
© COURTESY OF IBM ARCHIVES

Comment savoir d'un seul coup d'œil si un objet ou un concept a fait l'objet d'un grand nombre de brevets ? Des outils mathématiques permettent de dessiner des cartes semblables aux cartes géographiques, sur lesquelles les technologies forment des régions facilement identifiables.

La cartographie des brevets

Antoine Blanchard,
Ingénieur information brevet chez Syngenta Crop Production.

Les brevets ne servent pas qu'à protéger les inventions. Les bases de données dans lesquelles ils sont consignés permettent aussi de brosser un état de l'art des technologies, sur lequel les entreprises peuvent s'appuyer dans leurs processus d'innovation. On considère en effet que 70 % à 90 % de l'information contenue dans les brevets ne se trouvent nulle part ailleurs, puisque, pour être brevetable, une invention doit être nouvelle et n'avoir jamais été rendue publique.

Deux stratégies

Un brevet peut permettre à une entreprise de sortir d'une impasse technologique ou d'invalider celui déposé par un concurrent. Encore faut-il qu'elle trouve exactement celui qu'elle cherche (s'il existe) dans les bases de données disponibles gratuitement sur Internet [1] ou proposées par des éditeurs privés. Il s'agit là de formuler la requête la plus précise possible pour interroger la base. Dans le cas où l'entreprise souhaite visualiser un portefeuille de brevets afin d'en optimiser la gestion, ou de comprendre les stratégies de recherche et développement de ses concurrents, le problème est tout autre. Il faut « prendre de la hauteur » en examinant un large corpus de documents et s'en faire une idée d'ensemble. Dans ce but, des algorithmes analysent des centaines voire des milliers de brevets pour en produire une image synthétique et fidèle : c'est le travail de la fouille de texte ou « text mining ».

Fouiller le texte

Un nombre croissant de logiciels est disponible pour fouiller le texte de tous types de documents. Ils traitent des textes et des phrases en langage naturel — dans le cas des brevets, ce sont surtout le titre, le résumé et les revendications (qui forment le cœur du brevet) qui sont analysés. Ces logiciels reposent sur une idée formulée en 1958 par un pionnier Hans Peter Luhn (en médaillon), selon laquelle *« la fréquence d'un mot dans un document donne une mesure utile de sa significativité »*. Il en découle que, dans une approche dite du « sac de mots » (« bag of words »), on peut faire fi des relations sémantiques et lexicales pour traiter les mots comme des objets mathématiques, les trier par un algorithme et en déduire un classement des documents.

Le poids des mots

La méthode la plus courante se déroule suivant deux étapes. Dans un premier temps, on collecte les mots qui constituent le corpus de documents et on filtre leurs déclinaisons. Par exemple, on fusionne les mots « stable », « stables » et « stabilité ». Puis on attribue un poids à chacun de ces groupes de déclinaisons selon un algorithme appelé « tf × idf » (« term frequency × inverse document frequency »), qui consiste à faire le produit de deux mesures complémentaires : la fréquence du mot (ou de l'ensemble de ses déclinaisons) dans le corpus et l'inverse de la proportion de documents qui contiennent ce mot. Avec cet algorithme publié en 1988 par Gerard Salton, à l'université Cornell et Chung-Shu Yang, à l'université de l'Iowa, le poids augmente quand le mot est utilisé abondamment et il diminue quand il apparaît dans de nombreux documents. On élimine ensuite les mots dont le poids est inférieur à un seuil fixé — dans la pratique, cela revient à éliminer des prépositions et des verbes qui reviennent dans tous les documents (« le », « un », « est », etc.), ainsi que les mots très rares. Il ne reste alors plus que les *n* mots pertinents qui apparaissent suffisamment de fois pour être considérés comme porteurs d'information significative.

Avec la deuxième étape, on passe du niveau des mots à celui des documents. L'algorithme parcourt chaque document, fait la liste des mots significatifs qu'il contient et exprime sa position dans un espace à *n* dimensions à l'aide d'un vecteur composé de 0 (mot absent du document) et de 1 (mot présent dans le document). Ainsi, les documents dont le contenu est similaire se retrouvent à proximité les uns des autres.

Carte de densités

On ramène l'espace de *n* à deux dimensions par une transformation de type « Self Organizing Map » (SOM), un mode de représentation développé par le Finlandais Teuvo Kohonen au début des années 1980 et qui fait appel aux réseaux de neurones. On obtient alors une carte où la distance entre les documents augmente quand leur similarité diminue. On y ajoute aussi des courbes topographiques pour transcrire la densité de documents ainsi que des dégradés de couleurs suivant cette densité, afin d'obtenir un rendu proche de celui des cartes géographiques.

