HAL open science

Caractérisation du rôle de vecteur de transfert de connaissances d'un périodique fondée sur l'analyse des citations

Nathalie Vedovotto, Ivana Roche, Dominique Besagni, Claire François, Marianne Hörlesberger, Edgar L. Schiebel

To cite this version:

Nathalie Vedovotto, Ivana Roche, Dominique Besagni, Claire François, Marianne Hörlesberger, et al.. Caractérisation du rôle de vecteur de transfert de connaissances d'un périodique fondée sur l'analyse des citations. 4ème Séminaire de Veille Stratégique, Scientifique et Technologique, VSST 2013, Oct 2013, Nancy, France. hal-00962362

HAL Id: hal-00962362

https://hal.science/hal-00962362

Submitted on 21 Mar 2014

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

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

Distributed under a Creative Commons Attribution 4.0 International License

CARACTERISATION DU ROLE DE VECTEUR DE TRANSFERT DE CONNAISSANCES D'UN PERIODIQUE FONDEE SUR L'ANALYSE DES CITATIONS

Nathalie VEDOVOTTO(*), Ivana ROCHE(*), Dominique BESAGNI(*), Claire FRANÇOIS(*), Marianne HÖRLESBERGER(**), Edgar SCHIEBEL(**)

nathalie.vedovotto@inist.fr; ivana.roche@inist.fr; dominique.besagni@inist.fr; claire.francois@inist.fr; marianne.hoerlesberger@ait.ac.at; edgar.schiebel@ait.ac.at

(*)CNRS, INIST, UPS76, 2 allée du Parc de Brabois, 54519 Vandœuvre-lès-Nancy Cedex, France

(**)AIT, Austrian Institute of Technology GmbH, Donau-City-Strasse 1, 1220 Vienna, Austria

Mots clefs :

revue scientifique, analyse des citations, catégorisation des mots-clés, expertise assistée par ordinateur, analyse diachronique, transfert de connaissances

Keywords:

ST journal, citation analysis, keywords categorization, computer-aided expertise, diachronic analysis, knowledge transfer

Palabras clave :

revista científica, análisis de citas, categorización de palabras clave, pericia asistida por ordenador, análisis diacronico, transferencia de conocimientos

Résumé

Dans cette étude, nous cherchons à caractériser le rôle d'une revue scientifique en tant que vecteur de transfert de connaissances entre différentes communautés de recherche. Dans le but de détecter l'étendue de l'audience d'une revue, nous nous appuyons sur l'analyse des citations, en considérant deux corpus scientifiques construits à partir de la production d'un même périodique : le premier est constitué des références *ex ante*, à savoir les références citées dans la bibliographie de tous les articles publiés dans la revue pour une année de publication donnée, le second est formé par ses références *ex post*, à savoir les travaux citant ces mêmes articles. Chacun de ces corpus fait l'objet d'une étape de classification qui produit une représentation thématique de leur contenu, rendue possible par une indexation automatique. Une catégorisation thématique assistée, validée par un expert, fournit une représentation de chacun des corpus par un vecteur de catégories, dont chaque élément indique la mesure de la représentativité de la catégorie dans le corpus. Un indicateur de rôle est alors défini, fondé sur un calcul de la similarité cosinus entre les deux vecteurs de catégories associés à la revue. Cet indicateur qualifie le comportement de la revue en termes de vecteur de transfert de connaissances. L'étude de cas a porté sur deux revues du domaine des STIC (Sciences et Technologies de l'Information et de la Communication).

1 Introduction

La littérature scientifique et technique est un support habituel de diffusion des résultats scientifiques, permettant aux chercheurs de soumettre leurs travaux à l'évaluation de leurs pairs et de les mettre à disposition de la communauté scientifique pour une éventuelle réutilisation dont les citations sont la trace. Dans ce processus de diffusion, chaque périodique joue un rôle de vecteur de transfert de connaissances. Les travaux consacrés à l'étude de ce processus de diffusion se fondent pour la plupart sur les informations fournies par les grands producteurs de bases de données bibliographiques, qui classent chaque périodique dans au moins une catégorie thématique choisie dans une liste connue. Cette catégorie est par extension associée à chacun des articles publiés dans le périodique. En outre, ces travaux développent une méthodologie fondée sur l'analyse des réseaux de citations ou de co-citations [1-4].

Dans cette étude, nous avons appliqué des techniques d'analyse de contenu aux deux corpus correspondant aux réseaux de citations situés en amont et en aval des travaux publiés dans chaque périodique. En pratique, notre analyse porte sur la comparaison de deux paysages scientifiques construits à partir de la production d'un même périodique : le premier est constitué des références *ex ante*, à savoir les références citées dans la bibliographie de tous les articles publiés dans la revue pour une année de publication donnée, le second est formé par ses références *ex post*, à savoir les travaux citant ces mêmes articles. Chacun des deux corpus, après une étape de classification automatique, est représenté par une carte thématique illustrant les domaines scientifiques présents qui, selon une métaphore géographique, constitue son paysage scientifique. Ces domaines remontent directement du contenu des articles de chaque corpus et non pas de la classification thématique des périodiques proposée par les grands producteurs de bases de données.

Nous faisons l'hypothèse qu'un périodique possède un lectorat limité à son périmètre éditorial ou qu'il dépasse ces frontières en intéressant d'autres communautés scientifiques. Nous cherchons à savoir s'il est possible de formaliser une typologie caractérisant ce rôle de vecteur de transfert de connaissances. Plus précisément, nous essayons de répondre à la question suivante : l'analyse des citations peut-elle nous aider à détecter l'étendue de l'audience d'une revue ?

