Évaluation du potentiel d'applicabilité d'un projet de recherche: vers une méthodologie fondée sur l'analyse de contenu

Ivana Roche, Nathalie Vedovotto, Claire François, Dominique Besagni, Pascal Cuxac, Marianne Hörlesberger, Dirk Holste, Edgar L. Schiebel

EVALUATION DU POTENTIEL D'APPLICABILITE D'UN PROJET DE RECHERCHE : VERS UNE METHODOLOGIE FONDEE SUR L'ANALYSE DE CONTENU[1]

Ivana Roche (*), Nathalie Vedovotto (*), Claire François (*), Dominique Besagni (*), Pascal Cuxac (*), Marianne Hörlesberger (**), Dirk Holste (**), Edgar Schiebel (**)
ivana.roche@inist.fr, nathalie.vedovotto@inist.fr, claire.francois@inist.fr, dominique.besagni@inist.fr, pascal.cuxac@inist.fr, marianne.hoerlesberger@ait.ac.at, dirk.holste@ait.ac.at, edgar.schiebel@ait.ac.at

(*) INIST-CNRS, 2 allée du Parc de Brabois, 54519 Vandoeuvre-lès-Nancy Cedex, France
(**)AIT, Austrian Institute of Technology GmbH, Donau-City-Strasse 1, 1220 Vienna, Austria

Mots clefs :
Processus de sélection, évaluation de projet, expertise, analyse de contenu, applicabilité, indicateur
Keywords:
Selection process, project evaluation, expertise, content analysis, applicability, indicator
Palabras clave:
Proceso de selección, evaluación proyecto, trabajo experto, análisis de contenido, aplicabilidad, indicador

Résumé

Lorsque la production scientifique devient trop complexe et sophistiquée pour être appréhendée par un évaluateur, les méthodes infométriques peuvent s'avérer utiles, soit en appui du processus de prise de décision, soit pour évaluer ce dernier. En fait, l'évaluation infométrique pourrait répondre au besoin croissant de suivi des résultats de la science. C'est dans ce contexte que s'inscrit ce travail, en proposant une méthodologie d'évaluation du potentiel d'applicabilité d'un projet de recherche soumis à une agence de financement. Notre approche repose sur une analyse de contenu opérée à l'aide d'outils de traitement automatique du langage (TAL) et de classification automatique. L'objectif est de faciliter l'étape d'expertise, qui reste néanmoins incontournable. Notre travail est illustré par le traitement d'un cas réel extrait des résultats d'une prestigieuse agence de financement européenne qui met en œuvre un processus de sélection, fondé sur l'excellence scientifique de la recherche exploratoire comme seul critère de décision.

[1] Ce travail est partiellement réalisé dans le contexte du projet DBF (*Development and Verification of a Bibliometric Model for the Identification of Frontier Research*), action de soutien (CSAs) du programme IDEAS du 7ème PCRD de la Commission Européenne (référence du projet n° 240765). Les auteurs lui savent gré de ce soutien.

Introduction

La question que nous abordons dans ce travail est l'évaluation de l'applicabilité potentielle des travaux qu'un chercheur présente lors de conférences et/ou publie dans la littérature scientifique et technique. Nous nous sommes confrontés à cette problématique dans le cadre d'un projet européen [1] où nous avons développé une méthodologie d'analyse fondée sur la modélisation de critères définis par l'ERC (European Research Council) et mis en œuvre par leurs experts scientifiques lors du processus de sélection de projets de recherche en vue de leur financement.

Bien sûr, nous ne pouvons pas nous attendre à ce qu'un modèle numérique se substitue à l'expertise humaine ou à la reconnaissance de la communauté scientifique, toutes deux étant très difficiles à quantifier. Des modèles de ce type peuvent néanmoins servir à vérifier des décisions, fournir des informations complémentaires ou mettre en évidence des biais dans le processus de sélection [2]. La nature et les objectifs des critères mis en œuvre dans la sélection des projets de recherche sont très diversifiés : identité du porteur du projet, risque associé au projet, sujet peu conventionnel… Ceci est particulièrement évident dans le choix des indicateurs définis pour modéliser la stratégie, les missions et la politique scientifique des agences de financement pour établir une relation de cause à effet interprétable et utile. On observe des divergences entre les choix des experts et les propositions issues des indicateurs infométriques. Plusieurs études en ont approfondi les raisons sous-jacentes [2-4].

