Détection de techniques prometteuses à partir de méthodes bibliométriques

Ivana Roche, Claire François, Dominique Besagni

Détection de techniques prometteuses à partir de méthodes bibliométriques

Ivana Roche (1)
ivana.roche@inist.fr

Claire François (1)
claire.francois@inist.fr

Dominique Besagni (1)
dominique.besagni@inist.fr

(1) INIST, CNRS, 2 allée du Parc de Brabois, CS 10310, 54514 Vandoeuvre les Nancy cedex

Mots-clés : technologies prometteuses, bibliométrie, approche experts, plan de classement, indicateur de croissance.

Keywords: promising technologies, bibliometrics, expert approach, classification scheme, growth indicator.

Résumé : Ce travail s'inscrit dans un projet européen dont le but est l'identification des technologies émergeantes et prometteuses ayant des liens forts avec le domaine de la Physique, et l'identification des communautés scientifiques développant ces technologies. Nous présenterons les critères qui ont guidés notre choix de la base documentaire utilisée pour cette étude, puis nous décrirons notre approche méthodologique qui combine les analyses bibliométriques fondées sur la notion de plan de classement et les analyses d'experts pour sélectionner les thématiques par étapes successives partant d'une centaine de candidates pour aboutir in fine à 10 technologies prometteuses. Nous conclurons par une discussion sur les problématiques associées à ce type de projet : comment caractériser la dynamique d'une thématique, voire des émergences ?

Abstract: This work is part of an European project the goal of which is to identify emerging and promising technologies having strong links with the domain of Physics, and to identify the scientific communities developing these technologies. We present the criteria we followed to choose the bibliographic database exploited in this study. Then we describe our methodological approach that combines both bibliometric analysis based on classification schema and expert analysis in order to select themes, step by step, starting from one hundred of candidates to 10 promising technologies. Finally, we discuss about the problems related to this kind of project: how to characterize the dynamics of a domain, or even the emergences?

Introduction

Plusieurs analyses économiques montrent que l'innovation technologique contribue considérablement à la croissance économique. Cet effet a augmenté cette dernière décennie. De nouvelles technologies comme la microélectronique, les technologies de l'information ou les biotechnologies ont induit l'émergence, ou au moins une croissance importante, de secteurs industriels, et ont changé le profil des pays industriels. Il y a donc un intérêt grandissant pour l'identification de champs technologiques qui auront une exploitation économique à plus ou moins long terme.

L'analyse des technologies émergentes actuelles montre qu'elles sont de plus en plus liées à de récents résultats en Physique, le potentiel de contribution de ce domaine scientifique étant considérable.

Les travaux que nous présentons s'inscrivent dans le cadre d'un projet européen, «PROM-TECH», dont les objectifs sont :

- l'identification des technologies émergeantes et prometteuses ayant des liens forts avec le domaine de la Physique ;
- l'identification des communautés scientifiques développant ces technologies ;
- l'aide à la sélection de thématiques qui seront soutenues dans les programmes de R&D du 7ème Programme-Cadre de la Commission Européenne.

ROCHE I., FRANCOIS C., BESAGNI D., (2007) Détection de techniques prometteuses à partir de méthodes bibliométriques. CIDE, 10ème Colloque International sur le Document Electronique, Nancy 4-7 juillet 2007 pp 193-197.