Comme dans un « jeux des 7 erreurs », nous cherchons alors à savoir s'il existe une différence entre les paysages *ex ante* et *ex post* d'un périodique en termes de représentation des domaines scientifiques. Plus précisément : le second paysage inclut-il uniquement des thématiques déjà présentes dans le premier ou comporte-t-il de nouveaux domaines ? Cette analyse nous aidera à savoir si la revue joue son rôle de vecteur de transfert de connaissances exclusivement à l'intérieur de son périmètre éditorial, comportement que nous qualifions d'autarcique, ou si elle en dépasse les limites et constitue une passerelle vers des domaines absents de son périmètre éditorial, dans une approche qualifiée d'expansionniste.

Habituellement cette analyse comparative est réalisée par un expert scientifique. La méthodologie proposée vise, entre autres, à alléger sensiblement la tâche d'expertise grâce à l'introduction d'une étape de catégorisation assistée des mots-clés indexant les articles.

2 Méthode

2.1 Extractions des données

Les données ont été extraites de la base Scopus d'Elsevier, dont les caractéristiques en termes d'articles cités et citants correspondent à nos besoins. Pour un journal J donné, nous avons choisi une année de publication charnière Y suffisamment ancienne pour que les articles correspondants aient pu être cités dans des publications plus récentes. Après avoir cherché dans Scopus les articles publiés dans J pour l'année Y, nous avons extrait de cette base deux corpus de données.
Le premier, appelé BIB et représenté par une ellipse verte dans la figure 1, contient le signalement de la bibliographie de ces articles, c'est-à-dire les références des articles cités par ceux qui ont été publiés dans J durant l'année Y. Le second corpus, appelé CIT et représenté par une ellipse orange dans la figure 1, comprend le signalement des articles ayant cité les travaux publiés dans J durant l'année Y.
Notre méthodologie est fondée sur les hypothèses suivantes :

- le corpus BIB est l'image du paysage scientifique sur lequel s'appuient les travaux publiés dans le journal J durant l'année Y
- le corpus CIT est l'image du paysage scientifique qui s'est inspiré des travaux publiés dans le journal J durant l'année Y

2.2 Traitement des données

Une analyse de contenu est réalisée sur chacun des corpus au moyen de l'outil Stanalyst [5], station d'analyse de l'information scientifique et technique développée à l'INIST.

Cette analyse est précédée d'une phase de reformatage et d'indexation automatique des notices, fondée sur des techniques de traitement automatique du langage. Elle est opérée au moyen de l'outil IRC3 [6, 7] et de ressources terminologiques propres à l'INIST et spécifiques aux disciplines scientifiques identifiées par un expert comme représentatives des thématiques couvertes par le journal J.

Les corpus indexés font ensuite l'objet d'une classification automatique, représentée sous forme d'une carte faisant apparaître des réseaux thématiques de classes, chaque classe regroupant des notices similaires au regard de leur indexation. Les mots-clés liés à une classe renseignent de ce fait sur son contenu, en termes de thématique scientifique. Selon une métaphore géographique, la carte ainsi obtenue pour un corpus représente son paysage scientifique.

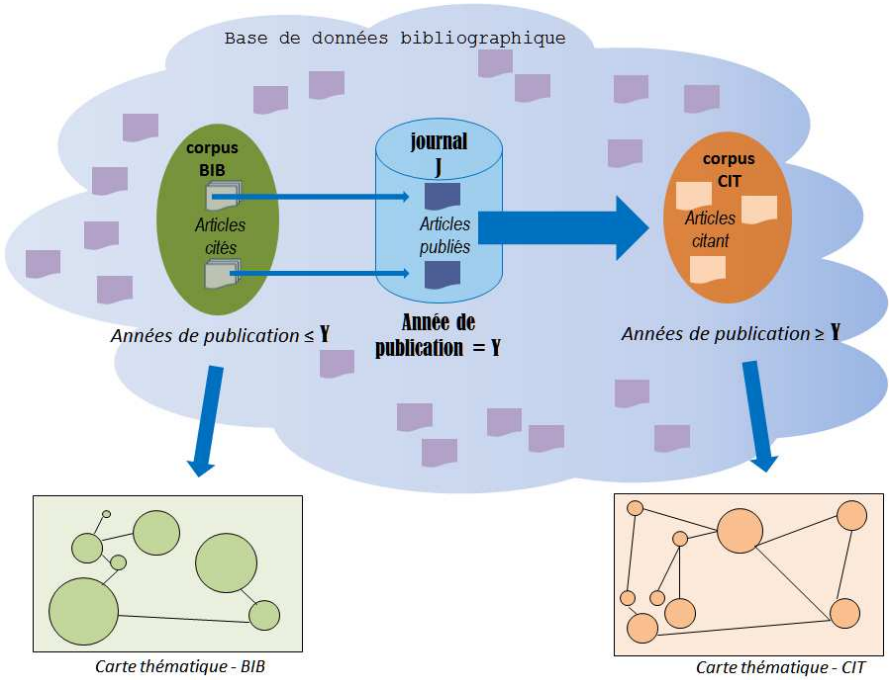

Figure 1 : Illustration de la méthode employée pour l'extraction et le traitement des données

