Co-publications scientifiques, analyse des réseaux latents

Eric Boutin, Gabriel Gallezot, Daphné Duvernay

Co-publications scientifiques, analyse des réseaux latents

Boutin Eric
Université du Sud Toulon Var
Maître de Conférences laboratoire I3M IUT TC
BP 132 83957 la Garde Cedex FRANCE
boutin@univ-tln.fr
+33 4 94 14 23 56

Gabriel Gallezot
Université de Nice - Sophia Antipolis, Urfist PACA-C
Laboratoire I3M,
gallezot@unice.fr

Daphné Duvernay
Université du Sud Toulon Var
Maître de Conférences laboratoire I3M IUT TC
BP 132 83957 la Garde Cedex FRANCE
duvernay@univ-tln.fr
+33 4 94 14 22 30

Résumé :
L'analyse des co-publications scientifiques a fait l'objet de travaux nombreux en mobilisant l'analyse relationnelle à l'étude de corpus infométriques. Tous ces travaux valorisent l'existence d'une relation entre auteurs comme composant élémentaire du réseau d'association. Dans ce travail, notre objectif consiste à mettre en évidence l'absence remarquable de collaboration entre auteurs. L'absence de relation peut se révéler beaucoup 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 collaborations potentielles. Cette approche permet de révéler des non associations remarquables. Une association latente entre deux auteurs correspond au fait que ces deux auteurs n'ont jamais collaboré ensemble alors que sur d'autres plans, ils ont entre eux une certaine proximité. Par exemple deux auteurs n'ont jamais publié ensemble alors que leurs recherches sont voisines. 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 qui viendrait renforcer ou tel ou tel champ de recherche ?
Dans ce travail, nous décrivons la méthode infométrique utilisée pour révéler les réseaux latents avant de la déployer de manière expérimentale sur un corpus documentaire composé de près de 900 articles publiés dans l'archive ouverte d'Archivesic.

Mots clés :
Analyse relationnelle, infométrie, réseau latent, archivesic

Abstract :
Analysis of co-scientific publications has been the main subject of numerous works mobilizing relational analysis to the study of infometric corpora. All these works focus on the existence of a relationship between authors as elementary component of the network association. In this work, our aim is to highlight the remarkable lack of collaboration between authors. The lack of relationship can be more significant than his presence. The latent network is based on the use of empty. The approach by the latent networks is interested in the identification of potential collaborations. This approach helps to reveal new associations. A latent association between two authors means that these two authors have never worked together while in other areas, they have between them a proximity. For example, two writers have never published together while their research is similar. Maybe there is an incompatibility between those people who can not work together? Maybe the latent association revealed by the analysis focus a emerging association in the future.
In this work, we describe the infometric method used to reveal latent networks. We then propose an experimental work composed of nearly 900 papers published in the French open archive Archivesic.

Keywords :
Network analysis, infometric, latent network, archivesic

INTRODUCTION

Des travaux menés en SIC (Dumas et al – 2005, Gallezot et al.-2006, Loneux et al. -2005) se sont interrogés sur l'association française des « sciences de l'information ET des sciences de la communication ». Ces études reposent, dans leur démarche exploratoire, sur l'analyse de corpus d'articles scientifiques ou de thèses des membres de la communauté. Ces études tant qualitatives que quantitatives débouchent sur l'analyse rétrospective de la communauté des SIC. Dans cette communication, nous nous inscrivons dans la logique de ces travaux de caractérisation de notre discipline par une étude des réseaux latents. Nous souhaitons nous positionner non plus seulement dans une attitude d'analyse rétrospective d'une situation, comme précédemment évoqué, mais dans une démarche pro-active, démarche qui souhaite, par une étude des liens possibles entre chercheurs (réseaux latents), donner à lire les potentiels de l'unité plurielle des SIC.

Notre travail consistera tout d'abord à dresser les contours théoriques du concept de réseau latent pour introduire la méthodologie infométrique induite. Cette dernière sera ensuite déployée de manière expérimentale sur un corpus documentaire composé de près de 900 textes déposés dans Archivesic[1]. Nous évaluerons ainsi les capacités de ce concept.

