La cartographie des brevets dans l'industrie et la recherche : outils et pratiques

Antoine Blanchard

La cartographie des brevets dans l'industrie et la recherche : outils et pratiques

Antoine Blanchard*

*Deuxième labo, 56 bd Auguste Blanqui, 75013 Paris
antoine.blanchard@gmail.com
http://www.deuxieme-labo.fr

Résumé. L'information brevet est souvent perçue comme une mine d'or pour l'industrie et comme un puits sans fond pour la recherche académique. La demande de l'industrie pour des outils lui permettant d'exploiter cette mine d'or en ont fait un utilisateur routinier de nombreuses technologies de cartographie de l'information, de fouille de texte ou de données, quand les chercheurs se contentent souvent d'une information brevet géographiquement limitée et des méthodes d'analyse qui leur sont familières. Malgré leurs différences d'approches et d'objectifs, les deux domaines peuvent apprendre l'un de l'autre. Dans cette communication, nous tenterons de rendre compte, de manière aussi exhaustive que possible, des pratiques qui ont cours dans l'industrie ou dans la recherche et de montrer comment certains dispositifs de navigation entre niveaux méso- et macroscopique pourraient les rapprocher.

1 L'information brevet et ses différents niveaux d'analyse

L'information brevet est souvent la plus redoutée des non-spécialistes, et même des spécialistes, car elle présente plusieurs difficultés :

- elle est disparate : chaque pays possède son propre système de brevets et donc ses propres brevets (les brevets relatifs à une même invention déposés dans différents pays formant ce que l'on nomme une famille de brevets, *cf.* figure 1), publiés dans sa propre langue et son propre format (base de données électronique ou non, résumés ou plein texte etc.) ;
- elle est protéiforme : des figures détaillées côtoient un vocabulaire technico-juridique ou «*patentese*» dont le but est souvent d'être le plus englobant et le moins explicite possible ;
- elle est riche de méta-données : chaque brevet va de pair avec un inventeur et un déposant mais aussi une ou plusieurs classes de la Classification internationale des brevets (CIB), un statut juridique[1], une place dans l'histoire de sa famille[2]…

À l'inverse, les publications scientifiques sont un matériau d'acccès plus facile et immédiat et donc plus couramment utilisé dans les recherches sur la sociologie et l'histoire des sciences et techniques. Pourtant, l'information brevet est irremplaçable pour comprendre notamment les processus d'innovation. Les entreprises de R&D la manient tous les jours pour des raisons évidentes et se sont constitué une expertise qui peut être valorisée dans d'autres contextes. Cette expertise relève de deux approches distinctes :

- chercher et utiliser l'information brevet avec l'œil expert de l'ingénieur brevets ou du scientifique qui s'intéresse à sa portée juridique ou son contenu technique (analyse microscopique) ;
- chercher et utiliser l'information brevet selon l'approche plus holistique du veilleur qui s'intéresse aux motifs d'ensemble, aux signaux faibles etc. (analyse macroscopique). Anthony J. Trippe (2002) a proposé de nommer «*patinformatics*» cette deuxième approche, sur le modèle de disciplines ayant récemment bouleversé la méthode scientifique par leur utilisation de données massives comme la bio-informatique ou la chemo-informatique.

Les outils de la *patinformatics*, dont la cartographie des brevets, sont venus soutenir en routine l'approche macroscopique du veilleur ou du consultant en innovation mais sont aussi déployés aujourd'hui pour l'analyse mésoscopique par les sociologues ou chercheurs en économie, gestion et propriété intellectuelle.

[1] Ainsi, un brevet dont la redevance cesse d'être payée perd son effet légal au bout de plusieurs mois.

[2] On distingue en particulier entre la demande de brevet, qui n'a aucun effet légal, et le brevet proprement dit qui est délivré après examen et peut différer substantiellement de la demande de brevet. Les États-Unis sont aussi fameux pour leur système complexe qui donne naissance à des familles de brevets arborescentes plutôt que linéaires.