2.3 Etape d'expertise

A ce stade intervient un expert scientifique, chargé d'analyser les cartes thématiques et d'en effectuer une analyse comparative pour en déterminer les similitudes et les différences. L'expert compare les cartes BIB et CIT pour identifier l'éventuelle apparition dans CIT de domaines scientifiques qui seraient absents de BIB, dans le but de qualifier et quantifier le rôle joué par le journal J en tant que vecteur de transfert de connaissances.
Cette tâche d'expertise est toutefois assez lourde à gérer, puisque l'expert doit examiner une à une chacune des classes de chaque carte, dans le but d'identifier son périmètre scientifique par analyse des titre, résumé et mots-clés associés aux notices qui y sont regroupées. Pour alléger ce travail, nous avons donc mis au point un outil d'expertise assistée, en utilisant le potentiel des ressources terminologiques internes à l'INIST.
Cet outil, fondé sur la catégorisation thématique des mots-clés indexant une classe, fournit pour chacune d'entre elles une image en termes de catégories, associées à des disciplines scientifiques. Il est alors uniquement demandé à l'expert de valider la liste des catégories, en opérant si besoin des regroupements ou des éclatements, et d'ajouter une catégorie aux mots-clés qui n'auraient pas été reconnus par l'outil.

2.4 Calcul de la similarité entre deux corpus

Par l'étape de catégorisation, chacun des mots-clés de BIB et CIT s'est vu attribuer une catégorie, validée par l'expert selon la procédure décrite en 2.3. Le résultat de cette étape détermine la liste des *n* catégories que nous considérons comme représentatives du périmètre scientifique du journal J, tel qu'il est défini par ses corpus BIB et CIT.
D'autre part, pour chacun des corpus BIB et CIT associés au journal J, la classification produit une carte thématique formée de *m* classes. Celles-ci sont constituées de notices bibliographiques indexées par des mots-clés. Chaque mot-clé est caractérisé par un poids calculé dans l'étape de classification.

Dans l'étape suivante, les cartes thématiques obtenues à partir de ces corpus sont analysées séparément, en identifiant l'ensemble des mots-clés de chaque catégorie présents dans chaque classe de chacune des deux cartes. Si chaque mot-clé peut être présent dans plusieurs classes de la même carte il n'est lié qu'à une seule catégorie, qui correspond à sa discipline scientifique d'appartenance, déterminée lors de la catégorisation.

L'étape suivante consiste à calculer, pour chaque classe, la représentativité de chacune des catégories, en utilisant le poids des mots-clés constituant la classe. Cette valeur, notée p_{ij}, est définie comme la somme des poids des mots-clés de la catégorie, divisée par le nombre total des mots-clés qui forment la classe :

$$p_{ij} = \frac{D_{ij}}{D_j} \times \frac{\sum_{\substack{k=1,D_{ij} \\ i=1,n \\ j=1,m}} w_{kj}}{D_{ij}} = \frac{\sum_{\substack{k=1,D_{ij} \\ i=1,n \\ j=1,m}} w_{kj}}{D_j}$$

avec :
i = nombre de catégories
j = nombre de classes
D_{ij} = nombre de mots-clés de la classe *j* liés à la catégorie *i*
D_j = nombre total de mots-clés dans la classe *j*

$\sum_{\substack{k=1,D_{ij} \\ i=1,n \\ j=1,m}} w_{kj}$ = somme des poids de tous les mots-clés de la classe *j* liés à la catégorie *i*

La représentativité de chaque catégorie dans la classe *j* est ensuite pondérée par la proportion des notices spécifiques à *j*, à savoir ne contribuant qu'à la formation de cette classe. Cette pondération vise à accentuer la contribution de cette classe à la catégorie, par rapport à d'autres classes dont le contenu est moins spécifique. Cette représentativité pondérée est notée v_{ij} :

$$v_{ij} = p_{ij} \times \frac{q_j}{Q}, \text{ avec :}$$

q_j = nombre de notices spécifiques à la classe *j*

Q = nombre total de notices du corpus

Chacune des *m* classes présentes dans une carte est donc représentée par un vecteur $\overrightarrow{v_j}$ à *n* composants :

$$\overrightarrow{v_j} = \begin{bmatrix} v_{1j} & v_{2j} & \dots & v_{nj} \end{bmatrix} \text{ pour } j = 1, m$$

La valeur de chacun de ces composants représente l'importance dans la classe de la représentativité des mots-clés liés à la catégorie correspondante, une valeur nulle indiquant par exemple que la classe n'est indexée par aucun des mots-clés relevant de cette catégorie. Après calcul pour toutes les classes, on obtient une matrice de dimension *n* x *m* :

$$\begin{bmatrix} v_{11} & v_{12} & \cdots & v_{1m} \\ \vdots & \vdots & \ddots & \vdots \\ v_{n1} & v_{n2} & \cdots & v_{nm} \end{bmatrix} = \begin{bmatrix} \overrightarrow{v_1}^T & \cdots & \overrightarrow{v_m}^T \end{bmatrix}$$

La représentativité de la catégorie *i* dans une carte, notée r_i, est alors calculée en sommant les valeurs obtenues par cette catégorie dans chacune des *m* classes de la carte :

$$r_i = \sum_{j=1}^{m} v_{ij}$$

Après calcul pour chaque catégorie, cet ensemble de valeurs constitue un vecteur de catégories de dimension *n*, noté $\vec{r}$, donnant la représentativité des *n* catégories pour l'ensemble de la carte thématique, donc pour le corpus.