1 : ETAT DE L'ART : REFERENCES THEORIQUES DE L'ETUDE

Les techniques de fouilles de données (datamining) ont pour objectif d'extraire, de corpus volumineux, des indicateurs ou des représentations synthétiques qui fournissent une ou des grilles de lectures des documents primaires constituant le corpus. Il s'agit souvent de faire revivre un domaine et de présenter les résultats d'un corpus non pas comme un ensemble de documents disjoints mais comme un corpus organisé qui donne à voir l'articulation entre les différents éléments le constituant. La fouille de données débouche alors sur des vues qui

[1] http://archivesic.ccsd.cnrs.fr/

permettent, selon le cas, de mieux comprendre un domaine, de catégoriser les éléments d'un ensemble. Pour y parvenir, la fouille de données met en œuvre différentes méthodes de type analyse de contenu ou analyse relationnelle. Ces analyses font l'objet de descriptions approfondies dans les travaux de la communauté en fouille de données, en infométrie et bibliométrie (on peut renvoyer notamment aux colloques Isko, Vsst ou au contenu de la revue *Scientometrics*).

Dans ce travail, nous proposons une méthode permettant une lecture plus prospective que rétrospective d'une situation. Il ne s'agit pas d'étudier un corpus daté pour mieux comprendre la situation actuelle mais d'identifier des vides dans une situation passée pour mieux appréhender l'évolution à venir d'un phénomène. Cette analyse s'insère dans une logique de pensée asiatique (Jullien : 2002). Le point de départ est contenu dans l'expression : « la nature a horreur du vide ». Cet aphorisme Aristotien, devenu expression populaire, décrit le fait qu'un espace vide est rapidement occupé en évoquant une certaine forme de prédisposition naturelle des actants à remplir l'espace laissé libre. Par «le vide appelle le plein », les asiatiques désignent une chose analogue. Cette expression signifie que le vide contient en lui-même un potentiel de plein : les vides d'aujourd'hui sont potentiellement les pleins de demain. Avoir conscience d'un élément « vide » à la source alors que le plein n'est pas venu le remplir, c'est disposer du pouvoir d'anticipation et d'action sur les choses en amont. Les asiatiques se plaisent à prendre la métaphore de l'eau. La petite source d'eau peut être déviée facilement. Par contre, la source transformée en rivière devient plus difficilement domptable. Il s'agit donc dans la pensée orientale d'identifier le potentiel d'une situation le plus en amont possible comme signes d'une action à venir. Cela correspond à ce qui est appelé « signal faible » dans le domaine de l'intelligence économique. En Occident, on observe une inclinaison plus franche à voir le plein plutôt que le vide : dans l'analyse des réseaux sociaux par exemple, on privilégie souvent les relations entre auteurs comme matérialisation d'une collaboration passée. On pourrait de la même façon s'intéresser à l'absence de collaboration entre certains auteurs comme porteur de germe des collaborations futures. Cette « nouvelle » orientation de l'esprit, cette inclinaison vers le « rien » qui accorde plus d'importance à l'absence de quelque chose qu'à sa présence est notamment traquée dans les études d'usage par exemple, où la non-utilisation de tel service ou tel objet est révélateur d'un manque d'ergonomie, de formation… Ainsi donc, l'absence de quelque chose est parfois jugée plus signifiante que sa présence. Le principe du réseau latent repose sur cette exploitation des vides. Il s'agit de s'intéresser à l'identification de traces qui n'existent pas et qui pourraient, devraient exister. Cette approche permet d'identifier des non associations remarquables. Par exemple deux auteurs n'ont jamais publié ensemble alors que leurs recherches sont voisines. Peut être y a t-il dissension, concurrence, incompatibilité d'humeur entre ces auteurs ? Peut être la non association révélée par l'analyse des réseaux latents préfigure-t-elle aussi une association future, une association émergente?

La démarche que nous conduisons semble paradoxale à plusieurs titres. Tout d'abord, le fait de privilégier l'étude des vides conduit à un premier paradoxe dans la mesure où nous évoluons dans un contexte de surcharge informationnelle. Conceptuellement, l'aboutissement de l'étude conduira à la production de textes supplémentaires qui viendront ainsi enfler plus encore les corpus disponibles. Toutefois, cet aspect est le propre de la science[2] et infère une compartimentation naturelle d'une discipline. Ce que nous souhaitons mettre en avant avec cette analyse du vide, ce sont bien les liens transversaux entre les compartiments produits par la segmentation, la localisation, la méconnaissance des corpus, … bref tous les aspects de la

[2] L'accumulation des connaissances sur un thème, un objet, un processus permet d'en préciser les contours, la définition, …

surcharge informationnelle. Il ne s'agit pas seulement de produire des connaissances supplémentaires, il convient de les produire en « liaison » en pensant « unité plurielle » et complétion.

