BASES

POUR UNE RECHERCHE INTELLIGENTE D'INFORMATION

N°395 • Septembre 2021

SOMMAIRE

IA et bases de données brevets : avec IPRally, les professionnels seront-ils bons pour le musée ?

Philippe Borne

IA, deep learning, recherche sémantique, classification automatique : ces termes sont de plus en plus fréquents dans le monde des bases de données brevets.

Vont-ils renvoyer les tenants de la recherche traditionnelle au rayon des archives du monde des professionnels de l'information brevet ? Les codes CIB, CPC, les mots-clés représenteront-ils bientôt des techniques démodées à remiser au placard ? Qui sont ces nouveaux outils et condamnent-ils réellement des techniques éprouvées depuis plusieurs dizaines d'années, ou au contraire ne font-ils que les complémenter ? Enfin, quel est leur niveau de performance et comment les utilise-t-on ?

Dans cet article, nous faisons le point sur le sujet en partant de l'exemple d'IPRally, un nouvel outil de recherche brevets qui place l'IA au cœur de son produit. Après en avoir présenté les fonctionnalités, nous évaluerons les performances de l'outil à partir des tests que nous avons réalisés.

Les informations ici présentées résultent également d'un échange avec les équipes d'IPRally.

IPRally : une sorte de retour à «l'esprit startup»

IPRally fait partie des nouveaux acteurs récemment arrivés sur le marché de l'information brevet proposant des produits payants, appuyés uniquement sur les nouvelles technologies ; celles-ci résumées par les termes cités plus haut : IA, sémantique, *deep learning*.

La startup a été fondée il y a 3 ans et demi en Finlande par Sakari Arvela, à l'origine, conseil en PI (Propriété Intellectuelle). Constatant la manière dont il décompose les revendications en concepts essentiels dans le cadre de son travail journalier de recherche de brevetabilité ou de liberté d'exploitation, il a eu l'idée d'apprendre à une IA cette méthode et de tenter d'automatiser un processus jusque-là intellectuel.

IP Rally annonçait 12 collaborateurs en 2020 et emploie aujourd'hui 25 personnes. Il a bénéficié en janvier dernier d'une levée de fonds de 2 M€.

IA et bases de données brevets : avec IPRally, les professionnels seront-ils bons pour le musée ? *suite*

Un Knowledge Graph et 3 modes de définition de la question permettant une certaine transparence

Le moteur de recherche est fondé sur un Knowledge Graph (graphe de connaissance) généré pour chaque question et chaque demande de brevet chargée dans la base de données.

Pour en savoir plus sur les Knowledge Graphs, vous pouvez lire ou relire notre article « Les Knowledge Graphs vont-ils enfin révolutionner la recherche d'information professionnelle ? » (BASES n°394 – juillet/août 2021).

Ce graphe de connaissance correspond peu ou prou au tableau des caractéristiques essentielles – dénommé en anglais *claim chart* – que l'on construit par exemple pour déterminer les mots-clés dans la préparation d'une recherche traditionnelle, ou pour comparer des brevets entre eux.

Ces caractéristiques essentielles - par exemple « *moteur* », « élément luminescent », « *gicleur* » - peuvent s'imbriquer les unes dans les autres selon une hiérarchie, avec des caractéristiques principales et des caractéristiques secondaires qui dépendent des principales.

« *Washing machine* », caractéristique principale, peut par exemple couvrir plusieurs caractéristiques filles comme « *rotatable water jet member* » (car la machine à laver contient un « *rotatable water jet member* »), ou comme « *dishwashing machine* » (car la machine à laver est une machine à laver la vaisselle).