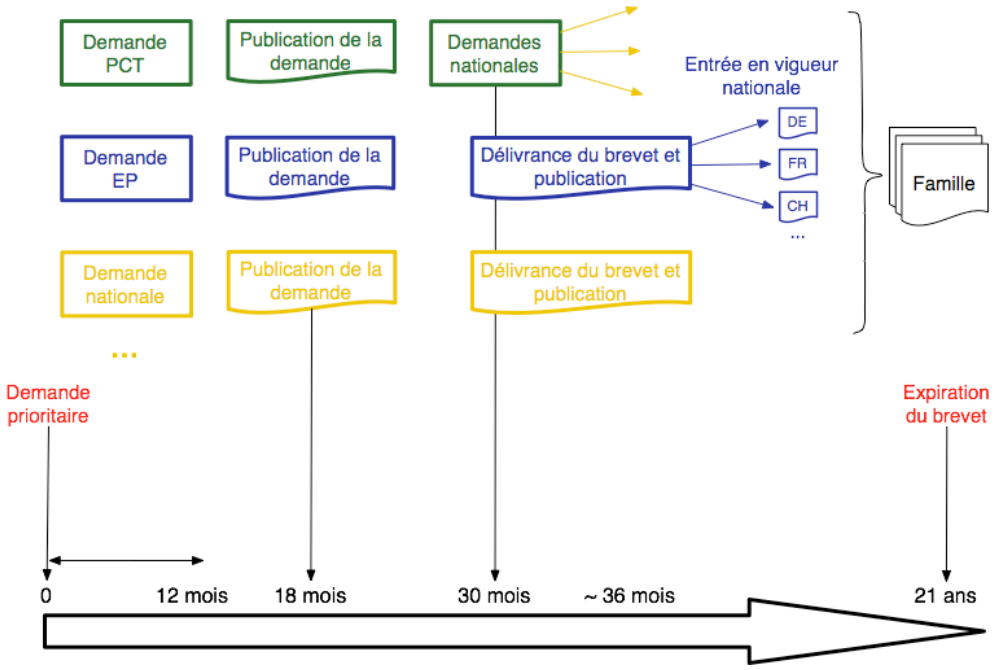

FIG. 1 – *Cheminement typique d'une demande de brevet résultant en une famille de brevets, depuis le dépôt au niveau national ou international («EP» pour le système européen, «PCT» pour le système de l'Organisation mondiale de la propriété intellectuelle…) dans les 12 mois suivant la demande prioritaire jusqu'à la délivrance d'une famille de brevets et leur expiration.*

2 Différents outils et pratiques de cartographie des brevets

La cartographie des brevets est définie par l'Office européen des brevets comme la représentation visuelle du résultat d'analyses statistiques ou de fouille de texte appliquées aux brevets, permettant de comprendre et évaluer facilement de larges volumes d'information brevet[3]. Pour Anthony J. Trippe (2002), cela revient essentiellement à représenter graphiquement ou physiquement dans un espace-plan l'art relevant d'un domaine particulier. Techniquement, la cartographie regroupe l'ensemble des représentations graphiques de données groupées, allant notamment de l'analyse dimensionnelle comme l'analyse en composantes principales à la classification non-supervisée en passant, dans une moindre mesure, par la construction d'arbres ou dendrogrammes (Porter et Cunningham, 2005, chapitre 10.3).

Parmi ses traits distinctifs, on trouve en particulier la présence de regroupements appelés «*clusters*», qu'ils soient des partitions définies rigoureusement ou des représentations graphiques lâches, séparés par une certaine distance ou des arêtes. Un *cluster* peut être obtenu de façons très diverses et ne se limite pas ici à la classification non-supervisée comme le voudrait sa définition stricte.

La cartographie des brevets recouvre donc une palette d'outils et de techniques, complémentaires, qui peuvent être utilisés aussi bien par l'industrie que par la recherche. Nous allons les détailler à présent, en gardant à l'esprit que chaque technique peut être déclinée selon une typologie transversale, celle de la dimension temporelle :

- la cartographie statique donne une image figée à un certain point dans le temps, qui est généralement le moment de l'analyse mais représente souvent le passé en raison des délais propres à l'information brevet (18 mois entre le dépôt de la demande prioritaire et la première publication, plus le temps d'enregistrement et d'indexation dans les bases de données) ;
- la cartographie dynamique peut être utilisée pour retracer le développement au cours du temps d'une technologie. Elle consiste souvent en une suite de cartes statiques, qui découpent la période temporelle étudiée.