Le second paradoxe se trouve peut-être dans le déploiement de méthodologies infométriques qui balaient des corpus documentaires pour identifier l'absence d'une information. La réponse à cette question réside dans le fait que la méthode que nous utilisons mobilise une approche transitive permettant ainsi de faire émerger des éléments potentiellement nouveaux. Notre approche s'intéresse aux méthodes de génération de l'émergent ou de l'innovant, essentiellement développées dans le domaine du *knowledge discovery* et beaucoup expérimentés au sein du domaine biomédical (Swanson : 1998, Weeber et al. : 2000, Gordon et al. : 2002, Srinivasan : 2004). Dans ce travail, nous changeons de registre applicatif et nous nous intéressons non pas à l'identification d'éléments innovants mais d'associations innovantes.

On pourrait caractériser notre démarche par quelques qualificatifs qui fondent sa spécificité :
- Une approche plus orientée vers l'analyse prospective que rétrospective
- Une logique plutôt influencée par la pensée orientale qu'occidentale
- L'application d'une méthode abductive issue du *knowledge discovery* et appliquée à un autre domaine scientifique.

Nous pouvons maintenant préciser la notion de réseau latent. Dans cette communication, nous nous inscrivons dans le champ de l'infométrie par une analyse des collaborations qui passe par la co-signature d'articles. L'analyse réseau est alors un outil précieux pour restituer ces interactions (Wasserman et al. : 1994). Notre travail consistera à mettre en évidence non pas l'interaction entre des auteurs qui collaborent mais l'absence remarquable de collaboration entre auteurs. Un réseau latent est composé d'associations. Ces associations latentes entre deux auteurs correspondent au fait que ces deux auteurs n'ont jamais collaboré ensemble alors que sur d'autres plans, ils ont entre eux une certaine proximité. Quelques exemples de telles proximités peuvent être donnés à titre d'illustration. Deux auteurs n'ont jamais publié ensemble alors :
- qu'ils utilisent des mots clés analogues pour décrire leurs articles
- qu'ils renvoient aux mêmes citations.
- que leurs papiers sont souvent associés (commandés ensemble, téléchargés ensemble par les usagers, cités ensemble)
- qu'ils ont publié chacun avec des auteurs communs.

Ces exemples de proximité peuvent être combinés ensemble rendant encore plus probable l'association, l'appariement entre ces chercheurs.

2 : EXPERIMENTATION :

Cette expérimentation a été réalisée à partir d'un corpus de 886 notices bibliographiques issues de documents déposés sur Archivesic de Janvier 2003 à Septembre 2007. Les notices sont structurées autour des champs : auteurs, mots clés français, mots clés anglais, titre, date de publication, date de dépôt dans Archivesic, références bibliographiques. Les mots clés français et anglais sont saisis par le déposant : le langage n'est donc pas

contrôlé[3]. On peut noter que toutes les notices bibliographiques ne sont pas renseignées pour chacun des champs.

Notre objectif est d'identifier des binômes d'auteurs qui, bien que n'ayant jamais publié ensemble, se caractérisent par une certaine forme de proximité dans leur production scientifique. Ces binômes sont appelés associations latentes. Notre analyse consiste donc à identifier ces relations potentielles avant que les binômes ne se forment.

Plusieurs éléments permettent d'identifier ces associations latentes. Deux chercheurs présents dans archivesic n'ont jamais publié ensemble mais :

- la production scientifique de ces deux chercheurs est décrite par des mots clés semblables
- ces auteurs ont toujours publié séparément mais ils ont des collaborations communes avec d'autres chercheurs
- dans le parcours des usagers, ces auteurs ont des travaux qui font l'objet de la même attention de la part des internautes.
- Les articles de ces deux auteurs renvoient aux mêmes références bibliographiques.