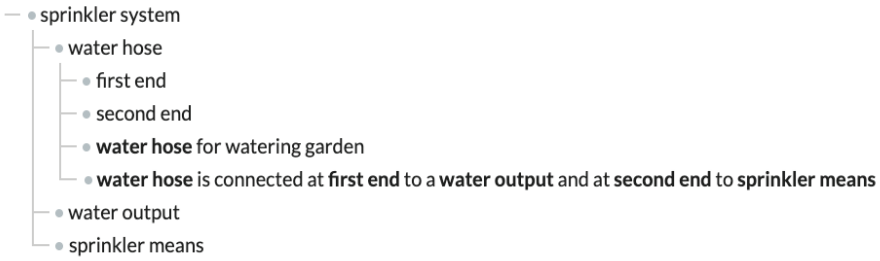

Figure 1. Exemple du Knowledge Graph d'IPRally

Le Knowledge Graph utilisé par IPRally comporte également des « relations », qui permettent de décrire des liens de nature technique entre les caractéristiques essentielles. Ces relations sont hiérarchiquement positionnées dans le graphe sous les caractéristiques essentielles.

À titre d'exemple, dans la figure 1. (cf. Figure 1. Exemple du Knowledge Graph d'IPRally), on a une caractéristique essentielle parente (*sprinkler system*) couvrant 3 caractéristiques secondaires («*water hose*», «*water output*» et «*sprinkler means*»), 4 relations se trouvant sous «*water hose*».

Ces Knowledge Graphs sont chargés dans un réseau de neurones utilisé pour comparer le graphe de la question avec les graphes de la base de données, les résultats étant triés en fonction de leur similarité avec la question.

La question peut être entrée de 3 manières différentes :

1. En construisant son propre graphe ;
2. En saisissant un numéro de publication de demande de brevet ;
3. Ou en entrant un texte, par exemple un projet de revendication.

Dans les deux derniers cas, le Knowledge Graph généré pour la question peut être affiché et même édité, ce qui permet d'avoir un certain aperçu de la manière dont fonctionne la recherche.

La saisie d'un numéro de publication est bien entendu présentée comme très utile dans une recherche d'invalidation. L'utilisateur peut demander que le texte intégral de la publication soit pris en compte, ou seulement une des revendications. Par défaut, seuls sont recherchés les documents publiés avant la date de dépôt de la demande utilisée en entrée, cette date limite pouvant être changée.

Une fonction de surveillance est disponible qui permet de réexécuter automatiquement une recherche et d'être informé des nouvelles réponses.

Une couverture satisfaisante

Chaque réponse obtenue correspond à une famille de brevets. IPRally couvre environ 80 millions de documents, demandes de brevet et brevets issus de 25 offices (en particulier US, EP, WO, CN, JP et KR).

IPRally se fournit chez LexisNexis et IFI Claims.

Les documents disponibles dans une langue autre que l'anglais font l'objet d'une traduction machine vers l'anglais permettant ensuite de générer les graphes de connaissance. Les familles affichées pour chaque document retrouvé proviennent quant à elle de la base de données Inpadoc de l'OEB.

Une page dédiée dans l'aide (http://help.iprally.com/en/articles/3270009-data-coverage) donne des informations détaillées sur les périodes couvertes pour chaque type de document.

Cette page annonce aussi que « typiquement » les documents sont présents dans la base de données 14 jours après leur date de publication. Cette valeur est probablement une moyenne, le délai pour chaque office dépendant de la fourniture par lesdits offices de brevet de l'information à IFI Claims et Total Patent de LexisNexis, fournisseurs d'IPRally.

Attention toutefois, IPRally ne couvre pas la littérature non brevet.

Un affichage des résultats intuitif et adapté à une recherche basée sur l'IA

Par défaut, 50 réponses sont affichées, mais ceci peut être paramétré jusqu'à un maximum de 500. Dans une telle recherche, il est bien entendu moins aisé de déterminer ce qui est pertinent et ce qui ne l'est pas que dans une recherche traditionnelle basée sur mots-clés et codes de classification.