2.1 Les cartes sémantiques ou «*topic maps*»

L'approche des cartes sémantiques consiste à laisser parler le corpus par lui-même en se reposant uniquement sur son contenu lexical, plutôt que sur des catégories définies *a priori* (classifications des bases de données bibliographiques,

[3] http://www.epo.org/patents/patent-information/business/stats/faq.html#mapping

mots-clés attribués par les auteurs) ou *a posteriori* (jugement des experts). Cette approche permet de se débarasser des idées préconçues et donc de mieux rendre compte du front de recherche, en perpétuel renouvellement (Mogoutov et al., 2008). Elle permet aussi d'ajuster le degré de détail au niveau voulu, comme nous le montrerons dans la partie 3.

Méthodologie. Il faut en principe deux étapes pour obtenir une carte sémantique. La première découpe le corpus étudié en *n* entités (termes et expressions) bien délimitées tandis que la seconde représente le résultat obtenu sous la forme de carte, typiquement en projetant en deux dimensions l'espace de vecteurs (représentant chacun un document) à *n* dimensions. Ainsi, plus le contenu lexical des documents est proche, plus la distance qui les sépare est faible (les deux axes de la carte n'ont aucune signification en soi), la densité des documents pouvant être représentée en sus par des « montagnes » (Blanchard, 2006). Certains logiciels « clés en mains » combinent les deux étapes, qui deviennent indiscernables pour l'utilisateur.

Trousse à outils. Deux familles de méthodes permettent la première étape de découpage du corpus, même si les frontières tendent à se brouiller avec le développement de logiciels empruntant le meilleur des deux mondes :

- la fouille de texte ou « *text mining* » utilisant exclusivement une approche statistique (fréquences d'occurrence, de co-occurrence etc.), qui réussit d'autant mieux que le corpus de documents est volumineux. C'est la méthode par défaut des logiciels de cartographie clés en mains, notamment ceux qui sont spécialisés dans l'information brevet comme STN AnaVist[4] ou Aureka[5] ;

- la fouille de texte à base de traitement automatique des langues ou « *natural language processing* » (NLP) qui utilise l'analyse syntaxique des phrases, caractérise la relation entre différents termes, résout les ambiguïtés du langage et peut utiliser un dictionnaire spécialisé. Cette méthode est l'apanage de logiciels souvent professionnels, comme SPSS LexiQuest Mine[6] utilisé par Mogoutov et al. (2008), auxquels il faut adjoindre un logiciel de représentation des données (ReseauLu[7] en l'occurrence).

Pratique des cartes sémantiques. Les outils clés en main fleurissent pour la cartographie sémantique des brevets, et sont utilisés avidemment par l'industrie (Eldridge, 2006). Ils sont faciles à prendre en main, faciles à interpréter pour la prise de décision et permettent de se passer de la CIB, employée par l'ensemble des examinateurs des offices des brevets à travers le monde, qui manque souvent de pertinence pour ses besoins (Fattori et al., 2003) et de clarté pour ses utilisateurs finaux. Mais la recherche s'en est également emparée à quelques occasions, comme dans cette étude des brevets dans le domaine des biotechnologies pour l'agriculture (Graff et al., 2003). Le logiciel utilisé, Aureka, permet notamment de créer automatiquement des tranches temporelles dans les données et de comparer l'activité technologique entre plusieurs périodes, comme illustré par Trippe (2001).

Informations pouvant enrichir une carte sémantique. De nombreuses informations peuvent être ajoutées à une carte sémantique en utilisant des couleurs ou la forme des points, par exemple l'information sur le déposant du brevet ou le fait qu'il appartienne au corpus citant ou cité, pour lui faire répondre à des questions encore plus larges. Chez Syngenta Crop Protection, entreprise du secteur agrochimique, nous avons utilisé cette approche pour repérer où se concentrent les citations par rapport au portefeuille de brevets et caractériser l'activité des concurrents ou de l'industrie pharmaceutique autour de certaines classes de molécules brevetées par Syngenta.

2.2 Les cartes macro-thématiques

Ces cartes mobilisent ce que Zitt et Bassecoulard (2004) qualifient de dimension macro-thématique de l'information scientifique, c'est-à-dire aussi bien l'indexation manuelle offerte par les bases de données à valeur ajoutée que les diverses classifications appliquées aux documents. La CIB a ainsi été utilisée par Meyer (2007) dans une étude de cas sur les nanotechnologies et les codes de la base de données Derwent[8] par Trippe (2001) dans l'étude du portefeuille de l'entreprise pharmaceutique Vertex.