Parmi les critères de l'ERC, nous trouvons ainsi l'applicabilité, caractérisée comme suit : "… *may well be concerned with both new knowledge about the world and with generating potentially useful knowledge at the same time. Therefore, there is a much closer and more intimate connection between the resulting science and technology, with few of the barriers that arise when basic research and applied research are carried out separately.*" [5].

Donald Stokes [6] a introduit une approche permettant de distinguer entre recherche fondamentale et appliqué en définissant un schéma bidimensionnel appelé Quadrant de Pasteur. Il s'agit d'une caractérisation des travaux de recherche permettant de classer les relations existant entre les sciences fondamentales et l'innovation technologique. Les travaux de Louis Pasteur sont considérés comme un parfait exemple d'une recherche à la fois scientifiquement importante et susceptible d'applications pratiques. Le résultat est une caractérisation de trois types distincts de recherche :

- la recherche purement fondamentale, illustrée par les travaux de Niels Bohr, physicien danois du début du 20ème siècle ;
- la recherche purement appliquée, illustrée par les travaux de Thomas Edison, inventeur et industriel nord-américain ;
- la recherche fondée sur la théorie mais à visée appliquée, décrite comme le Quadrant de Pasteur, scientifique français, chimiste et physicien de formation, pionnier de la microbiologie.

Dans cet article, nous définissons d'abord l'approche utilisée pour évaluer le potentiel d'applicabilité d'un projet, puis nous décrivons le traitement des données. Pour conclure nous présentons et commentons les résultats obtenus.

1 Contexte

Une solution classiquement adoptée pour déterminer le degré d'application des travaux d'un chercheur amène à s'intéresser aux éventuels brevets au développement desquels il(elle) a pris part. [e.g. 7-9]. La soumission d'un brevet est souvent l'aboutissement d'une démarche de transfert de technologie et peut en effet être considérée comme l'accomplissement pratique de travaux de recherche présentant des caractéristiques résolument appliquées. Une autre

possibilité est d'examiner directement les travaux publiés par le chercheur pour caractériser leur contenu en appliqué ou fondamental. Dans une étude précédente, nous avons modélisé ces deux critères à l'aide d'indicateurs calculés à partir des données affichées par le chercheur dans les documents consignés lors de la soumission de son projet, à savoir, le nombre de brevets auxquels il(elle) a contribué et les titres des périodiques où ses travaux, présentés dans son *Curriculum Vitae*, ont été publiés.

Le premier indicateur est fondé sur un simple dénombrement et présente des valeurs entières comprises dans l'intervalle $[0, \infty[$. Or, le nombre de brevets dans les *Curriculum Vitae* est souvent faible, ce qui produit un impact négatif sur la justesse de l'indicateur concerné. Le second résulte du calcul de la part des travaux du chercheur publiés dans des périodiques dont le contenu est catégorisé comme appliqué. Cet indicateur affiche des valeurs réelles variant de 0 à 1. Pour les deux indicateurs nous supposons alors que plus haute est leur valeur, plus affirmée est l'applicabilité des travaux du chercheur. Si cette approche paraît pragmatique, elle présente néanmoins quelques faiblesses dues en particulier à l'étape de catégorisation des périodiques. En effet, en appliquant à ces derniers une catégorisation binaire "appliquée ou fondamentale" il paraît *a priori* aisé de transposer automatiquement la catégorie du périodique à tous les articles que y sont publiés. La première difficulté consiste en la détermination des critères permettant d'obtenir la catégorisation des périodiques car, si beaucoup a été fait pour les caractériser selon un classement hiérarchique arborescent plus ou moins détaillé des domaines scientifiques, le problème reste entier lorsqu'il faut déterminer lesquels sont résolument appliqués ou fondamentaux.