$$\vec{r} = \begin{bmatrix} r_1 & r_2 & \dots & r_n \end{bmatrix}$$

Ce calcul est réalisé pour les deux corpus étudiés :

- BIB : références des articles cités par ceux qui ont été publiés dans le journal J durant l'année Y
- CIT : références des articles ayant cité les travaux publiés dans le journal J durant l'année Y

Pour le journal J, il est alors possible de déterminer un indicateur, noté S^J, fondé sur le calcul de la similarité cosinus entre les vecteurs de catégories obtenus pour chacun de ces deux corpus et qui qualifie le rôle de J en tant que vecteur de transfert de connaissances :

$$S^J = \frac{\vec{a}^J \cdot \vec{b}^J}{\left\|\vec{a}^J\right\| \cdot \left\|\vec{b}^J\right\|} = \frac{\sum_{i=1}^{n} a_i^J \times b_i^J}{\sqrt{\sum_{i=1}^{n} (a_i^J)^2} \times \sqrt{\sum_{i=1}^{n} (b_i^J)^2}}, \text{ où :}$$

J est le journal considéré
n est le nombre de catégories déterminées pour J
$\vec{a}^J$ est le vecteur de catégories obtenu à partir du corpus BIB de J
$\vec{b}^J$ est le vecteur de catégories obtenu à partir du corpus CIT de J

Selon ce mode de calcul, l'indicateur de rôle S^J peut être compris entre 0 et 1 (cf. figure 2) :

- s'il est égal à 0 la similarité thématique entre BIB et CIT est nulle, le journal est qualifié d'expansionniste : l'univers dans lequel il joue son rôle de transfert de connaissances est orthogonal à son périmètre éditorial
- s'il est égal à 1 la similarité thématique entre BIB et CIT est totale, le journal est qualifié d'autarcique : l'univers dans lequel il joue son rôle de transfert de connaissances reste limité à son périmètre éditorial
- toutes les valeurs comprises entre 0 et 1 qualifient des comportements intermédiaires, où le journal dépasse plus ou moins les frontières de son strict périmètre éditorial

Figure 2 : Echelle de comportement d'une revue, selon la similarité entre BIB et CIT

Pour affiner l'interprétation des résultats, nous calculons également pour chaque périodique J :

- la variation de la représentativité de chaque catégorie entre les deux corpus BIB et CIT : $\delta_i^J = b_i^J - a_i^J ; \forall i = 1, n$
- la moyenne et l'écart-type des δ_i^J, notés $\bar{\delta}^J$ et σ^J.

3 Etude de cas

3.1 Matériel d'étude

L'univers choisi pour réaliser cette étude est celui des STIC (Sciences et Technologies de l'Information et de la Communication), pour lequel nous disposions de ressources terminologiques qualifiées, en termes de disciplines scientifiques.
Cette étude a plus particulièrement porté sur deux revues : *Electronics Letters* (ISSN 0013-5194) et *Neural Networks* (ISSN 0893-6080). L'année de publication charnière choisie est 2010. La volumétrie des corpus BIB et CIT, extraits de Scopus, pour ces deux revues est présentée dans le tableau 1.

Tableau 1 : Volumétrie des corpus pour les revues et l'année charnière choisies (source Scopus)

	Nombre d'articles cités (BIB) Année de publication ≤ 2010	**Nombre d'articles ayant cité (CIT)** Année de publication ≥ 2010
***Electronics Letters* (EL)** – Y = 2010	5936	1872
***Neural Networks* (NN)** – Y = 2010	3362	749

3.2 Cas de Neural Networks

Après examen des descripteurs générés au cours de l'étape d'indexation automatique et des catégories thématiques liées, l'expert a retenu pour ce titre 19 catégories.

Tableau 2 : Représentativité des catégories thématiques dans les corpus BIB et CIT, pour la revue Neural Networks

catégorie		a_i^J	b_i^J	δ_i^J
BIO	Biochimie	0,0121899	0,01786375	0,00567386
COG	Cognition	0,14220829	0,15211148	0,0099032
COM	Informatique	0,07661242	0,09815581	0,02154339
CTL	Automatique	0,1258361	0,15450974	0,02867364
DOC	Sciences de l'information	0,00515232	0,00455181	-0,00060051
ECO	Economie	0,00988045	0,02359533	0,01371487
ELN	Electronique	0,03479926	0,12706631	0,09226706
ENE	Energie	0,01069008	0,01315519	0,00246512
MAT	Mathématiques	0,60507238	0,49071034	-0,11436205
MDI	Multidisciplinaire	0,17557234	0,05790487	-0,11766747
MED	Médecine	0,01514929	0,0256638	0,01051451

MES	Métrologie	0,07165169	0,04444137	-0,02721032
PHQ	Physique	0,10369352	0,07115716	-0,03253636
PHY	Physiologie	0,30200853	0,13330214	-0,16870639
PSY	Psychologie	0,07230644	0,03466091	-0,03764553
SIP	Traitement du signal	0,13827144	0,07321064	-0,06506079
SOC	Sociologie	0,05314277	0,03458953	-0,01855324
TEL	Télécommunications	0,01265186	0,01055698	-0,00209487
UNI	Sciences de l'univers	0,01350254	0,01232	-0,00118254

La représentativité de chacune de ces catégories dans les cartes des corpus BIB et CIT est indiquée dans le tableau 2. Ce tableau précise aussi la différence entre ces deux valeurs, qui indique l'écart de la représentativité de chaque catégorie entre les deux corpus, illustrée également par la figure 3. L'indicateur S^J obtenu pour *Neural Networks* est égal à 0,94.