[4] http://www.stn-international.de/stninterfaces/stnanavist/stn_anavist.html
[5] http://www.thomsonreuters.com/products_services/scientific/Aureka
[6] http://www.spss.com/fr/lexiquest/lexiquest_mine.htm
[7] http://www.aguidel.com/fr/?sid=6
[8] http://www.stn-international.de/stndatabases/databases/wpidswpx.html

Carte de co-occurrences. Meyer (2007) identifie 5407 sous-classes de la CIB présentes dans son corpus puis range par deux les sous-classes qui apparaissent ensemble sur un ou plusieurs brevets, avec leur fréquence de co-occurrence. Il utilise le logiciel de statistique SPSS[9] pour représenter cette matrice sur une carte, par une méthode de «*Multi-Dimensional Scaling*» (MDS) présente dans la routine SYSTAT de l'algorithme ALSCAL. Chaque sous-classe de la CIB apparaît d'autant plus grosse qu'elle est fréquente dans le corpus et proche des autres sous-classes avec lesquelles elle co-occurre souvent. On visualise ainsi des *clusters* de technologies composant le champ des nanotechnologies, où les dispositifs de mesure se distinguent des nano-matériaux, du monde de la chimie ou pharmacie et des nano-couches ou semi-conducteurs.

Regroupement ou «clustering» des brevets. Trippe (2001) utilise le logiciel de fouille de données TWID Expert[10] pour regrouper les brevets et demandes de brevets qui sont codés de façon similaire par les indexeurs Derwent. Pour ce faire, le motif d'ensemble des codes de classification est pris en compte («*many to many*») et pas uniquement leurs co-occurrences deux à deux. Les 128 documents ainsi traités se retrouvent regroupés dans 25 *clusters* (nombre fixé par l'analyste par essai et erreur), avant d'être rangés et étiquetés manuellement.

2.3 Les réseaux de citation

À l'inverse des méthodes présentées précédemment, les réseaux de citation reposent sur des liens directs et directionnels entre documents, à savoir les liens de citation. La cartographie consiste principalement à les représenter graphiquement et à en tirer des informations complémentaires par le jeu de la transformation et de l'analyse, comme cela se fait couramment dans l'analyse des articles scientifiques (Small, 1999).

Représentation des liens de citation. Les citations entre brevets sont un indicateur de la proximité technologique entre différentes inventions, dans certaines limites soulignées par Oppenheim (2000)[11]. En agrégeant les données par déposant et les mobilisant dans des réseaux de citations à grande échelle, on peut mettre en évidence certaines relations entre les acteurs — les plus liés entre eux étant les plus susceptibles d'entrer en compétition. En agrégeant les brevets en *clusters* ayant un profil de citation similaire, on peut aussi représenter les flux de connaissances («*knowledge flows*») à un niveau mésoscopique. C'est ainsi que Igami (2008) fait apparaître le rôle de pivot des capteurs et actionneurs pour les nanotechnologies alors que des technologies comme les cristaux photoniques ou la nanolithographie sont beaucoup plus isolées.

Le réseau peut être calculé sur la base des brevets individuels plutôt que sur celle des déposants, à condition de fixer un seuil ou une limite temporelle pour diminuer le nombre d'observations et conserver un résultat lisible. On peut ainsi être amené à s'intéresser à un brevet en particulier, comme celui de l'entreprise japonaise Rohm dont Sternitzke et al. (2008) constatent qu'il cite pas moins de 229 familles de brevets appartenant en partie au concurrent Nichia, et ce en raison de ses revendications particulièrement larges. On retombe ici dans une analyse plus microscopique et Aureka offre cette possibilité en routine, très appréciée en entreprise.

Abstraction des liens de citation. Une autre approche consiste à utiliser les liens de citation pour calculer un facteur d'attraction ou de répulsion entre chaque brevet. En utilisant un algorithme développé pour la représentation de molécules en chemo-informatique, Masatsura Igami (2008) obtient ainsi des *clusters* dont ont disparu les liens de citation à proprement parler, remplacés par une distance plus ou moins grande entre *clusters* et en leur sein. Cette étude de cas sur les nanotechnologies lui permet par exemple d'observer quinze domaines constitutifs, contenant de 43 à 550 demandes de brevets. Une série temporelle lui permet également de mettre en évidence l'évolution de ces domaines technologiques entre 1990 et 2000, notamment dans leurs interactions.