Les différentes méthodes utilisées pour détecter ces nouvelles tendances, sont en particulier les méthodes Delphi basées sur la confrontation d'avis d'experts et les méthodes de scénarios qui visent à balayer les différents futurs possibles [1]. Une approche complémentaire est l'étude des bases de publications scientifiques et de brevets par analyse bibliométrique : l'application des méthodes d'analyse de données textuelles permet de détecter les grandes tendances thématiques [2]. Dans ce domaine, les méthodologies permettant la détection de nouvelles thématiques et l'analyse de leurs évolutions se multiplient [3]. Le programme de recherche TDT « Topic Detection & Traking » lancé par la DARPA en 1996 a induit le développement de technologies de détection d'événements [4]. Ces dernières se concentrent sur l'efficacité informatique pour traiter des flux de dizaines ou centaines de milliers de dépêches d'agences, ou d'enregistrements issus d'entrepôts de données ou d'Internet. Pour permettre une analyse des évolutions détectées, des techniques de visualisation de l'évolution ont été développées : l'outil ThemeRiver utilise la métaphore de la rivière pour montrer les fluctuations des fréquences de termes dans le temps [5] ; l'outil TCRIP (Temporal Graph dRawing with Intelligent placement) [6] montre les évolutions de catégories de documents par une superposition de réseaux de catégories calculés à intervalles réguliers. De même, l'outil VisuaGraph [7] permet de suivre l'évolution d'un graphe de données relationnelles (co-signatures, co-dépôts de brevets, co-occurrences). Cependant, la lisibilité des résultats reste difficile dès que l'on essaie de visualiser des évolutions de structure sur les volumes importants de données.

Pour notre projet, nous avons choisi de travailler à partir des notices d'une base de données bibliographique multidisplinaire et de nous appuyer sur son plan de classement pour analyser les dynamiques des différentes thématiques recherchées. Cette approche est combinée avec l'analyse réalisée par des experts qui nous apportent leurs visions prospectives de leurs domaines de compétence. Dans cet article, nous présenterons les critères qui nous ont guidés pour le choix de la base de données, puis nous décrirons notre méthodologie qui combine les approches bibliométriques fondées sur la notion de plan de classement et les approches par experts pour sélectionner les thématiques prometteuses par étapes successives partant d'une centaine de thématiques, puis limitant à une quarantaine et, enfin aux 10 technologies prometteuses. Nous conclurons par une discussion sur la méthodologie employée.

1 Choix des données

A première vue, les brevets semblent les documents les plus appropriés pour analyser les technologies émergentes. Cependant, les bases de brevets n'ont pas de classification adéquate pour faire le lien entre les technologies décrites dans les brevets et les connaissances scientifiques ayant concouru à leur développement. Au contraire, les publications scientifiques offrent un important potentiel pour repérer les avancées technologiques très en amont, en particulier dans le domaine des technologies basées sur les connaissances. En effet, il est commun de trouver, associés dans la même publication, des résultats innovants et leurs fondements scientifiques. Le choix des membres du projet s'est donc porté sur les bases de données de publications scientifiques.

Une premier problème est la pauvreté des schémas de classement des publications dans la plupart de ces bases de données, ce qui implique des analyses portant sur des hauts niveaux d'agrégation et ne permet pas d'identifier le développement de champs spécifiques. De plus, la plupart des bases sont spécialisées sur un domaine (Sciences de la Vie, Physique, Chimie…) ce qui interdit d'identifier des liens vers les technologies appliquées à des domaines variés.

Nous avons choisi la base PASCAL de l'INIST, base multidisciplinaire d'environ 15 millions de notices bibliographiques, pour les caractéristiques suivantes :

- sa multidisciplinarité permet d'accéder à la fois aux domaines de la Physique et aux domaines d'applications technologiques ;
- la finesse de son plan de classement offre la possibilité d'analyser des domaines très spécialisés. Par exemple, il est plus fin que le «Category codes» du Science Citation Index ;
- son mode de constitution permet d'établir via les codes de classement multiples, des passerelles entre la Physique et ses applications technologiques. En effet, chaque notice bibliographique est affectée d'un ou plusieurs codes de façon individuelle.

Après analyse du plan de classement de la base, nous avons sélectionné des notices bibliographiques ayant à la fois un code en Physique, quelque soit la spécialité, et un code correspondant à un domaine d'application technologique. Les deux domaines applicatifs sélectionnés sont : les Sciences de l'Ingénieur et les Sciences de la Vie.