De plus, la catégorie d'un périodique peut ne pas être unique mais dépendre du domaine scientifique du chercheur qui y publie ses travaux. Considérons, par exemple, le domaine de la Biologie comme *a priori* fondamental. Toutes les sources classées dans ce domaine reçoivent donc la catégorie "fondamentale" de même que tous les articles qui y sont publiés. Or cela ne correspond pas toujours à une réalité. En effet, si cela reste vrai pour les publications des biologistes dans une de ces sources, celle d'un informaticien qui apporterait un développement logiciel appliqué à la Biologie devrait, elle, recevoir la catégorie "appliquée".

Aussi, dans ce travail, nous nous affranchissons de cette étape de catégorisation en proposant un indicateur plus évolué, fondé sur une analyse de contenu opérée par un expert scientifique sur, d'une part, l'ensemble des publications qui citent au moins une des publications du chercheur et, d'autre part, l'ensemble des publications qui possèdent au moins une référence citée commune avec l'ensemble des références citées par le chercheur dans la bibliographie de son projet et sur lesquelles ce dernier est supposé trouver, en partie, ses fondements.

Concernant le premier ensemble, notre hypothèse est que, par la voie de la citation, les publications citantes expriment l'exploitation des publications du chercheur et que, par conséquent, elles constituent une source d'information réelle et pragmatique sur l'utilisation des résultats de ses travaux passés dans le cadre de nouvelles recherches. L'analyse de contenu réalisée sur cet ensemble nous donne le moyen d'apprécier l'utilisation des travaux du chercheur publiés jusqu'au moment de la soumission de son projet. Pour le second ensemble, nous supposons que les publications ayant une ou plusieurs citations communes avec le projet du chercheur peuvent représenter le contexte scientifique dans lequel celui-ci vient également s'inscrire. L'analyse de contenu opérée sur ce corpus nous permet de qualifier le degré d'application de ces travaux qui, se fondent sur un même socle de connaissances que le projet. Alors, par analogie, on associe au projet ce même degré d'application. Finalement, la comparaison des résultats de ces deux analyses apporte une réponse à notre interrogation sur l'aspect évolutif du degré d'application des travaux d'un chercheur et nous permet de déduire leur potentiel d'applicabilité.

Cette démarche nous permet, en outre, de considérer deux niveaux d'application : le premier correspondant à la simple utilisation des résultats publiés dans des nouveaux travaux s'y référant et le second amenant, *via* l'analyse de leur contenu, à la caractérisation de ces derniers à l'aune d'un critère ne comprenant que deux valeurs possibles : appliqué ou fondamental.