Pour aller plus loin. Les citations allant toujours du plus récent au plus ancien, elles permettent aussi de retracer la trajectoire d'un domaine et l'évolution des centres d'intérêt des inventeurs (en corrélant par exemple la taille des points du réseau à l'âge des brevets, comme chez Sternitzke et al. (2008), ou en indiquant le sens de la citation). Pour identifier selon ce principe le progrès des technologies d'actionnement variable de soupape dans l'industrie automobile, von Wartburg et al. (2005) ont constitué un corpus selon une méthode itérative qui s'appuie sur la chaîne des citations de brevets européens. Le résultat montre par exemple que l'ajout d'un arbre à cames par l'industrie est venu compléter aussi bien les solutions hydrauliques que mécaniques, lesquelles forment deux *clusters* bien distincts. Dans un second temps, les auteurs vont encore plus loin en tenant également compte du couplage bibliographique (le fait pour deux brevets de

[9] http://www.spss.com/fr/statistics/

[10] http://www.synthema.it/textmining

[11] Et il est exagéré d'en faire autre chose, comme Sun et Morris (2008) qui y voient la marque de l'utilisation d'une technologie par une autre.

citer un même document) et des co-citations (le fait pour deux brevets d'être cités par un même document), pour obtenir une carte de proximité qui confirme les observations précédentes.

Le couplage bibliographique est utilisé également par Sun et Morris (2008) pour regrouper les brevets qui ont des motifs de citation similaires, formant 10 *clusters* relatifs à la technologie des têtes de lecture optiques. Ils décomposent alors chaque *cluster* selon un axe temporel pour montrer l'évolution des fronts de recherche, dont on voit qu'ils se relaient les uns les autres. En complétant la représentation graphique avec des donnés de citation (chaque point apparaît d'autant plus gros que le brevet a été beaucoup cité et d'autant plus foncé qu'il est encore cité aujourd'hui), ils montrent en même temps lesquelles de ces lignes de recherche ont été les plus fécondes.

Il ne faut pas oublier non plus les liens d'auto-citation qui permettent de visualiser l'importance de l'innovation endogène face au transfert de technologies extérieures (Jaffe, 1998) ou l'enchevêtrement des brevets d'une entreprise se citant abondamment les uns les autres (qui forment des «*patent thickets*» défavorables aux compétiteurs qui voudraient entrer dans le domaine).

2.4 Les cartes de réseaux sociaux

Les brevets peuvent aussi être utilisés non pas pour eux-mêmes, mais pour faire parler les acteurs qui se cachent derrière, individus ou institutions. C'est ce que viennent de réaliser Sternitzke et al. (2008) dans une étude de cas portant sur les diodes électroluminescentes et les diodes lasers.

Trousse à outils. Certains logiciels tournés vers l'information brevet proposent l'analyse clés en mains des réseaux de collaboration entre inventeurs ou déposants, à l'instar de Matheo Patent[12] (Dou, 2004), de Vantage Point[13] ou de Thomson Data Analyzer[14]. Mais l'utilisation de logiciels d'analyse et de cartographie de réseaux sociaux comme ReseauLu, Pajek[15] ou UCINET[16] offre une plus grande souplesse à l'analyste, à condition qu'il prenne la peine de reformater l'information issue des bases de donnés de brevets dans des formats compatibles, à l'aide par exemple de PATONanalist[17] (Sternitzke et al., 2007).

Préparation et représentation des données. Les données sont susceptibles d'être faussées par la présence d'homonymes ou synonymes et doivent de préférence être nettoyées. Une solution proposée par Sternitzke et al. (2008) consiste à croiser les données avec la base INPADOCDB[18] de l'Office européen des brevets, laquelle contient le nom complet des inventeurs et pas seulement l'initiale de leur prénom. Selon l'analyse recherchée, les nœuds du réseau seront les brevets, les inventeurs ou les déposants, dont la loi de Lotka (1926) prédit que l'immense majorité ont en fait une contribution anecdotique. Il est donc conseillé de se limiter aux inventeurs ou déposants les plus actifs, par exemple ceux du dernier décile, ce qui permet en même temps de rendre le résultat plus lisible.