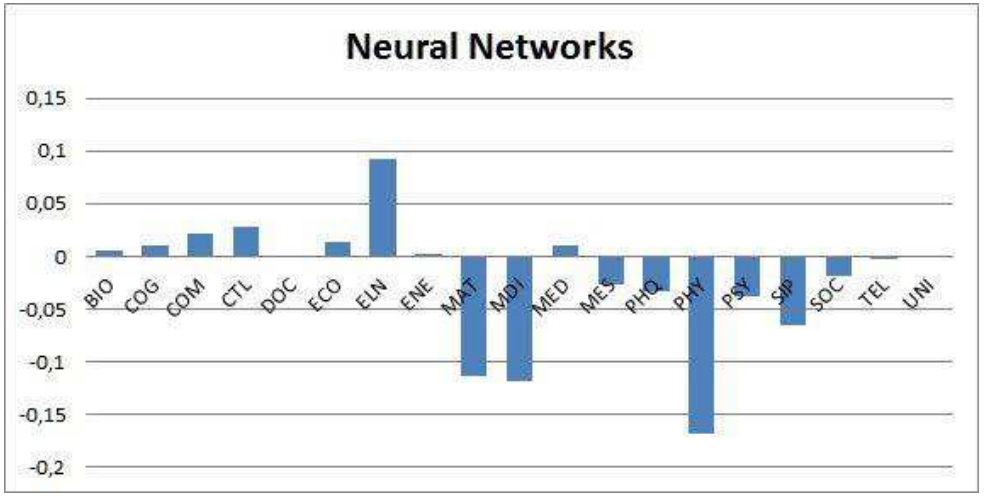

Figure 3 : Ecart de la représentativité de chaque catégorie entre les deux corpus BIB et CIT (δ_i^J), pour la revue Neural Networks

Ce calcul de différence permet de qualifier chaque catégorie selon l'évolution de sa représentativité.

- Progression : les trois catégories qui montrent la plus forte progression sont l'Electronique, l'Automatique et l'Informatique.
- Déclin : les principales diminutions sont observées pour les catégories Physiologie, Mathématiques, Psychologie et Physique (en négligeant la catégorie Multidisciplinaire, jugée non significative dans ce contexte).

Cette évolution illustre le déclin des thèmes liés à l'observation et à la modélisation du cerveau animal, en faveur de l'essor des thèmes s'intéressant au cerveau « artificiel » (intelligence artificielle, robotique…).

3.3 Cas de Electronics Letters

Après examen des descripteurs générés au cours de l'étape d'indexation automatique et des catégories thématiques liées, l'expert a retenu pour ce titre 17 catégories. Les valeurs des composantes de chacune de ces catégories dans les cartes des corpus BIB et CIT sont indiquées dans le tableau 3. Ce tableau précise aussi la différence entre ces deux valeurs, qui indique l'écart de la représentativité de chaque catégorie entre les deux corpus, illustrée également par la figure 4.

L'indicateur S^J obtenu pour *Electronics Letters* est égal à 0,99.

Ce calcul permet de qualifier chaque catégorie selon l'évolution de sa représentativité.

- Progression : les quatre catégories qui montrent la plus forte progression sont la Métrologie, la Physique, les Télécommunications et le Traitement du signal (en négligeant la catégorie Multidisciplinaire, jugée non significative dans ce contexte).
- Déclin : les principales diminutions sont observées pour l'Electromagnétisme et les Mathématiques.

Tableau 3 : Représentativité des catégories thématiques dans les corpus BIB et CIT, pour la revue Electronics Letters

catégorie		a_i^J	b_i^J	δ_i^J
BIO	Biologie	0,02320842	0,0308884	0,00767998
CHM	Chimie	0,00618036	0,01116862	0,00498826
COM	Informatique	0,03958024	0,03344178	-0,00613846
CTL	Automatique	0,02990794	0,02335954	-0,00654839
DOC	Sciences de l'information	0,00541428	0,00712537	0,00171109
ECO	Economie	0,00098372	0,003328	0,00234428
ELM	Electromagnétisme	0,25186083	0,23827099	-0,01358984
ELN	Electronique	0,42736978	0,43759714	0,01022736
ENE	Energie	0,07372052	0,04277666	-0,03094387
MAT	Mathématiques	0,16034992	0,14081303	-0,01953689
MDI	Multidisciplinaire	0,01873591	0,04272215	0,02398624
MEC	Mécanique	0,00386401	0,01332592	0,00946192
MES	Métrologie	0,05477072	0,09700894	0,04223822
OPT	Optique	0,14652549	0,1508635	0,00433802
PHQ	Physique	0,1107711	0,13764546	0,02687436
SIP	Traitement du signal	0,18099671	0,20250144	0,02150473
TEL	Télécommunications	0,19908708	0,21710672	0,01801964