Pour notre première tentative de caractérisation des réseaux latents, nous avons privilégié, les deux premiers points. En effet, l'information sur les usages est actuellement mise de coté car son exploitation suppose la mise à disposition de fichiers de traces de parcours des internautes sur le site d'Archivesic. Cet aspect représente à lui seul des protocoles longs à mettre en place.

Le traitement des références bibliographiques des documents a été écarté car ce champ n'est pas ou mal renseigné pour toutes les notices. De plus ce champ, quand il existe, reprend la bibliographie de l'auteur sans avoir fait l'objet d'une homogénéisation de style dans l'ordonnancement des éléments de la référence bibliographique[4]. Ces problèmes de curation, bien connu des bibliomètriciens, pour obtenir une base de qualité, prête au traitement automatique, représentent un travail de longue haleine qui ne pouvait être entrepris pour cette première analyse. Si les résultats de cette dernière sont encourageants, les deux derniers points pourraient être intégrés aux critères de constitution des réseaux latents.

Nous allons successivement présenter la méthode d'identification des réseaux latents avant d'évaluer dans un second temps le caractère prédictif de la méthode proposée.

3 : DETERMINER LES ASSOCIATIONS LATENTES

3.1 Raisonnement sur le champ des mots clés français

Modèle de départ

La démarche obéit à la logique suivante : soit deux auteurs qui n'ont jamais publié ensemble. Ces deux auteurs ont écrit des articles qui sont décrits par des mots clés identiques. On va qualifier l'intensité de la relation entre deux auteurs par le nombre de chemins différents qu'il existe pour passer de l'auteur 1 à l'auteur 2 en passant par l'espace des mots clés. Ce nombre sera d'autant plus élevé que l'auteur sera prolifique, que ses articles seront décrits par un grand nombre de mots clés, et que son travail scientifique sera décrit par des termes voisins de l'autre auteur. De façon générale, la méthode consiste à identifier un champ intermédiaire (ici le champ mot clé) qui permet d'établir une passerelle entre deux auteurs.

Le schéma de la figure 1 décrit le processus d'identification de ces relations latentes.[5]

---

[3] A cet égard il y a une grande proximité avec le *social taging* où des mots clés sont librement attribués pour « référencer » une ressource.

[4] Date, auteur ou auteur, titre, date, …

[5] Ce travail est effectué avec le logiciel Access

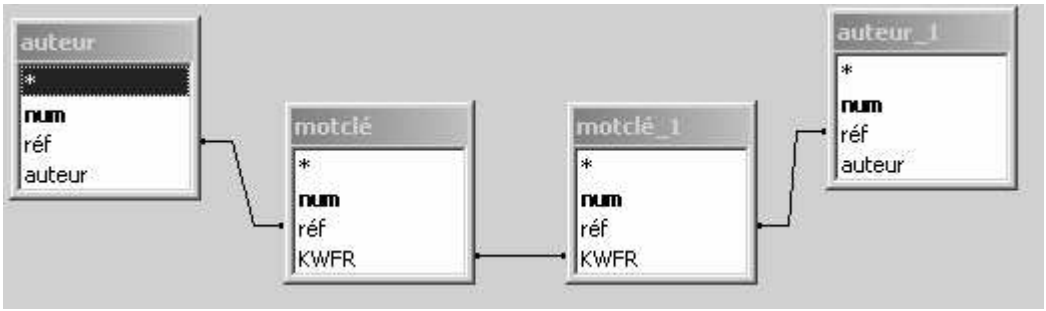

Figure 1 : principe de détermination des associations latentes

A l'issue de l'analyse, on obtient un tableau de résultats dont le tableau1 nous fournit un extrait :

Tableau 1 : identification des associations latentes

paires	intensité de latence
Poissenot, C.Fourmentraux, J.-P.	89
Fourmentraux, J.-P.Froissart, P.	6r
Fourmentraux, J.-P.Dumas,!P.	62
Fourmentraux, J.-P.Maignien, Y.	52
Fourmentraux, J.-P.Bertacchini, Y.	46
Dumas, P.Herbaux, P.	43
Froissart, P.Dumas, P.	35
Fourmentraux, J.-P.Le Roux, L.	3[3]
Fourmentraux, J.-P.Deboin, M.-C.	33
Fourmentraux, J.-P.Courbet, D.	33
BOUKACEM, C.Poyssenot, C.	32
Bertacchini, Y.Venturini, M.-M.	32
…	
Gallezot< G.Michel, C	15

- Chaque ligne de ce tableau correspond à une association latente. On explicite dans le tableau 2 la méthode de calcul permettant de trouver la valeur de latence entre les auteurs Gallezot et Michel. Les trois premières colonnes à gauche du tableau 2 décrivent l'univers de Gallezot et celles de droite celui de Michel. Les deux colonnes du milieu sont communes. Il y a 15 lignes à ce tableau donc 15 chemins permettant d'aller de Gallezot à Michel en passant par l'univers des mots clés.