Les réponses peuvent être triées selon un score de pertinence, mais aussi selon la date de publication, le nom du déposant, et le titre par ordre alphabétique. Par défaut, on obtient un affichage avec les mosaïques c'est-à-dire les figures (*cf.* Figure 2. Mode d'affichage complet avec mosaïque) qu'il est possible de masquer pour n'avoir que le titre. L'affichage peut être étendu aux champs habituels (abrégé, CIB, CPC, inventeurs, etc.).

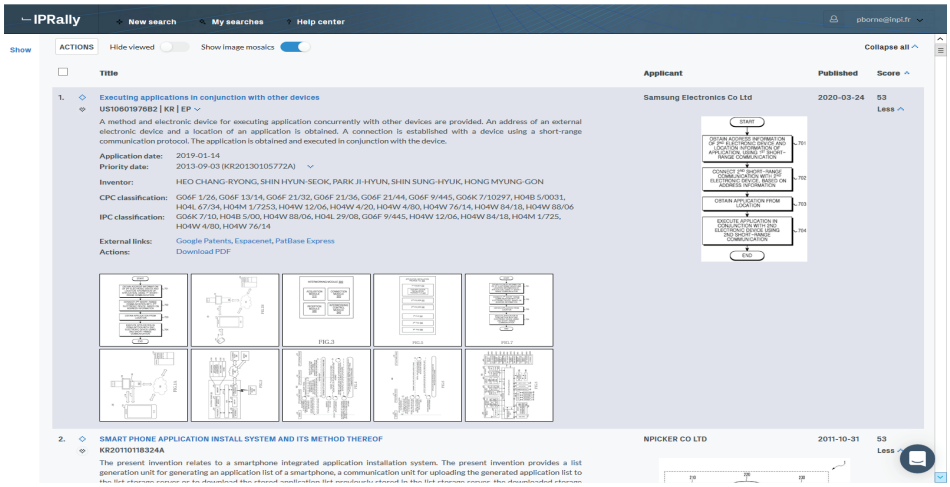

Figure 2. Mode d'affichage complet avec mosaïque

Des liens vers Google Patents, Espacenet et – pour ceux qui disposent d'un accès – Patbase Express permettent d'accéder au document complet.

Un mode d'affichage particulier obtenu en cliquant sur le titre du document permet d'obtenir deux volets : un à gauche comportant les titres des documents obtenus, et un à droite affichant pour chaque document sélectionné les données bibliographiques et les figures suivies du texte complet du document où sont surlignées dans des nuances de jaunes les parties du texte pertinentes par rapport à la question. Le degré de pertinence est reflété par l'intensité de la couleur jaune.

L'utilisateur peut en outre lui-même décider de surligner jusqu'à 24 termes spécifiques dans des couleurs différentes. Une même couleur peut être affectée à un groupe de termes.

Pour les documents publiés dans une langue autre que l'anglais, titre et abrégé sont affichés en anglais *via* traduction machine ou humaine, selon la pratique des offices. Le texte intégral affiché en anglais provient d'une traduction machine.

Des possibilités de préciser la recherche dans la logique du système

Selon les résultats de l'analyse des réponses obtenues, il est possible d'affiner la recherche de deux manières différentes.

D'une manière très classique en premier lieu, *via* codes CIB ou CPC, mots-clés, mais aussi intervalle de dates de publication, noms de déposant et d'inventeur, ou codes pays de l'office de dépôt.

Il est aussi possible de demander au moteur de recherche de prendre en compte des réponses favorites (fonction « *Zoom to favorites* »), considérées donc comme particulièrement pertinentes, sélectionnées en cours d'affichage.

Il faut préciser que si le nombre maximum de réponses que IPRally permet d'afficher est limité à 500, en réalité le moteur de recherche basé sur l'IA retrouve environ 10 000 documents qu'il considère comme étant les plus proches de la question. Seuls donc les 500 premiers sont affichables. Lorsque l'on précise la question *via* la fonction « *Zoom to favorites* » ou les filtres (codes CIB ou CPC par exemple), le moteur de recherche prend en compte ces « favoris » ou ces filtres pour réexécuter la recherche non pas seulement sur les 500 premiers résultats, mais sur la totalité des 10 000 – environ – retrouvés au départ.