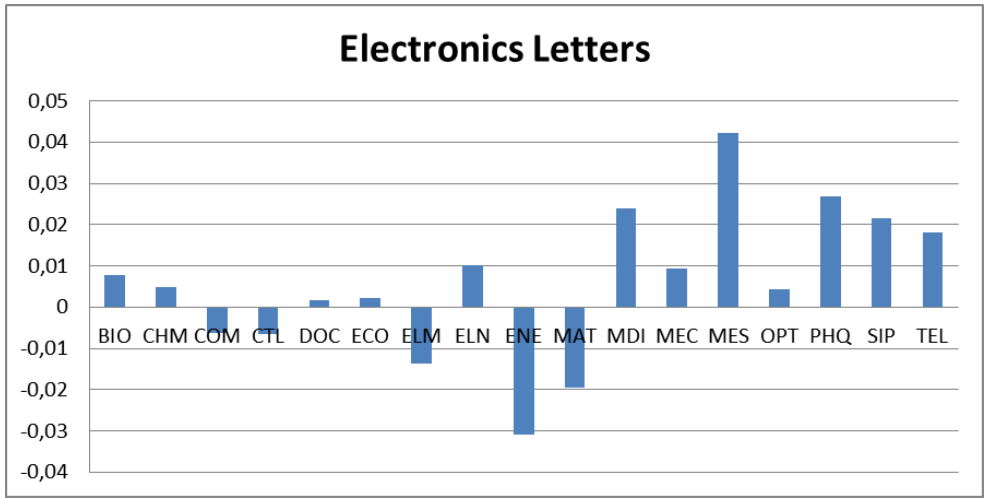

Figure 4 : Ecart de la représentativité de chaque catégorie entre les deux corpus BIB et CIT (δ_i^J), pour la revue Electronics Letters

Même si parmi les domaines en croissance figurent des thématiques appliquées au traitement et à la transmission du signal, la croissance conjointe de composantes plus fondamentales, comme la physique, ne permet pas de mettre en évidence une nette tendance d'évolution entre les deux corpus. En outre, les variations se situent dans un intervalle de valeurs beaucoup plus réduit que pour *Neural Networks*, comme nous le verrons dans la section 3.4.

3.4 Evolution comparée des deux revues

La mise en regard des valeurs de S^J, $\bar{\delta}^J$ et σ^J obtenues pour les deux revues est illustrée par le tableau 4 et la figure 5.

Tableau 4 : Valeurs de l'indicateur de rôle S^J, de la moyenne $\bar{\delta}^J$ et de l'écart-type σ^J pour chacune des revues

	S^J	$\bar{\delta}^J$	σ^J
Neural Networks	0,94	0,041	0,037
Electronics Letters	0,99	0,015	0,009

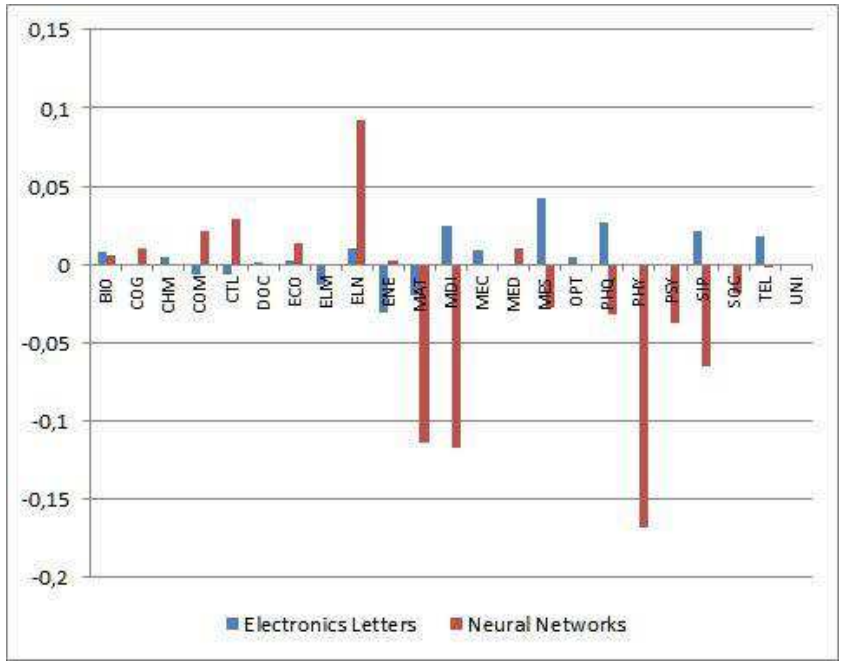

Figure 5 : Comparaison de la variation de la représentativité de chaque catégorie, pour les deux revues

L'analyse comparative des valeurs de S^J, $\bar{\delta}^J$ et σ^J obtenues pour les deux revues nous amène à des constatations nous aidant à interpréter les résultats obtenus :

- Une première constatation porte sur l'ordre de grandeur de S^J : même si la valeur calculée dans cette étude par la similarité cosinus peut être comprise entre 0 et 1, les deux résultats sont très proches de 1. Ceci est évidemment logique si l'on considère la philosophie du calcul. En effet, une valeur proche de 0 signifierait que la majorité des travaux publiés dans une revue inspirerait des études quasi exclusivement orientées vers un tout autre horizon scientifique que celui relatif aux études ayant contribué à la réalisation des travaux qui y sont publiés, ce qui paraît très peu probable.
- Un éclairage complémentaire nous est apporté par les valeurs de $\bar{\delta}^J$ et σ^J, qui nous renseignent sur l'échelle de dispersion de la variation de la représentativité de chaque catégorie entre les deux corpus BIB et CIT associés à une revue. A ce sujet, nous constatons que la valeur de $\bar{\delta}^J$ obtenue par *Neural Networks* est largement supérieure à celle d'*Electronics Letters*, ce qui montre qu'en moyenne cette variation est plus importante pour le premier. La comparaison des écarts-types vient confirmer cette tendance.