Auteur départ	Num de la référence	mot clé	mot clé	Numéro de la référence	Auteur arrivée
Gallezot, G.	943	usage	usage	342	Michel, C.
Gallezot, G.	943	usage	usage	339	MICHEL, C.
Gallezot, G.	600	publication électronique	publication électronique	339	MICHEL, C.
Gallezot, G.	1416	usage	usage	339	MICHEL, C.
Gallezot, G.	1416	usage	usage	342	Michel, C.
Gallezot, G.	989	recherche d'information	recherche d'information	341	MICHEL, C.
Gallezot, G.	989	recherche d'information	recherche d'information	342	Michel, C.
Gallezot, G.	989	recherche	recherche	336	MICHEL, C.

Auteur départ	Num de la référence	mot clé	mot clé	Numéro de la référence	Auteur arrivée
		d'information	d'information		
Gallezot, G.	989	recherche d'information	recherche d'information	335	MICHEL, C.
Gallezot, G.	989	recherche d'information	recherche d'information	328	MICHEL, C.
Gallezot, G.	689	recherche d'information	recherche d'information	328	MICHEL, C.
Gallezot, G.	689	recherche d'information	recherche d'information	342	Michel, C.
Gallezot, G.	689	recherche d'information	recherche d'information	335	MICHEL, C.
Gallezot, G.	689	recherche d'information	recherche d'information	341	MICHEL, C.
Gallezot, G.	689	recherche d'information	recherche d'information	336	MICHEL, C.

Tableau 2 : mesure du niveau de latence entre Gallezot et Michel

3.2 : Affinement 1 : prise en compte de la production scientifique des chercheurs

Nous avons précisé que l'intensité de la latence dépendait du nombre d'articles de chaque auteur, du nombre de mots clés décrivant chaque article et de la proximité des mots clés d'un auteur avec ceux des autres auteurs. L'indicateur précédent survalorise donc les auteurs qui produisent beaucoup[6] d'articles décrits par beaucoup de mots clés appartenant au langage des autres auteurs. Nous souhaitons gommer cet effet. Pour cela, nous allons rapporter la latence des deux auteurs au maximum du nombre de mots clés caractérisant leurs deux productions scientifiques. Ce maximum correspond en effet au nombre de chemins maximum que l'on peut concevoir entre l'auteur A et l'auteur B. Exemple : deux auteurs ont une intensité de latence de 42. Le premier auteur a 50 mots clés caractérisant sa production scientifique et le second 100. 100 correspond au nombre de chemin maximum que l'on peut concevoir entre ces deux auteurs. Sur ces 100 chemins possibles, 42 sont observés soit 42%. Plus ce pourcentage est fort, plus on a affaire à des chercheurs qui décrivent leur production scientifique avec les mêmes mots clés. La liste fournie dans le tableau 3 privilégie les indicateurs de plus de 50%

binômes	intensité de latence	nbre mots clés auteur 1	nbre mots clés auteur 2	indicateur
Fourmentraux, J.-P.Poissenot, C.	89	118	94	0,7542373
Vanhuele, M.Intartaglia, J.	4	7	6	0,5714286
Lavigne, F.Intartaglia, J.	4	7	6	0,5714286
Herbaux, P.Goria, S.	22	40	33	0,55
Herbaux, P.Geffroy, P.	22	40	33	0,55
froissart, P.Fourmentraux, J.-P.	62	85	118	0,5254237
Dumas, P.Fourmentraux, J.-P.	62	119	118	0,5210084

Tableau3 : affinement de l'indicateur en considérant le nombre de mots clés de chaque auteur

6 Et dans notre expérimentation déposent beaucoup sur Archivesic

3.3 Affinement 2 : prise en compte du pouvoir discriminant du mot clé

Tous les mots clés n'ont pas le même poids. Certains mots clés ont un pouvoir discriminant faible (internet, SIC…) alors que d'autres ont un pouvoir discriminant plus fort (mots clés singuliers). On pourrait considérer que le pouvoir discriminant d'un mot clé est inversement proportionnel à sa fréquence d'apparition dans le corpus. Nous avons choisi une vision dichotomique plus simple qui consiste à retenir certains mots clés et à en exclure d'autres. Le tableau 4 fournit les mots clés que nous avons éliminés avec pour chacun leur fréquence d'apparition dans le corpus. Il s'agit souvent de termes triviaux.