L'outil fait l'objet d'améliorations régulières, la dernière en date (juin 2021) étant la possibilité d'exclure les documents comprenant un mot-clé particulier lorsque l'on précise la recherche.

Des fonctions d'analyse statistique et d'export minimalistes

IPRally ne permettant pas d'afficher plus de 500 réponses, les fonctions d'analyse statistique sont donc minimalistes.

Nous dirons que ce n'est pas ce qui motivera l'utilisation du produit.

Elles sont accessibles *via* un onglet dédié qui affiche sous-groupes CPC, déposants, inventeurs, offices de dépôt et années de publication les plus fréquents. Une fonction permet d'exporter sous format Excel un fichier comprenant les champs bibliographiques classiques, de même que le score de pertinence.

Des aides et tutoriels permettent de prendre en main le produit.

L'épineuse question de la sécurité des données

C'est une question qui ne se pose pas que pour IPRally, elle se pose dès que l'on entre une requête sur un moteur de recherche et que cette requête transite *via* divers réseaux dont on ne connaît pas le niveau réel de sécurisation.

Dans le cas de produits comme IPRally, la question est plus aiguë peut-être, car un mode de saisie de la question permet d'entrer une requête non pas constituée uniquement de mots-clés ou de codes de classification, mais de la description détaillée de l'invention elle-même. IPRally est basé en Finlande, en Europe donc, et même dans l'UE, ce qui peut certainement apaiser les craintes que pourraient susciter d'autres localisations.

Pour les grands comptes, notamment les offices de brevet, IPRally propose une solution plus rassurante : l'installation sur le site même de l'utilisateur.

Les autres utilisateurs peuvent tirer avantage d'un chiffrage à 3 couches présenté comme garantissant un haut niveau de sécurité. Des paramétrages permettent par ailleurs d'éviter de sauvegarder des informations considérées comme sensibles sur le système, en particulier les questions.

Les outils de recherche à base d'IA : des performances intéressantes

La question centrale est bien entendu celle de la performance de ce type d'outil. Peuvent-ils remplacer ou seulement compléter les techniques de recherche traditionnelles, ou les résultats sont-ils simplement totalement hors sujet ?

Burkhard Schlechter, qui travaillait encore récemment pour l'Office autrichien des brevets, a mené des tests comparant la performance des techniques classiques à celles de plusieurs outils de recherche basés sur l'IA, en particulier IPRally. Ces tests montrent qu'en moyenne les outils retrouvent 10% des documents pertinents retrouvés par les techniques classiques, IPRally se situant plutôt autour de 30%.

Pour une technique jeune, c'est plutôt prometteur.

Nous avons de notre côté réalisé deux types de tests, bien entendu plus rapides. Mais les résultats sont également intéressants.

D'une part, nous avons utilisé deux inventions caractérisées par des mots-clés que nous qualifierions de « très parlants ».

Le premier sujet de recherche concernait : *Adhesive fly trap for capturing insects comprising a sheet on a surface of which an adhesive is spread, where the surface to be exposed is treated with at least one ink which contains luminescent pigments.* Les caractéristiques essentielles sont ici « *adhesive* », « *trap* », « *fly* » et « *luminescent* ». C'est le texte qui précède qui a été utilisé comme requête : de manière bluffante, la réponse cible apparaît d'emblée en tête de la liste de résultat.

Le second sujet concernait un `Portable casing for tennis balls that has the ability to pressurize the balls in order to maintain them longer usable. The idea is that it should look like a normal tennis ball case (Cylinder with 2 to 8 balls), but it should incorporate a small pressurizing mechanism (preferably battery driven).` Ce sujet provient d'une formation récemment organisée par l'OEB.