2 Méthodologie

2.1 Analyse des facteurs d'irrégularité

Afin de pouvoir interpréter les évolutions, il est important de s'assurer que la dynamique des corpus étudiés est bien due à une évolution thématique indépendante de tout événement lié au mode de constitution de la base de données ou au mode de diffusion de l'information (ex : congrès pluri-annuels).

Nous avons analysé différents évènements touchant à la production de la base Pascal :

- début et fin de deux importantes collaborations d'achat de notices bibliographiques dans les domaines de la Physique et des Sciences de l'Ingénieur. Ceci nous a conduits à choisir la période entre 1994 et 2003 où ces collaborations étaient stables ;
- à partir de 2004, la procédure de constitution des notices de la base Pascal a été en partie automatisée, ce qui conduisit à affecter des codes plus génériques à l'ensemble d'une revue et non plus notice par notice comme c'était le cas auparavant. La sélection des notices à partir de l'année de production 2004 est donc rendue plus difficile.

Nous avons conservé uniquement les articles publiés directement dans les revues, en négligeant les articles initialement publiés dans des actes de congrès et ensuite repris dans une revue. En effet, d'une part, le rythme de parution des conférences est par nature irrégulier ou, obéit à une fréquence pluriannuelle et, d'autre part, les thématiques des articles publiés varient d'une conférence à l'autre. Ce type de publication induit donc des variations de volumétrie de la production scientifique qui influencent les taux de croissance de façon irrégulière et rendent difficile l'explication des courbes d'évolution de la production scientifique.

2.2 Croisement des catégories du plan de classement, et sélection du premier ensemble de thématiques

Nous avons exploré les associations entre la Physique, considérée dans sa globalité (code 001B de l'arborescence du plan de classement), et les domaines applicatifs. Pour chacun de ces derniers, nous avons défini un niveau dans l'arborescence du plan de classement afin de déterminer la liste des différentes thématiques appliquées devant être examinées. Pour les Sciences de l'Ingénieur (code 001D), nous avons découpé l'arborescence au niveau « 7 caractères » et nous obtenons par exemple les catégories suivantes :

- 001D03F : électronique des semi-conducteurs
- 001D11C : transformation de matériaux métalliques

Pour les Sciences de la Vie (code 002), nous avons découpé l'arborescence au niveau « 6 caractères » et nous obtenons par exemple les catégories suivantes :

- 002A04 : biologie moléculaire et cellulaire
- 002B26 : sciences médicales - chirurgie.

L'ensemble des croisements possibles correspond à 201 catégories (133 en Sciences de l'Ingénieur et 68 en Sciences de la Vie).

Pour chaque catégorie nous avons étudié son évolution globale annuelle et nous avons sélectionné celles ayant un volume minimum de 10 références chaque année soit 112 catégories ou thématiques.

2.3 Sélection des thématiques candidates

Cette sélection est réalisée essentiellement selon des critères statistiques avec une validation finale de la part des experts du Fraunhofer ISI [8].

La première sélection est basée sur le nombre de références existant la dernière année (2003) , en l'occurrence, 30 minimum. En effet, un nombre trop faible de références ne permet pas de faire un traitement statistique. Les catégories ayant plus de 500 notices en 2003 sont éclatées en descendant à un niveau plus fin dans la hiérarchie du plan de classement du point de vue applicatif afin de distinguer plus finement les différentes technologies ou domaines scientifiques applicatifs. Les 108 catégories restantes ont été analysées par les méthodes statistiques décrites ci-dessous.

Trois indicateurs de croissance sont utilisés pour étudier les évolutions des catégories :

- l'indice de croissance qui est une relation simple entre le nombre de publications dans la période la plus récente (2003) et une année de base considérée comme la référence; cet indice est un indicateur de l'évolution globale au cours d'une période donnée ;

- le taux moyen de croissance annuelle qui permet de tenir compte des variations annuelles dans la croissance d'une thématique et complète ainsi l'indice de croissance qui ne tient compte que des bornes de l'intervalle ;
- le « Sharp ratio » complète le taux moyen de croissance annuelle en le pondérant par la variabilité des croissances annuelles.