Régions de montagnes

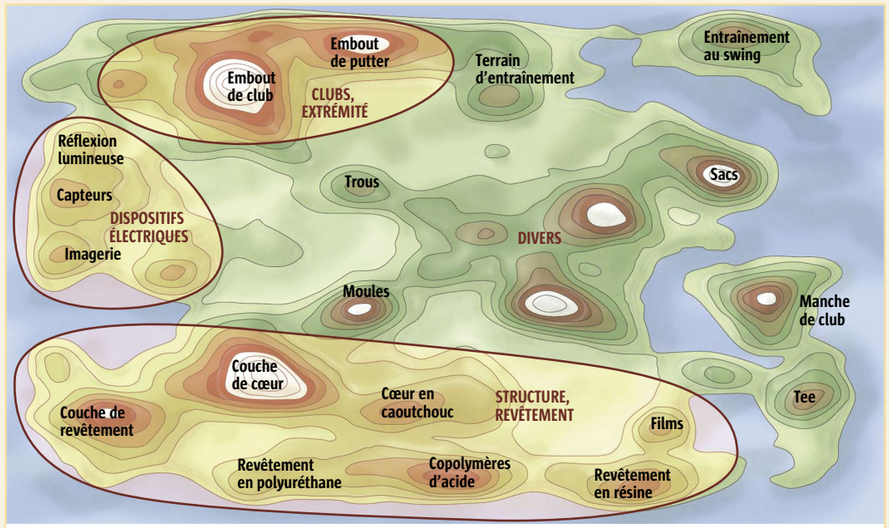
Sur le document final, on voit des « montagnes » denses de brevets relatifs à une même technologie. Les montagnes les plus proches forment des régions qui décrivent un ensemble de technologies apparentées. Entre différentes régions, on peut trouver des montagnes ou même des documents isolés. Si l'on nomme automatiquement les montagnes à partir des mots qui caractérisent leurs documents, il devient possible de lire le résultat d'un seul coup d'œil (en gardant à l'esprit les problèmes éventuels liés à la polysémie).

Le résultat de cette cartographie, ou « patent mapping », est-il fiable ? Comme elle fait abstraction du sens des mots rencontrés, la méthode est censée fonctionner quel que soit le sujet traité à condition que le nombre de documents soit assez grand. Mais dans certains cas, pour optimiser le résultat, l'utilisateur lui-même peut être amené à ajouter des mots que l'algorithme devra ignorer parce qu'ils sont triviaux dans le contexte considéré. Il s'agit alors d'une démarche empirique qui échappe au formalisme mathématique, avec les incertitudes que cela comporte.

Compléter l'information

Pour aller au-delà, l'information obtenue à partir de la fouille de texte peut être complétée par l'information contenue dans les champs numériques des brevets (inventeur, année de dépôt, classification internationale) : il s'agit alors de fouille de données ou « data mining ». Celle-ci permet par exemple d'obtenir facilement des graphiques traduisant les tendances de dépôt de brevets au cours du temps. On peut aussi analyser les réseaux de citations entre brevets de la même façon que l'on étudie les réseaux de citations entre articles scientifiques, afin d'estimer quels sont les brevets les plus importants (en première approximation, les plus cités) ou retracer l'évolution d'une technologie.

Cette approche cartographique permet d'appréhender comment une technologie se construit. Ainsi, l'équipe de Gregory Graff s'est récemment appuyée sur une carte des brevets dans le domaine des biotechnologies pour l'agriculture afin de comprendre comment l'innovation y est partagée entre les secteurs public et privé [2]. Quant aux investisseurs et aux consultants, ils font de plus en plus appel à ce type de cartes pour anticiper les futurs développements et les tendances de l'innovation. ■■



SUR CETTE CARTE apparaissent 16 579 brevets et demandes de brevets liés au concept de « balle de golf ». Les principales technologies sont regroupées en régions : « Extrémités de clubs », « dispositifs électriques » et « structure et revêtement des balles ». Les courbes topographiques indiquent la densité de documents et les couleurs imitent celles des cartes géographiques. © INFOGRAPHIE BRUNO BOURGEOIS/D'APRÈS THOMSON SCIENTIFIC

[1] www. uspto.gov/patft ; http://ep. espacenet.com

[2] G. Graff, *et al., Nature Biotechnology, 21,* 989, 2003.