Par conséquent, nous sommes amenés à conclure que le rôle de vecteur de transfert de connaissances d'*Electronics Letters* reste majoritairement confiné dans son périmètre éditorial : le paysage scientifique des articles auxquels se réfèrent les auteurs qui y publient leurs travaux est très proche de celui des travaux qui s'appuient sur les publications de la revue. Son comportement peut donc être qualifié de largement autarcique.

A l'opposé, *Neural Networks* joue son rôle de vecteur de transfert de connaissances au-delà des limites définies par son périmètre éditorial. Nous supposons que cette revue sert de passerelle entre les scientifiques travaillant sur des domaines différents mais liés, via entre autres la recherche de la transposition du comportement animal à la machine. Elle montre donc un comportement plutôt expansionniste.

3.5 Illustration des résultats par cartographie thématique des corpus

A la demande de l'expert, pour illustrer les résultats et les compléter, nous avons cartographié les deux corpus de chaque périodique à l'aide de l'outil VOSViewer [8], qui produit une carte thématique bidimensionnelle dans laquelle les termes sont positionnés de sorte que la distance entre deux termes donne une indication du lien entre eux : en général, plus cette distance est petite, plus fort est le lien que unit ces deux termes. La force du lien entre les termes est fondée sur leur cooccurrence dans les notices qui composent le corpus.

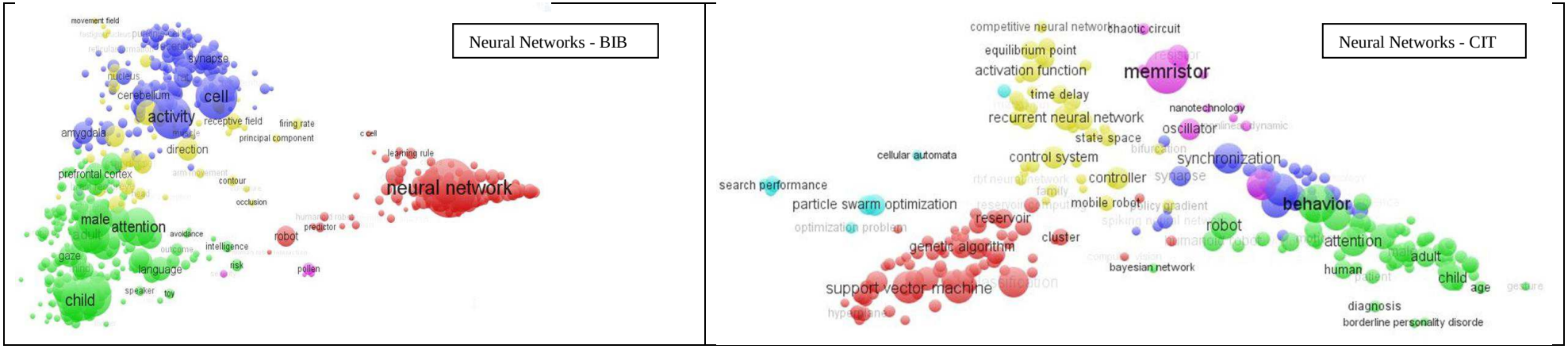

Figure 6 : Cartes obtenues pour les corpus BIB et CIT de Neural Networks

La figure 6 montre les cartes obtenues pour les corpus BIB et CIT de *Neural Networks*. L'expert a détecté dans la carte de BIB deux pôles thématiques presque indépendants : l'un autour du terme « *neural network* », positionné à droite de la carte, et l'autre combinant les aspects de la psychologie et la physiologie humaines, positionné sur la gauche. Il apparaît toutefois un signal d'interaction, par la présence de terme « *robot* » situé au milieu de la carte et qui semble jeter un pont entre les deux pôles. Dans la carte de CIT, le second pôle voit son périmètre réduit, comme on peut l'observer dans la zone inférieure droite de la carte. Malgré tout, il reste lié l'ensemble des autres classes, qui sont une combinaison des sujets présents dans les deux pôles thématiques détectés dans BIB, avec une visée beaucoup plus appliquée au développement du « cerveau artificiel ». Le terme « *neural network* » de BIB va alors s'effacer au profit d'autres plus spécifiques, comme « *competitive neural network* » ou « *recurrent neural network* » que l'on trouve dans CIT.

Cette observation corrobore les résultats issus de notre méthodologie, à savoir la décroissance notable des thématiques « Physiologie » et « Psychologie », au profit de l'émergence de thématiques comme « Automatique », « Electronique » ou « Informatique ». En effet, ces thématiques sont au cœur des applications liées au développement du « cerveau artificiel » (cf. §3.2).

La figure 7 montre les cartes obtenues pour les corpus BIB et CIT d'*Electronics Letters*. Bien que la morphologie des cartes soit différente, l'expert y a détecté une certaine stabilité au niveau des thématiques, illustrées par les termes, et de leur représentativité. L'apparition de la thématique « Métrologie », qu'avait révélée notre méthodologie, est confirmée par la présence dans la carte de CIT du terme « *sensor* », qui n'avait pas été mis en évidence dans BIB. Cette stabilité vient à l'appui des résultats de notre méthodologie, qui a mis en évidence un intervalle de variation très réduit des écarts de représentativité des thématiques pour la revue (cf. §3.3).