Mots clés supprimés	citation		Mots clés supprimés	citation
internet	68		Europe	5
communication	35		(cift 04)	5
information	28		université	5
document	28		tei	4
tic	27		sciences de l'information et de la communication	4
ntic	14		sic	4
web	14		cnrs département stic	4
france	11		mathématiques	4
la rochelle	10		rtp doc cnrs rapport as	4
sciences de l'information	10		science	4
formation	9		monde arabe	4
local	9		rtp 33 document et contenu : création	4
journée atala	9		support	3
publics	9		science de l'information	3
intranet	8		roumanie	3
développement	8		semaine du document numérique (sdn 2004) (cift 04)	3
image	8		semaine du document numérique (sdn 2004) conférence	3
publication	7		objet	3
création	7		méthode	3
cdi	7		méditerranée	3
enquête	6		ordre	3
statistique	6		caractérisation	2
théorie	5		technologies de l'information et de la communication	2
éducation	5		extraction d'	2
environnement	5			

Tableau 4 : mots clés supprimés

Ces mots clés ne sont donc plus pris en considération pour juger de la proximité des travaux scientifiques de deux auteurs. En mettant en œuvre la même démarche d'association latente que celle décrite précédemment, on aboutit au résultat présenté tableau 5.

paires	intensité de latence
Geffroy, P.Bertacchini, Y.	10
Michel, C.Geffroy, P.	6
Lafouge, T.Geffroy, P.	6

paires	intensité de latence
Gauthier, M.Poissenot, C.	6
chartron, G.Schöpfel, J.	6
Bertrand-Gastaldy, S.Poissenot, C.	6
Bergeron, P.Poissenot, C.	6

Tableau 5 : Associations latentes en supprimant certains mots courants

3.4 : Les amis de mes amis sont mes amis :

On a choisi d'illustrer la démarche de recherche des associations latentes en procédant à une analyse transitive par le biais des collaborations indirectes entre auteurs. A collabore avec B qui collabore avec C mais A n'a jamais publié avec C → donc on créé un lien potentiel entre A et C. Par transitivité on peut générer des associations qui pour certaines sont nouvelles. Chaque association est caractérisée par le nombre de fois où elle est obtenue.
Le résultat est présenté dans le tableau 6 ci-dessous :

auteur1	auteur2	transitivité
Lebreton, M.	Herbaux, P.	16
Lebreton, M.	Dumas, P.	16
Herbaux, P.	Dumas, P.	16
Fourquet-Courbet, M.-P.	Denis, S.	16
Fourquet-Courbet, M.-P.	Borde, A.	16
Ertzscheid, O.	Dumas, P.	16
Riqueau, C.	Bertacchini, Y.	14
Quoniam, L.	Dumas, P.	14
Labiche, J.	Héroux, P.	14
MICHEL, C.	Bador, P.	13
Gallezot, G.	Bertacchini, Y.	13
Ertzscheid, O.	Chartron, G.	13
Prime-Claverie, C.	Michel, C.	12
Noyer, J.-M.	Ertzscheid, O.	12
link-pezet, J.	Gallezot, G.	12
lacombe, E.	Gallezot, G.	12
Gallezot, G.	Riqueau, C.	10

Tableau 6 : associations latentes obtenues par transitivité

Certains de ces auteurs appartiennent à des réseaux:

- Ertzscheid, O, Gallezot, G, Noyer, J.-M., link-pezet, J. lacombe, E. , Chartron, G. >> Urfist
- Riqueau, C, Dumas, P, Bertacchini, Y, Quoniam, L >> Univ Toulon
- Prime-Claverie, Bador >> Univ Lyon
- Michel : Bordeau
- Labiche , Héroux: Rouen

Les valeurs sont très différentes de précédemment. La ligne « Gallezot Bertacchini 13 » se reurouve0schématisée par la figure 2.