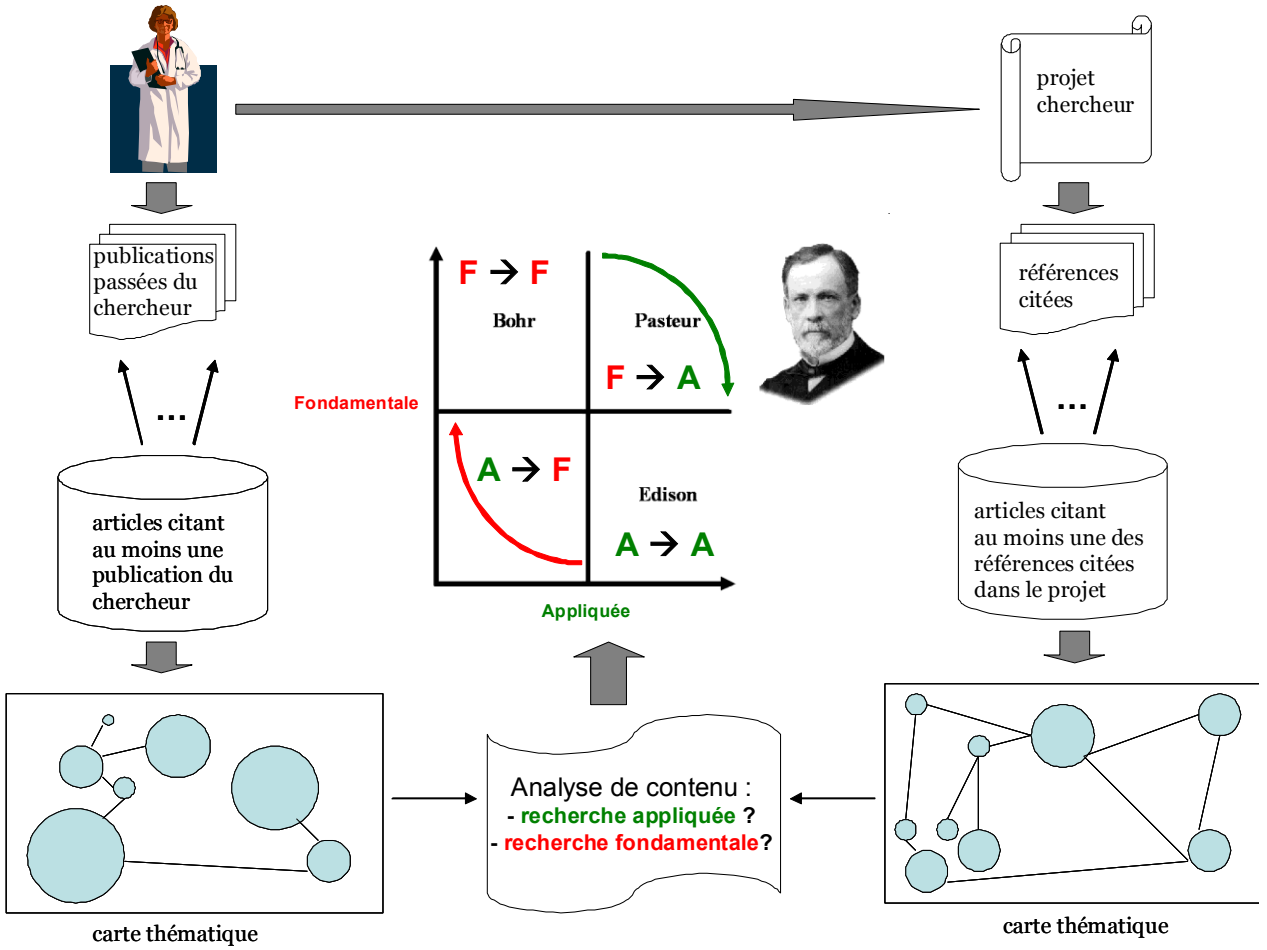

Figure 1. Schéma méthodologique d'un processus d'évaluation du potentiel d'applicabilité d'un projet de recherche

Notre analyse de contenu produit une représentation cartographique de chacun des deux ensembles considérés, sous la forme d'une carte thématique organisant leur contenu en classes de publications similaires. L'examen de ces deux cartes de classes permet à des experts du domaine scientifique concerné d'y détecter la présence de thématiques présentant des caractéristiques appliquées et d'en évaluer l'importance relative. De plus, la comparaison des résultats obtenus pour chacune des deux cartes thématiques rend alors possible une caractérisation de l'évolution du degré d'application des travaux du chercheur. Quatre types de « parcours » scientifiques peuvent être observés. Deux montrent une réelle stabilité : les travaux passés et présents du chercheur sont et demeurent à visée soit appliquée, soit fondamentale. Les deux autres parcours sont, de notre point de vue, bien plus intéressants car ils montrent à proprement parler une transition :

- soit vers l'applicatif : le chercheur passe du fondamental à l'appliqué ;
- soit vers le fondamental : le chercheur, à l'origine appliqué, passe au fondamental.

Nous avons représenté ces différents parcours à l'aide du Quadrant de Pasteur (cf. figure 1).