Chacun de ces indicateurs apporte une information complémentaire. Aussi nous avons utilisé l'ensemble pour sélectionner les thématiques candidates. La sélection est donc effectuée sur la base de la combinaison de ces indicateurs, mais également sur l'observation des courbes de croissance. Par exemple, pour une classe qui apparait uniquement en 2001, nous ne pouvons pas calculer ces indicateurs alors que la courbe de croissance est très prometteuse. Et enfin, un dernier critère est l'observation du contenu de la classe.

Cette étape a permis de sélectionner 43 thématiques candidates, que nous avons soumises à un panel d'experts par le biais d'une enquête.

2.4 Sélection et analyse des 10 thématiques prometteuses

A partir des corpus associés à ces 43 thématiques, nous avons repéré les principaux auteurs européens afin de définir un panel d'experts que nous avons consulté et à l'intention desquels nous avons préparé un questionnaire. Nous avons donc fourni à chaque expert un descriptif de chaque thématique le concernant accompagné d'un ensemble de questions lui permettant de nous faire part de son estimation des potentialités de la thématique.

Par exemple, pour chaque thématique, nos questions portaient sur :

- les développements technologiques les plus récents
- les développements les plus importants prévisibles dans les dix prochaines années
- l'importance économique de ces applications
- le délai de mise à disposition de ces premières applications

Une analyse des réponses a permis de dégager 10 thématiques prometteuses [9] présentées dans la figure 1.

Thématique	nombre de références	
	1996-1999	2000-2003
Medical imaging	723	1725
Magneto-electronics	28	253
Molecular biology	1680	2603
Light emitting diodes	1797	2074
Air pollution measurement	173	357
Molecular electronics	53	915
Semiconductors	5241	6112
Surface treatment of metals	93	298
Simulation in mechanical engineering	764	1502
Using neural networks	334	631

Figure 1. liste des 10 thématiques sélectionnées

Une analyse bibliométrique plus poussée permet de décrire précisément les thématiques sélectionnées. Notre approche consiste à réaliser des classifications thématiques sur deux périodes, 1996-1999 et 2000-2003, et d'analyser l'évolution du vocabulaire décrivant les classes. Une matrice de comparaison des mots-clés associés aux classes de chaque période est construite en se basant sur le pourcentage des mots-clés d'une classe de la 2nde période qui préexiste dans une classe de la première période. A partir de cette matrice, nous pouvons repérer différents comportements de classes (stabilité, fusion ou éclatement, changement de statut dans le réseau global des classes) qu'il est ensuite nécessaire de faire valider par les ingénieurs documentalistes spécialistes du domaine technologique. Cette analyse servira de base de discussion avec les experts lors de la conférence qui clôturera le projet.

3 Discussion

Pour ces travaux, dont les résultats finaux sont encore à venir, nous avons utilisé conjointement des techniques bibliométriques et des analyses réalisées par des experts. Cette approche nous permet de réduire la tâche d'expertise en calculant des propositions sur lesquels les experts se prononcent. Cette expertise est nécessaire pour valider et compléter les résultats que nous pouvons obtenir avec des méthodes bibliométriques.

Notre démarche bibliométrique est basée sur l'existence d'un plan de classement multidisciplinaire très fin et sur l'indexation des notices bibliographiques par des codes multiples. Cette contrainte est très forte, et n'est respectée que partiellement. En effet, les méthodes d'indexation sont différentes d'un domaine à l'autre et encore

différentes entre l'INIST et ses coopérants. La connaissance précise de ces pratiques est donc nécessaire pour réaliser l'interprétation des résultats bibliométriques.