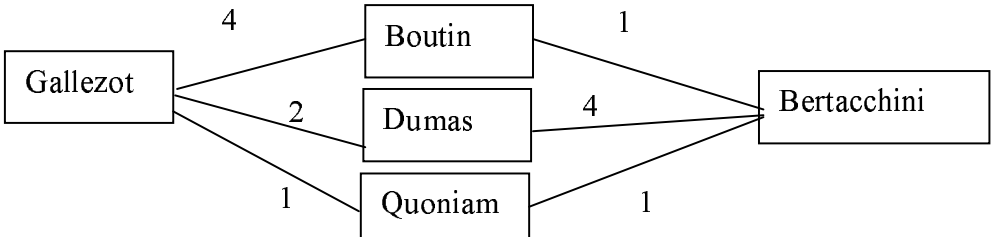

Figure 2 : fonctionnement transitif avec le champ auteur

Il y a 4 chemyns pour aller de Gallezot à Bertacchini en passant par Boutin, 8 en passant par0Dumas et 1 en passant par Quoniam soit 13 en tout.

4 : MESURES D'EFFICACITE DE LA METHODE

Dans le paragraphe prékédent, nous avons décrit deux méthodes permettant d'identifier des associations latentes$entre deux auteurs. Notre objectif est maintenant de juger de la pertinence des associations latentes révélées. Nous nous sommes intéressés à l'évaluation du caractère prédictif de la méthode présentée. Pour cela, nous avons procédé à un découpage temporel de notre corpus en deux périodes : articles déposés en 2003-2004 et articles déposés après 2004. La question est alors de savoir si les associations latentes identifiées à partir d'un corpus 2003-2004 se matérialisent par des associations réelles sur le corpus suivant. Si oui, alors notre méthode a un certain pouvoir prédictif. Nous avons donc confronté nos deux mesures au jugement de l'histoire.

4.1 : Mesure 2 : les amis de mes amis sont mes amis

On décompose la période d'étude en deux : documents déposés dans Archivesic avant 2005 et après 2005. On observe que la date de dépôt est renseignée pour 578 des 886 références[7].
La méthode consiste à étudier les associations latentes sur 2003-2004 et de voir quel est le % d'entre elles qui se traduisent à partir de 2005 par des collaborations scientifiques réelles. Les résultats ne sont pas très concluants. On avait identifié 886 associations latentes sur la base du corpus documentaires avant 2005. 5 seulement se sont concrétisées par des relations d'associations d'auteurs à partir de 2005. Elles figurent dans le tableau 7ci-après :

Association latente	force
Holzem, M.Trupin, E.	6
Dionisi, D.Holzem, M.	2
Knauf, A.Geffroy, P.	1
Knauf, A.Goria, S.	1
Quoniam, L.Boutin, E.	1

Tableau7 : associations latentes avant 2005 concrétisées à partir de 2005

Cette absence de pouvoir prédictif peut signifier que les réseaux d'auteurs ne sont pas dans la pratique les moyens par lesquels s'effectue la dynamique de la collaboration scientifique.
Cela peut vouloir dire également qu'il y a des décalages entre la perception d'une association latente et la production collaborative des deux auteurs en question.

---

7 Nous sommes tributaires de la qualité de la base.

4.2 : Mesure 2 avec mots clés communs

La méthode va consister à identifier les associations latentes sur le corpus 2003-2004. Il s'agira alors d'étudier si ces associations latentes se retrouvent dans les collaborations d'auteurs après 2004. Les résultats sont très satisfaisants. Considérons les 16 premières relations latentes identifiées sur la base des mots clés communs avant 2005. Ces résultats sont présentés tableau 8. 14 des 16 associations latentes se concrétisent en associations réelles lors de la période suivante. Cela représente un taux de transformation de 87.5% sur les 16 associations latentes les plus fortes.