Si le parcours vers l'appliqué est aisément interprétable, celui qui montre une transition vers le fondamental appelle un questionnement : cette transition ne peut-elle être interprétée comme un simple passage obligé, nécessaire pour acquérir des fondements théoriques qu'ensuite le chercheur peut mettre en œuvre dans le cadre de nouvelles applications ?

2 Méthodologie

Les données primaires sont directement extraites des documents déposés lors de la soumission du projet, en l'occurrence, le nom du chercheur et les références que celui-ci cite dans la bibliographie de son projet. A partir du nom du chercheur nous allons déterminer une liste de publications consistant en sa production publiée dans la littérature scientifique et technique et répertoriée dans une source de données bibliographiques permettant, en outre, d'accéder aux références citées. Cette liste obtenue, il est alors aisé de singulariser l'ensemble des publications qui citent au moins un des articles du chercheur. Ce premier corpus est considéré comme la représentation du contexte scientifique relatif aux publications qui emploient, à plus ou moins grande échelle, les connaissances véhiculées par les travaux passés du chercheur.

D'autre part, à l'exception des auto-citations, toutes les références citées par le chercheur dans la bibliographie de son projet sont considérées. Pour chacune nous recherchons toutes les publications citantes. L'ensemble de ces dernières constitue le second corpus, que nous considérons comme une représentation des connaissances qu'elles partagent avec le projet du chercheur. Une analyse de contenu est alors opérée sur chacun de ces deux corpus. Pour réaliser cette analyse, les notices bibliographiques du corpus sont reformatées et intégrées dans la station d'analyse de l'information Stanalyst. Une première étape de fouille de textes, appliquant des techniques fondées sur le traitement automatique des langues, opère une indexation assistée des notices en leur associant des mots-clés. Une approche classificatoire est ensuite appliquée à ce corpus de notices bibliographiques enrichies. L'outil logiciel employé, implémenté dans Stanalyst, met en œuvre un algorithme de classification automatique non supervisée et non hiérarchique, la méthode des K-means axiales, inspirée du formalisme des cartes auto-adaptatives de Kohonen. Cette méthode emploie les mots-clés comme des indicateurs du contenu des notices bibliographiques qui, à leur tour, sont considérées comme des indicateurs des thématiques de recherche. Cette étape est suivie d'une analyse en composantes principales permettant le positionnement des classes sur une carte 2D. Les relations entre les classes sont ensuite utilisées pour la construction de réseaux thématiques qui vont constituer, si l'on s'accorde une métaphore géographique, une carte du domaine de recherche représenté par le corpus.

L'expert procède alors, pour chacun des deux corpus, à l'analyse des résultats obtenus, à savoir, le contenu des classes ainsi que leur positionnement et leurs relations dans la carte de classes. Lors de cette analyse, l'expert doit appliquer une grille de lecture particulière, à même de l'amener à une évaluation du caractère majoritairement appliqué ou fondamental du contenu de chaque corpus. Il(elle) s'intéresse pour ce faire au contenu de chaque classe des deux cartes, examinant le titre et les mots-clés associés à chacune des notices afin d'évaluer le score d'application de ces classes. Pour chaque classe, il(elle) :

- détermine la proportion de sujets fondamentaux et appliqués, respectivement Pf et Pa (dont la somme est égale à 1)
- détermine le score d'application de chaque classe, égal à (Pa – Pf) et compris dans l'intervalle [-1, 1]. Une valeur égale à -1 correspond à une classe totalement fondamentale, une valeur égale à 1 à une classe totalement appliquée, et les valeurs intermédiaires indiquent des classes présentant ces deux caractéristiques.

La somme des valeurs attribuées à chaque classe d'une carte fournit le degré d'application du corpus. Si cette valeur est négative, le corpus traite majoritairement de recherche fondamentale, et inversement, si cette valeur est positive, le corpus est majoritairement concerné par la recherché appliquée.