Réseaux de coopération. En première approximation, la coopération entre inventeurs ou déposants revient à la co-signature ou le co-dépôt de brevets. L'utilisation de cette information contenue dans les champs bibliographiques des brevets permet de tracer les réseaux de coopération et mettre en évidence des sous-groupes fortement connectés («collèges invisibles»), des acteurs qui sont centraux et ceux qui font office de passeurs entre différents sous-groupes. En utilisant la CIB qui décrit l'appartenance de chaque brevet à un ou plusieurs domaines technologiques, Sternitzke et al. (2008) montrent notamment que ces passeurs ont une activité significativement plus variée, car ils ont accès au savoir de plusieurs collèges invisibles.

2.5 Les cartes géographiques

L'information brevet contient deux niveaux d'information d'intérêt géographique : le pays d'appartenance du déposant (ou inventeur) et le pays de dépôt du brevet. Igami (2008) a combiné le premier niveau avec une carte de densité pour comparer la spécialisation technologique de la Suisse et de la Corée, représentant chacune un sous-ensemble du corpus sur les nanotechnologies. Le deuxième niveau censé montrer quels pays sont incontournables pour un domaine donné apporte en fait une information moindre puisqu'on retrouve généralement les États-Unis, le Japon et les principaux pays européens en tête de liste.

[12] http://www.matheo-patent.com
[13] http://www.thevantagepoint.com
[14] http://www.thomsonreuters.com/products_services/scientific/Thomson_Data_Analyzer
[15] http://pajek.imfm.si
[16] http://www.analytictech.com/ucinet/ucinet.htm
[17] http://www.paton.de
[18] http://www.stn-international.de/stndatabases/databases/inpadocdb.html

3 Naviguer entre niveaux méso et macro

Nous avons indiqué précédemment que l'industrie se caractérise par son besoin à la fois d'information brevet microscopique guidant les actions légales (dépôt, opposition, plainte pour infraction…) et d'analyse macroscopique orientant l'innovation, tandis que la recherche s'intéresse surtout à l'information de niveau mésoscopique. Concrètement, cela signifie que l'industrie a les moyens de cartographier régulièrement le portefeuille complet d'une ou plusieurs multinationales ou d'un large domaine technologique en multipliant les sources d'information[19] tandis que la recherche académique s'intéresse souvent à un domaine plus restreint tel que représenté dans une unique juridiction (brevets américains, européens ou français) : elle peut se contenter d'un échantillon de l'information existante quand l'industrie a besoin d'exhaustivité. Alors que l'industrie cherche à agir, la recherche vise surtout à s'informer et les enjeux ne sont évidemment pas les mêmes.

Pourtant, ces deux approches sont loin d'être antinomiques. Nous avons montré comment les mêmes outils peuvent être utilisés par les uns ou les autres pour poser des questions semblables à un corpus plus ou moins large. Mais pour un corpus et une cartographie donnés, la représentation et la navigation peuvent aussi être personnalisées selon les besoins. Notre expérience nous a notamment amené à expérimenter l'effet des listes de mots vides ou «*stopwords*» sur la granularité des cartes sémantiques, tandis que d'autres auteurs ont misé sur le choix du système de pondération ou la reconnaissance des expressions par NLP.

3.1 Personnalisation des mots vides

Les mots vides sont nés dans les bases de données des années 1950, quand les ressources informatiques limitées nécessitaient que tous les mots ne soient pas indexés comme mots clés, particulièrement ces mots grammaticaux qui représentent jusqu'à la moitié des occurrences («le», «sur», «tant»…). Depuis, ils ont quasiment disparu des bases de données de brevets mais prospèrent dans les outils de cartographie sémantique, où ils contribuent grandement à la qualité et la pertinence du résultat final (Blanchard, 2007). Ces outils sont ainsi fournis avec une liste de mots vides, également appelé antidictionnaire, couvrant au moins les besoins de la langue anglaise. Mais ces listes par défaut manquent du vocabulaire propre aux brevets avec des termes juridiques comme «*embodiment*» ou «*comprising*» et des termes lexicalement pauvres comme «*exhibit*» ou «*demonstrate*». Ceux-ci peuvent être ajoutés manuellement dans la plupart des cas.

Mais l'analyste peut aller plus loin en ajoutant aux mots vides des termes spécifiques au domaine étudié qui n'apportent pas d'information dans le contexte. Par exemple, le terme «*protéine*» est superflu dans l'analyse de l'utilisation d'enzyme phytase pour la nutrition animale, dans la mesure où toutes les enzymes sont des protéines. L'expérience montre que cette méthode permet d'obtenir des *clusters* mieux séparés et d'affiner le niveau de l'analyse autour de concepts plus discriminants (Blanchard, 2007; Trippe, 2001), «moins macroscopiques». Elle peut même être automatisée chez certains outils comme OmniViz[20] et STN AnaVist.

