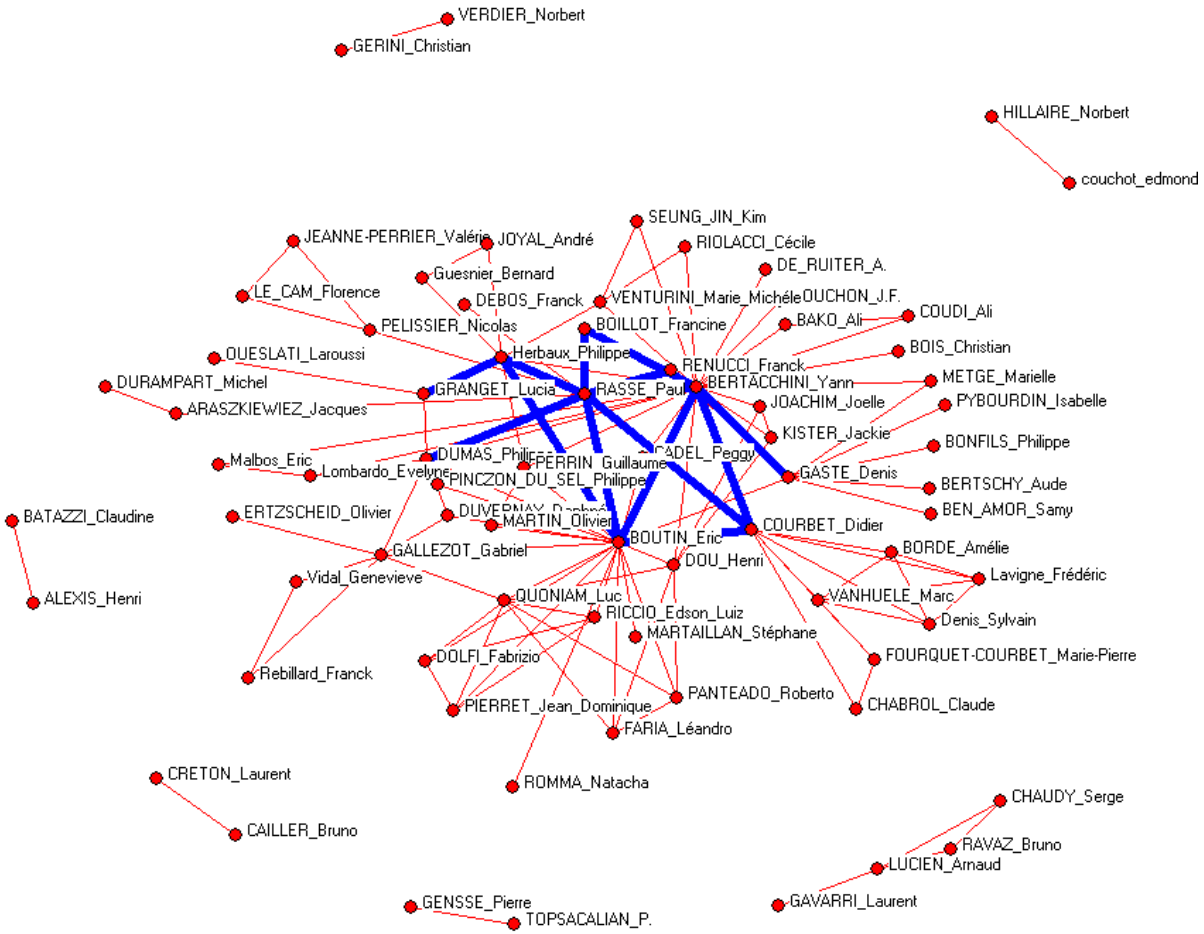

Figure 3 : réseau représentant les relations réelles (rouge) et latentes (bleu)

3.3 Interprétation des résultats :

Cette notion d'association latente peut être utilisée à un double niveau

- Pour identifier des relations potentielles entre des chercheurs qui ne travaillent pas encore ensemble. Les résultats de la méthode ont été soumis aux chercheurs du laboratoire. Certains résultats observés étaient évidents pour les personnes interrogées. Mais la méthode se révèle aussi assez originale en ce qu'elle permet de révéler des associations non triviales pour le chercheur du laboratoire I3M dont la production scientifique a été observée. Il

peut être important de regrouper des auteurs qui ont des profils semblables pour qu'ils publient ou travaillent ensemble. Cet outil est donc une méthode de génération possible de telles associations.

- D'identifier le caractère potentiellement innovant d'une collaboration scientifique. Prenons le cas de Boutin et Gasté qui sont tous deux de gros producteurs d'articles. Il existe une collaboration entre ces deux auteurs en 3 ans. Cette association n'est donc pas latente puisqu'elle s'est déjà matérialisée. Toutefois, on observe qu'entre ces deux auteurs, la valeur de l'association latente est de 10. L'étude zipfienne de la distribution statistique montre que l'association entre ces deux auteurs était improbable. Le fait que ces deux auteurs aient collaboré est intéressant car ces deux auteurs ont en réalité un profil recherche assez éloigné. Cette association a donc en germe un potentiel d'idées créatrices qui résultent de l'association originale entre des chercheurs aux problématiques différentes.

Conclusions, perspectives

Dans notre expérimentation, la méthode est utilisée à partir des champs auteur et titre. La méthode peut être utilisée à partir des champs auteur et mot clés ou auteur et références bibliographique (pour autant qu'on dispose d'information sur la citation). En effet, il y aurait matière à relier deux auteurs en fonction de la bibliographie qu'ils partagent. On pourrait même pour être plus fin considérer non plus deux mais 3 critères. On pourrait regrouper les auteurs qui ont au moins un mot clés commun et une référence commune. Cela permettrait d'être plus sélectif et de conforter la pertinence des associations latentes identifiées.

Bibliographie :

[1] Boutin, E. (1999). Le traitement d'une information massive par l'analyse réseau: méthode, outils et applications. Unpublished doctoral dissertation, Université Aix-Marseille III, Marseilles, France.

[2] Gordon, M., Lindsay, R.K, Fan, W. (2002), "Literature-based discovery on the World Wide Web", ACM Transactions on Internet Technology. Vol. 2, n°4, p. 261-275.

[3] Lhen J., Lafouge T.,Elskens Y., Quoniam L. , Dou H. (1995), « La statistique des lois de Zipf », *Actes du colloque : Les systèmes d'information élaborée*, Ile Rousse, 1995

[4] Srinivasan, P. (2004), "Text mining: generating hypotheses from MEDLINE", Journal of the American Society for Information Science. Vol. 55, n°5, p. 396-413

[5] Swanson, D.R. (1988), "Migraine and magnesium : eleven neglected connections", Perspectives in Biology and Medicine. Vol. 31, n°4, p. 526-557.

[6] Wasserman S, Faust K (1994). *Social Network Analysis : Methods and Applications*. Cambridge, England, and New York : Cambridge University Press

[7] Weeber, M.A., Klein, H., Aronson, A.R., Mork, J.G., de Jong – van den Berg, L.T.W., Vos, R. (2000), "Text-based discovery in biomedicine: the architecture of the DAD-system", Proceedings of the AMIA Symposium. p. 903-907.