Cette étape d'expertise réalisée sur les deux corpus va ensuite permettre de les comparer et de déterminer si le parcours scientifique du chercheur est stablement localisé dans le périmètre de la recherche soit appliquée, soit fondamentale, ou bien s'il montre une transition entre les deux.

3 Résultats

Dans cette étude, pour illustrer la méthodologie présentée nous l'avons appliquée à un cas d'étude issu de la campagne 2009 d'appel à projets de l'ERC. Aussi, parmi les 25 grands domaines couverts par l'ERC, nous avons choisi celui de l'Ingénierie de la communication et des systèmes. Dans ce domaine, 31 projets ont été soumis et nous en avons choisi un parmi les quatre qui ont été sélectionnés par le panel d'experts de l'ERC en charge de ce domaine scientifique. Pour des raisons de confidentialité, nous ne pouvons pas afficher des données nominatives ou bien des informations permettant, par recoupement, de les retracer. Appelons donc le chercheur porteur du projet CHE. La recherche sur le WoS (Web of Science) à partir du nom du chercheur a ramené, après vérification de l'inexistence d'homonymies ou d'artefacts, 24 publications s'étalant de 2000 à 2009. Pour leur part, toujours selon le WoS, ces publications reçoivent des citations de 663 publications, que nous avons collectées et qui constituent notre premier corpus que, par la suite, nous appellerons FC. D'autre part, dans son projet, CHE présente une bibliographie citant au total 45 références. Parmi elles, 5 sont des auto-citations, que nous écartons, et 25 ne sont pas présentes dans le WoS. Les 15 références restantes, dont la date de publication est comprise dans l'intervalle [2000, 2007], reçoivent des citations de 4612 publications, que nous avons extraites et qui vont former notre second corpus, appelé par la suite SC.

Les notices de FC et SC ont alors subi un reformatage leur permettant d'intégrer la plateforme d'analyse Stanalyst et ont été enrichies avec une indexation par des mots-clés obtenue grâce à l'application d'outils de fouille de texte fondés sur des techniques de TAL (traitement automatique de langue). Cette étape d'indexation assistée a, de plus, bénéficié d'une expertise scientifique à deux niveaux : lors de la constitution du vocabulaire servant de référentiel terminologique et, d'autre part, lors de la validation du résultat final avec l'élimination de termes jugés trop génériques. La table 1 résume les caractéristiques principales des deux corpus indexés.

Table 1. Caractéristiques des deux corpus indexés

Corpus	Nombre de références du corpus	Nombre de mots-clés d'indexation	Nombre de classes	Nombre de mots-clés dans les classes	Références dans les classes	
					nombre	% du corpus initial
FC	663	2210	20	1000	662	99,85%
SC	4612	4915	20	3010	4608	99,91%