3.2 Choix du système de pondération

Le système de pondération ou «*weighting system*» est un paramètre de l'algorithme de fouille de texte qui permet d'obtenir soit des gros *clusters* reposant sur des mots qui reviennent fréquemment, soit des petits *clusters* reposant sur des mots plus rares (Fattori et al., 2003). Un système fréquemment utilisé est l'algorithme «tf $\times$ idf» (pour «term frequency $\times$ inverse document frequency»), qui consiste à faire le produit de deux mesures complémentaires : la fréquence du mot dans le corpus et l'inverse de la proportion de documents qui contiennent ce mot (Salton et Buckley, 1988). Ainsi, le poids d'un mot augmente quand il est utilisé abondamment et il diminue quand il apparaît dans de nombreux documents. La personnalisation du système de pondération est permise par quelques outils comme PackMOLE, développé en interne pour les besoins de l'entreprise Tetra Pak (Fattori et al., 2003). Selon l'expérience de ces auteurs recherchant à obtenir des *clusters* à la fois homogènes et bien séparés, l'option de pondération préférée est celle favorisant des *clusters* spécialisés à base de mots plus rares.

3.3 Reconnaissance des expressions

Mogoutov et al. (2008) indiquent que le NLP permet de gagner un niveau de pertinence dans l'analyse par rapport à la fouille de texte statistique, en facilitant le travail préalable à la cartographie par la reconnaissance des expressions. Dans leur exemple, plutôt que de s'embarrasser d'un mot comme «cancer» qui est tellement courant dans leur corpus sur les puces à ADN qu'il en devient trivial, l'algorithme isole des expressions assez fréquentes comme «cancer du sein», «cancer du côlon» ou «épidémiologie du cancer». Ainsi, au lieu d'attendre que la cartographie fasse ressortir des *clusters*

[19] C'est ainsi que chez Syngenta, nous avons cartographié le portefeuille actif de l'entreprise, soit plus de 1 000 familles de brevets couvrant trois juridictions de délivrance de brevets.

[20] http://www.omniviz.com/content/omniviz

significatifs, on affine le découpage du corpus en amont et on gagne en finesse. À condition que cette finesse ne devienne pas excessive, auquel cas on peut toujours ajouter les nouvelles expressions jugées superflues à la liste de mots vides, selon le modèle précédent.

4 Conclusion

La cartographie des brevets désigne un ensemble varié de pratiques visant à représenter graphiquement un corpus de brevets et les informations issues de son analyse. Elle possède un fort pouvoir de séduction visuelle dans l'industrie et la recherche académique est convaincue de son intérêt heuristique depuis au moins les premières propositions de Derek J. De Solla Price (1965) visant à cartographier les articles scientifiques. On peut néanmoins ranger les outils et pratiques énumérés ici en plusieurs familles, selon qu'ils s'intéressent aux brevets (exemple des cartes sémantiques) ou aux acteurs (réseaux sociaux), aux données primaires (origine géographique) ou secondaires (réseaux de citation), aux données structurées (classifications macro-thématiques) ou non-structurées (cartes sémantiques). Certaines privilégient l'analyse quantitative des données quand d'autres cherchent à découvrir des motifs invisibles à première vue ou encore à mesurer la valeur qualitative d'un portefeuille de brevets grâce à l'analyse de citations.

Malgré tous ses attraits et retombées potentielles, il ne faut pas oublier que la cartographie des brevets est contrainte par les difficultés que nous soulignions dans le premier chapitre. Si les données sont inadaptées ou mal comprises, l'analyse ne vaut rien du tout. L'expérience montre que l'industrie tombe souvent dans ce piège, notamment à cause de la tentation du « tout, tout de suite », alors que la recherche est plus habituée à la rigueur et au doute systématique. La normalisation (Porter et Cunningham, 2005, chapitre 12.6), par exemple, ou la comparaison avec des études similaires effectuées dans d'autres domaines ou des études différentes effectuées dans le même domaine devraient faire partie des réflexes indispensables au spécialiste de la cartographie des brevets. Quant à la question de la reproductibilité de l'analyse, elle est résolue par certains auteurs comme Mogoutov et al. (2008) en utilisant des logiciels standards du commerce.