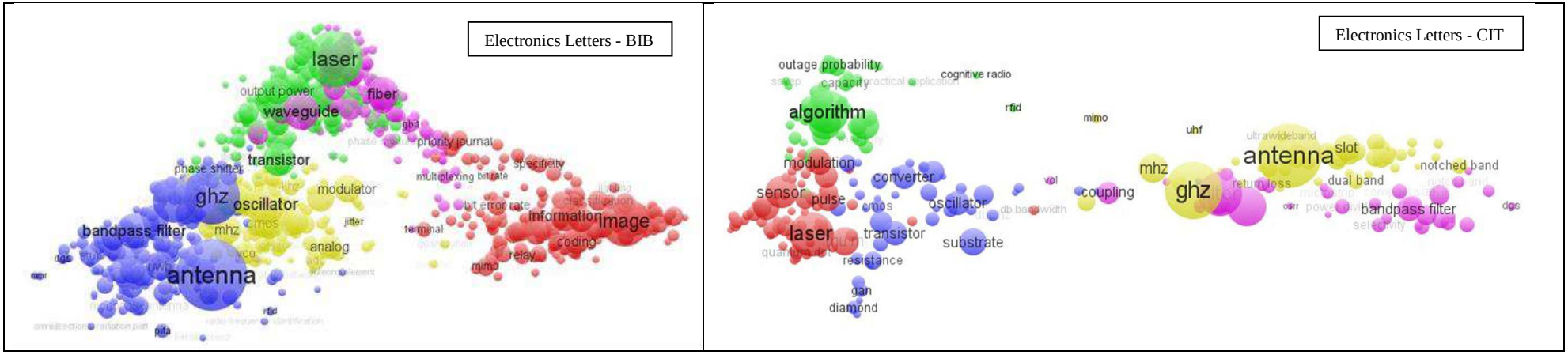

Figure 7 : Cartes obtenues pour les corpus BIB et CIT d'Electronics Letters

La cartographie thématique des corpus confirme donc les résultats obtenus précédemment. Cependant, la tâche de détermination du rôle joué par chaque périodique en tant que vecteur de transfert de connaissances incombant à l'expert aurait été sensiblement alourdie s'il avait dû fonder son analyse comparative des corpus BIB et CIT uniquement sur les cartes thématiques. Notre méthodologie permet ainsi d'assister l'expert dans cette tâche en mettant à sa disposition des informations calculées à partir d'une catégorisation thématique opérée sur les termes décrivant le contenu des notices de chaque corpus.

4 Conclusion et perspectives

La méthode que nous avons développée permet de construire un indicateur qualifiant le rôle d'une revue scientifique en tant que vecteur de transfert de connaissances.

Le travail se poursuit en suivant plusieurs pistes dont la difficulté de mise en œuvre est cependant variable :

- Etudier l'apport d'une catégorisation sémantique en remplacement de la catégorisation thématique
- Envisager une possibilité de hiérarchisation des catégories thématiques, pour offrir ce choix de catégorisation à l'expert

- Etudier les modalités de prise en compte les données classées dans la catégorie « Multidisciplinaire »
- Mettre en œuvre d'autres calculs de similarité
- Etendre l'expérimentation à d'autres disciplines scientifiques

Cet indicateur, expérimenté sur des revues, pourrait également avoir d'autres applications, comme la caractérisation en tant que vecteur de transfert de connaissances de la production scientifique d'un chercheur, d'un établissement, d'une région… La difficulté résidera alors dans la collecte des données. Elle est en particulier due à des problèmes d'homonymie, d'hétérogénéité dans l'écriture des affiliations des auteurs et d'obtention d'un corpus pouvant être considéré comme représentatif de la production scientifique de la population étudiée.

5 Bibliographie

[1] LEYDESDORFF L., CARLEY S., RAFOLS I.: *Global maps of science based on the new Web-of-Science categories*, Scientometrics, 2013, 94(2), p 589-593

[2] GUANG Y., MING-YANG W., DA-REN Y.: *Characterizing knowledge diffusion of Nanoscience & Nanotechnology by citation analysis*, Scientometrics, 2010, 84, p 81–97, DOI 10.1007/s11192-009-0090-2

[3] LARSEN K.: *Knowledge network hubs and measures of research impact, science structure, and publication output in nanostructured solar cell research*, Scientometrics, 2008, 74(1), p 123-142

[4] CARLEY S, PORTER A. L.: *A forward diversity index*, Scientometrics, 2012, 90(2), p 407-427

[5] POLANCO X., FRANCOIS C., ROYAUTE J., BESAGNI D., ROCHE I., *STANALYST: An integrated environment for clustering and mapping analysis on science and technology*, 8th ISSI conference, Sydney, Australia, Proceedings, 2001, Vol. 2, p 871-873

[6] ROYAUTE J., FRANCOIS C., ZASADZINSKI A., BESAGNI D., DESSEN P., LE MINOR S., MAUNOURY M-T. : *Approche terminologique et infométrique de fouille de données textuelles dans un corpus sur la génétique des cancers de la thyroïde*, Revue des Nouvelles Technologies de l'Information, 2004, II, p 465-476

[7] ROYAUTE J., FRANÇOIS C., ZASADZINSKI A., BESAGNI D., DESSSEN P., LE MINOR S., MAUNOURY M-T : *Mining corpora of texts on genes involved in thyroid cancers : a bioinformatic text mining and clustering process*, Proceedings of the European Conference on Computational Biology (ECCB), 2003, p 77-78

[8] VAN ECK, N.J., WALTMAN, L., *Software survey: VOSviewer, a computer program for bibliometric mapping*, Scientometrics, 2010, 84(2), p 523-538