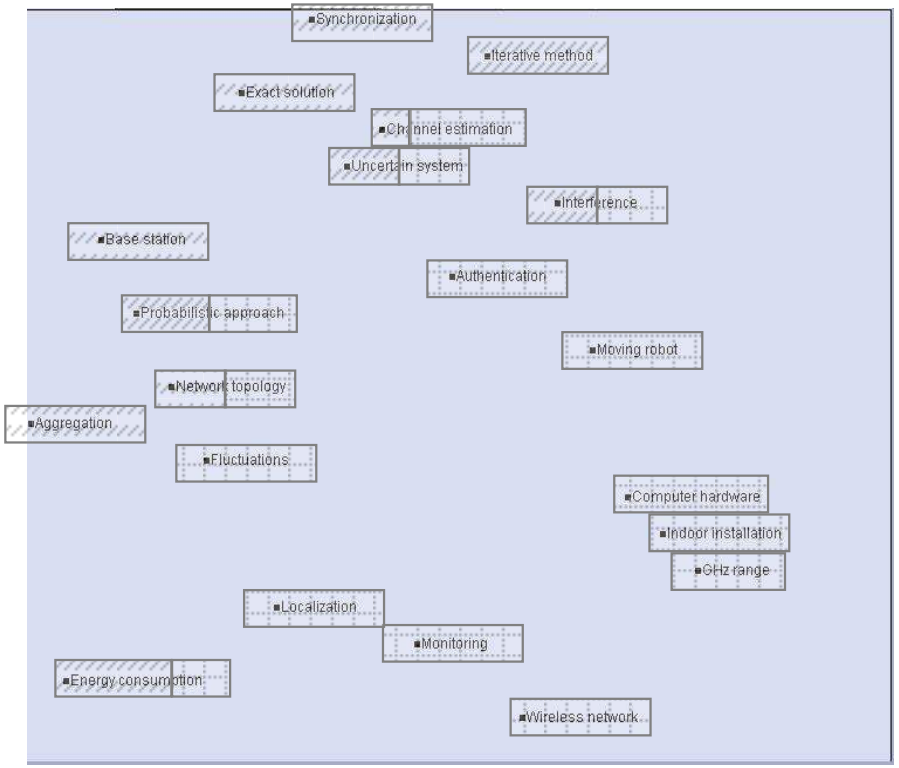

Figure 2. Cartographie du premier corpus (FC), illustrant la proportion fondamental-appliqué de chaque classe, sous forme de rectangles texturés (quadrillage= appliqué, hachures= fondamental)

Les cartes obtenues pour FC et SC ainsi que les résultats de l'expertise en termes de "niveau d'application" de chaque cluster sont présentés dans les figures 2 et 3. Par exemple, les classes "synchronization", "energy consumption", "interference", "channel estimation" et "monitoring" de la figure 2 obtiennent respectivement la valeur -1 ; -0,34 ; 0 ; 0,5 et 1.

Le degré d'application calculé pour chaque corpus est : 4.17 pour FC et 8.15 pour SC. Ces valeurs montrent que les deux corpus sont appliqués, mais on observe une nette évolution entre FC et SC. En effet, en comparant les deux cartes, nous constatons que FC présente cinq classes considérées comme totalement fondamentales, alors que SC n'en comporte aucune. De plus, certaines classes présentes dans les deux cartes, par exemple "energy consumption", montrent une augmentation notable de leur score d'application dans SC.

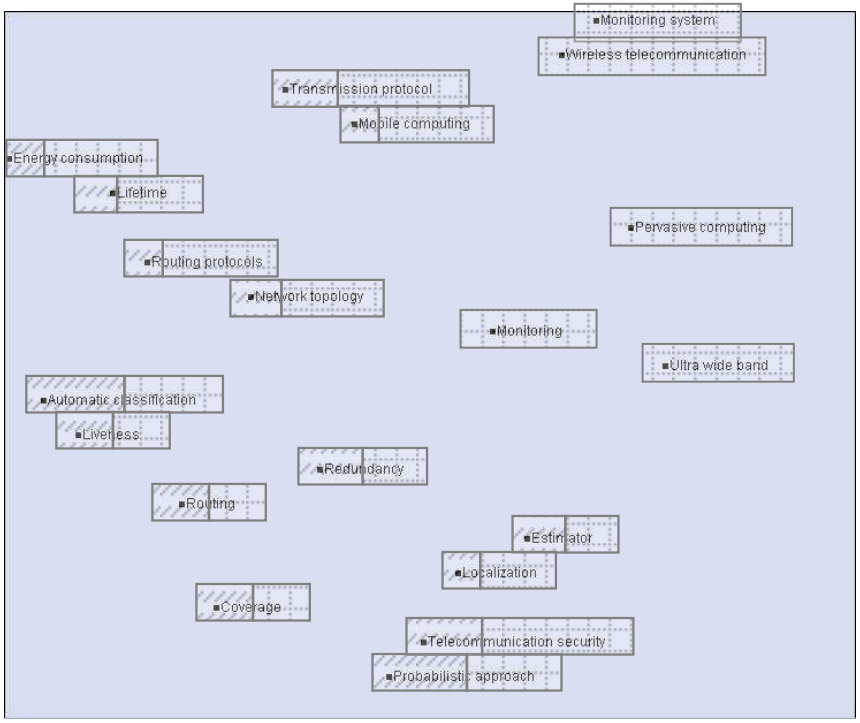

Figure 3. Cartographie du second corpus (SC), illustrant la proportion fondamental-appliqué de chaque classe, sous forme de rectangles texturés (quadrillage= appliqué, hachures= fondamental)