Références

Blanchard, A. (2006). La cartographie des brevets. *La Recherche 398*, 82–83.

Blanchard, A. (2007). Understanding and customizing stopword lists for enhanced patent mapping. *World Patent Information 29*(4), 308–316.

De Solla Price, D. J. (1965). Networks of scientific papers. *Science 149*(3683), 510–515.

Dou, H. J.-M. (2004). Benchmarking R&D and companies through patent analysis using free databases and special software : A tool to improve innovative thinking. *World Patent Information 26*(4), 297–309.

Eldridge, J. (2006). Data visualisation tools–a perspective from the pharmaceutical industry. *World Patent Information 28*(1), 43–49.

Fattori, M., G. Pedrazzi, et R. Turra (2003). Text mining applied to patent mapping : A practical business case. *World Patent Information 25*(4), 335–342.

Graff, G. D., S. E. Cullen, K. J. Bradford, D. Zilberman, et A. B. Benett (2003). The public-private structure of intellectual property ownership in agricultural biotechnology. *Nature Biotechnology 21*(9), 989–995.

Igami, M. (2008). Exploration of the evolution of nanotechnology via mapping of patent applications. *Scientometrics 77*(2), 153–171.

Jaffe, A. B. (1998). Patents, patent citations, and the dynamics of technological change. Technical report, National Bureau of Economic Research.

Lotka, A. J. (1926). The frequency distribution of scientific productivity. *Journal of the Washington Academy of Sciences 16*(12), 317–323.

Meyer, M. (2007). What do we know about innovation in nanotechnology ? Some propositions about an emerging field between hype and path-dependency. *Scientometrics 70*(3), 779–810.

Mogoutov, A., A. Cambrosio, P. Keating, et P. Mustar (2008). Biomedical innovation at the laboratory, clinical and commercial interface : A new method for mapping research projects, publications and patents in the field of microarrays. *Journal of Informetrics 2*(4), 341–353.

Oppenheim, C. (2000). Do patent citations count ? In B. Cronin et H. Barsky Atkins (Eds.), *The web of knowledge : A festschrift in honor of Eugene Garfield*, ASIS&T Monograph Series, pp. 405–432. Medford, NJ : Information Today Inc.

Porter, A. L. et S. W. Cunningham (2005). *Tech mining : Exploiting new technologies for competitive advantage*. New York : Wiley InterScience.

Salton, G. et C. Buckley (1988). Term weighting approaches in automatic text retrieval. *Information Processing and Management 24*(5), 513–523.

Small, H. (1999). Visualizing science by citation mapping. *Journal of the American Society for Information Science 50*(9), 799–813.

Sternitzke, C., A. Bartkowski, et R. Schramm (2008). Visualizing patent statistics by means of social network analysis tools. *World Patent Information 30*(2), 115–131.

Sternitzke, C., A. Bartkowski, H. Schwanbeck, et R. Schramm (2007). Patent and literature statistics — The case of optoelectronics. *World Patent Information 29*(4), 327–338.

Sun, T. et S. A. Morris (2008). Timeline and crossmap visualization of patents. In H. Kretschmer et F. Havemann (Eds.), *Proceedings of WIS 2008*, Humboldt Universität, Berlin, Allemagne.

Trippe, A. J. (2001). A comparison of ideologies : Intellectually assigned co-coding clustering vs ThemeScape automatic thememe mapping. In *Proceedings of the 2001 Chemical Information Conference*, Nîmes, France, pp. 61–79.

Trippe, A. J. (2002). Patinformatics : Identifiying haystacks from space. *Searcher 10*(9), 28.

von Wartburg, I., T. Teichert, et K. Rost (2005). Inventive progress measured by multi-stage patent citation analysis. *Research Policy 34*(10), 1591–1607.

Zitt, M. et E. Bassecoulard (2004). S&T networks and bibliometrics : The case of international scientific collaboration. In *4th Congress on Proximity Economics : Proximity, Networks and Co-ordination*, Marseille, France. GREQAM – IDEP.

Summary

The industry resorts to numerous information mapping, text mining and data mining technologies to tap into the patent information goldmine while academic researchers often content themselves with geographically limited information and familiar analysis tools. In this communication, we shall review as comprehensively as possible the practice of both domains and show how they could be brought closer to each other.