De même, les procédures automatiques d'indexation utilisées actuellement pour produire la base de données ne permettent plus de respecter cette contrainte. Afin de reproduire ou généraliser cette approche basée sur un plan de classement, il sera nécessaire d'envisager une méthode de catégorisation automatique des notices dans ce plan et de prévoir un protocole de mise à jour de ce dernier. La difficulté majeure sera la prise en compte du nombre de niveaux et de la finesse de la hiérarchie.

Une approche complémentaire de classification automatique est également à envisager, et plus particulièrement sous une forme incrémentale qui permet de suivre les évolutions des classes en fonction de la date de publication des références. Ce type d'algorithme est en cours d'élaboration [10] et des expérimentations portant sur un domaine scientifique unique, bien délimité ont déjà été réalisées Cependant, pour analyser une base de données dont la couverture est multi-domaines, nous devrons soit considérer une approche de classification à plusieurs niveaux pour définir des classes de plus en plus spécifiques, soit combiner une approche classification avec une méthode de catégorisation automatique.

Remerciements

Ces travaux sont réalisés dans le cadre du projet européen « PROM-TECH » (PROMising TECHnologies). Ce projet s'inscrit dans le cadre de l'Action Spécifique NEST (New and Emerging Science and Technology) du 6ème Programme-Cadre de l'Union Européenne. Le consortium est constitué par l'ARC System research GmbH (Vienne, Autriche), le Fraunhofer Institut für Systemtechnik und Innovationsforschung (Karlsruhe, Allemagne) et l'INIST-CNRS.

Nous remercions également nos collègues ingénieurs documentalistes de l'INIST-CNRS qui ont participé activement aux différentes étapes du projet en nous apportant leur expertise à la fois scientifique et documentaire.

Bibliographie

[1] R. Monti et F. Roubelat, La boîte à outils de prospective stratégique et la prospective de défense : rétrospective et perspectives, *Actes des Entretiens Science & Défense*, Paris 1998

[2] M. Rajman, V. Peristera, J.C. Chappelier, F. Seydoux, A. Spinakis, Evaluationof Scientific and Technological Innovation using Statistical Analysis of patents, *6ème Journées internationales d'Analyse statistique de Données Textuelles, JADT 2002*

[3] E. Noyons, Science maps within a science policy context. In *Handbook of quantitative Science and Technology Research*, Eds. Moed H.F., Glänzel W., Schmoch U., Kluwer Academic Publishers, London, 2004, pp 237-255

[4] H. Binsztok et P. Gallinari, Un algorithme en ligne pour la détection de nouveauté dans un flux de documents, Dans : *JADT 2002 : 6° Journées internationales d'Analyse statistique des Données Textuelles*

[5] S. Havre, E. Hetzler, P. Whitney, L. Nowell, ThemeRiver: visualizing thematic changes in large document collections, *IEEE transactions on visualization and computer graphics*, 2002, Vol. 8, N°1

[6] C. Erten, P.J. Harding, S.G. Kobourov, K. Wampler, G. Yee, Exploring the computing literature using temporal graph visualization, *Report, Department of Computer Science*, University of Arizona. 2003

[7] E. Loubier, W. Bahsoun, B. Dousset, Visualisation de l'évolution des informations relationnelles par morphing de graphe. Dans : *5° Aterlier Visualisation et Extraction de Connaissances (EGC 2007 Namur, Belgique, 23/01/2007-26/01/2007)*, http://www.info.fundp.ac.be/egc2007/actes.php, p. 43-54, janvier 2007

[8] R. Frietsch, Results of statistical evaluation, *Deliverable 03 for Project PROM-TECH (contract N° 15615)*, 12 pages, 2007

[9] J. v. Oertzen, Results of evaluation and screening of 40 technologies, *Deliverable 04 for Project PROM-TECH (contract N° 15615)*, 32 pages + appendix, 2007

[10] A. Lelu, P. Cuxac, J. Johansson, Classification dynamique d'un flux documentaire : une évaluation statique préalable de l'algorithme GERMEN, Dans : *JADT 2006*, Besançon, 19-21 Avril 2006