Cette augmentation du degré d'application entre FC et SC nous permet d'émettre l'hypothèse que, même si le "parcours" scientifique du chercheur est toujours localisé dans un périmètre appliqué, le potentiel d'applicabilité de son projet est important. Nous en concluons que ce projet répond favorablement au critère d'applicabilité défini par l'agence de financement.

4 Conclusions et perspectives

L'approche développée vise à fournir une méthodologie d'estimation du potentiel d'applicabilité d'un projet de recherche. Notons que cette étude repose sur un ensemble de données représentatives, car fournies par l'agence de financement elle-même. Par la suite, une source de données bibliographiques est employée afin d'enrichir cet ensemble, avec les publications citant celles du chercheur d'une part, et d'autre part avec celles partageant au moins une des références citées dans le projet du chercheur. L'analyse de contenu qui est alors appliquée à ces deux ensembles de données fournit une aide au travail d'expertise scientifique. Au final, cette méthodologie pourrait opérer une évaluation bibliométrique *a priori* du potentiel d'applicabilité des projets soumis, afin d'assister l'agence de financement lors du processus de sélection.

Dans une étape ultérieure, nos travaux pourraient porter sur l'amélioration de cette méthodologie, par
- une pondération du score d'application obtenu pour chaque classe, par un paramètre issu des résultats de classification automatique, afin de contextualiser la classe par rapport à son environnement, à savoir la carte des classes
- l'introduction d'un outil d'aide à la décision proposant à l'expert, pour validation, une valeur du score d'application de chaque classe, calculée en tenant compte d'une catégorisation sémantique existante des mots-clés.

Il n'en reste pas moins que la décision finale demeure la prérogative de l'agence de financement et ne s'appuie pas que sur ce seul indicateur.

5 Références bibliographiques

[1] HOLSTE D., ROCHE I., HÖRLESBERGER M., BESAGNI D., SCHERNGELL T., FRANCOIS C., CUXAC P., SCHIEBEL E., submitted to Scientometrics, 2012

[2] JUZNIC P., PECLIN S., ZAUCER M., MANDELJ T., PUSNIK M, DEMSAR F., Scientometric indicators: peer review, bibliometric methods and conflict of interest. Scientometrics, 2010, 85, p. 429-441

[3] BESSELAAR, P.v.d. & LEYDESDORFF, L., *Past performance, peer review and project selection: a case study in the social and behavioral sciences*. Research Evaluation, 2009, 18, p. 273-288

[4] BORMANN L., LEYDESDORFF L., BESSELAAR P.v.d., *A Meta-evaluation of Scientific Research Proposals: Different Ways of Comparing Rejected to Awarded Applications*. Journal of Informetrics, 2009, 4, p. 211-220

[5] EC – EUROPEAN COMMISSION, *Frontier research: The European Challenge*. High Level Expert Group Report, EUR 21619, 2005

[6] STOKES D., *Pasteur's Quadrant* - Basic Science and Technological Innovation, Brookings Institution Press, 1997

[7] GLÄNZEL W., MEYER M., *Patents cited in the scientific literature: An exploratory study of "reverse" citation relations*. Scientometrics, 2003, 58, p. 415-428

[8] MOED H.F., GLÄNZEL W., SCHMOCH U., *Handbook of quantitative science and technology research: The use of publication and patent statistics in studies of S&T systems*. Kluwer Academic Publishers, 2004

[9] GLÄNZEL W., ZHOU P., *Publication activity, citation impact and bi-directional links between publications and patents in biotechnology*. Scientometrics, 2011, 86, p. 505-525