Paire d'auteurs	Réalisation de latence ou non	Intensité de latence
Payette, L.Maurel, D.	Latence réalisée	16
Payette, L.Mas, S.	Latence réalisée	16
Payette, L.Da Sylva, L.	Latence réalisée	16
Froissart, P.Farchy, J.	Latence réalisée	14
Courbet, D.Lavigne, F.	Latence réalisée	13
Froissart, P.Boutin, E	Latence non réalisée	11
Geffroy, P.Bertacchini, Y	Latence non réalisée	10
Denis, S.Lavigne, F.	Latence réalisée	10
Krafft, D. B.Payette, S.	Latence réalisée	9
Maurel, D.Mas, S.	Latence réalisée	8
Maurel, D.Da Sylva, L.	Latence réalisée	8
Mas, S.Da Sylva, L.	Latence réalisée	8
Pinczon Du Sel, P.Boutin, E.	Latence réalisée	7
L'Hostis, D.Aventurier, P.	Latence réalisée	7
Fily, M.-F.Deboin, M.-C.	Latence réalisée	7
Ertzscheid, O.Boutin, E.	Latence réalisée	7

Tableau 8 : mesure du caractère prédictif de la méthode

CONCLUSION ET PERSPECTIVES

Nous avons, dans ce travail, proposé une méthode de calcul d'un indicateur de relation latente. Deux méthodes ont été proposées. La méthode qui construit les relations latentes d'après l'univers des mots clés des auteurs a un fort pouvoir prédictif. On peut toutefois relever certains biais de cette méthode. Les réseaux trouvés comme latents par l'analyse, ne peuvent être en fait que la résultante d'autres facteurs de proximité non évalués par l'analyse :

- Appartenance à un même labo, une même unité, une même association, une même université, … une même institution indiquant par là un biais lié aux réseaux relationnels que chaque acteur peut tisser au sein de toute organisation.
- Les déposants d'Archivesic peuvent déjà être considérés comme une communauté… évoquant ainsi un biais lié à l'épistémologie : la valeur accordée au partage des connaissances, les objets ou processus étudiés, …
- Une autre incidence liée au corpus choisi : rien n'indique que les binômes révélés ne préexistaient pas, notamment dans d'autres publications non déposées dans Archivesic.

Malgré ces réserves observées et liées à l'expérimentation, l'approche des réseaux latents peut servir plusieurs objectifs ÷

- Elle peut être utilisée dans une démarche de recherche d'informations. L'objectif est alors de présenter à l'usager d'un tel système, des articles voisins de l'article qu'il a trouvé ou des auteurs qui travaillent sur les mêmes domaines. En terme d'interface, la fonction pourrait alors s'apparenter à la fonction « pages similaires » de Google ou encore à une extension des fonctionnalités de Google scholar (classement des auteurs par collaboration, thèmes de cet auteur aussi travaillés par ...,)
- Elle peut être utilisée comme stimuli des échanges scientifiques. Un chercheur rentre alors dans une interface les mots clés de sa recherche, les auteurs fondamentaux qu'il cite dans ses travaux. Il obtient en retour les auteurs desquels il est le plus proche et avec lesquels il ne collabore pas encore.
- Elle peut rendre compte de thèmes inexplorés. En effet, la proximité latente de chercheurs peut indiquer le « thème maillon » manquant qui permettrait leur rapprochement.

A l'issue de ce premier bilan sur le concept de réseau latent, il conviendra d'étendre l'expérimentation à d'autres critères de proximité (analyse des parcours, des références bibliographiques, ...) et d'autre corpus documentaires (les revues SIC sur Francis, la totalité de HAL, ...), Un travail original pourrait aussi se préoccuper des différences de réseau latent issus d'un vocabulaire contrôlé ou librement choisi par l'auteur.

BIBLIOGRAPHIE

Dumas P., Boutin E., Duvernay D., Gallezot G., (2005) is communication separable from information, 2005, First european communication conference, Amsterdam, 2005

Gallezot G., Boutin E., Dumas P., (2006) "Les Sciences de l'Information ET de la Communication : une problématique du « et »", XVe Congrés SFSIC, Bordeaux, Mai 2006, Bordeaux : (2006)

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.

Jullien F (2002), traité de l'efficacité, Ed Livre de Poche

Loneux C., Bourdin S., Bouillon JL, (2005) Building the field of organisational communication in France : concepts, methods, institutions, First european communication conference, Amsterdam, 2005

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

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

Wasserman S., Faust K. (1994). *Social Network Analysis: Methods and Applications* : Cambridge University Press

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.