Dans ce cas-là, la réponse cible ne fut pas obtenue d'emblée, ce qui est lié au fait que le *Knowledge Graph* produit pour la requête ne prend pas en compte la caractéristique « *battery driven* ».

Comme expliqué plus haut, le graphe est éditable, et après transformation, on peut obtenir le graphe « question » que l'on peut voir en figure 6 (cf. Figure 6. Édition du Knowledge Graph pour mieux répondre à la question).

Ce graphe de connaissance modifié a alors retrouvé sans peine le document cible (DE102011016806A1).

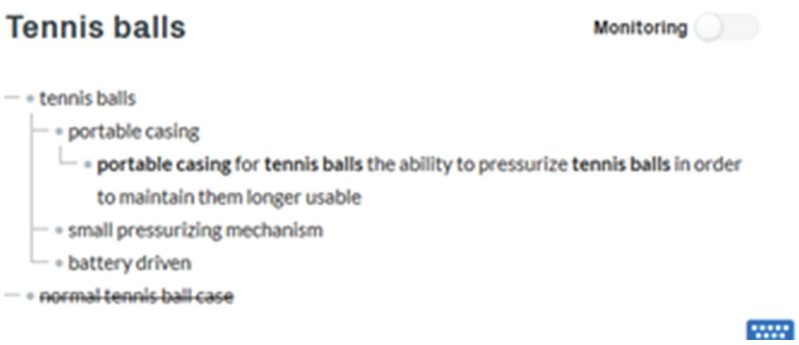

Figure 6. Édition du Knowledge Graph pour mieux répondre à la question

Dans un second temps, nous avons testé la recherche dite d'invalidation, consistant à entrer un numéro de publication, le système cherche alors les documents de l'art antérieur susceptibles de permettre une annulation du brevet correspondant. Il s'agissait de vérifier si IPRally retrouvait les documents X cités dans le rapport de recherche attaché au document servant de base au test.

Nous avons utilisé 15 demandes de brevet européen publiées en août 2021, comprenant un total cumulé de 52 documents cités avec code X. IPRally a retrouvé 10 de ces documents X, correspondant à 9 des 15 demandes de brevet européen de départ.

Attention « retrouvé par IPRally » signifie « faisant partie des 50 premières réponses retrouvées par IPRally ».

Cette performance est tout à fait honorable ; d'autant plus que, pour rappel, une fois un document pertinent identifié **on peut le marquer comme favori et demander une réexécution de la recherche prenant en compte ce favori, opération susceptible de faciliter l'obtention de nouveaux documents pertinents.**

Dans un second temps, nous avons utilisé 15 demandes européennes avec là encore des documents X dans leur rapport de recherche et publiés suffisamment récemment pour ne pas être encore chargées dans la base de données. Pour tenter de retrouver ces documents X, nous avons cette fois utilisé comme requête un texte libre constitué du jeu de revendications intégral de chaque document servant de base au test. Nous avons limité la recherche aux documents publiés avant la date de dépôt ou de priorité dudit document de départ.

Le résultat fut plus mitigé : sur les 53 documents X présents dans nos 15 demandes de départ, IPRally n'en a retrouvé que 5 correspondants à 4 des 15 documents en question (« retrouvé par IPRally » étant défini comme plus haut).

Nous avons toutefois poursuivi par une expérience intéressante : pour une des 15 demandes où IPRally ne parvenait pas à retrouver un des documents X lui étant associé, nous avons modifié le graphe de connaissance produit à partir de la revendication indépendante. L'impact de l'opération fut convaincant : un document X classé 92e lors de la première recherche dans la liste de réponses s'est retrouvé en 37e position ; et un second document X ne figurant pas parmi les 250 premières réponses de la liste s'est retrouvé en 167e position.

L'enseignement à en tirer est clair : l'apprentissage de la manipulation des graphes de connaissance peut significativement améliorer les résultats. Ce qui veut aussi dire qu'IPRally n'est pas uniquement réservé au débutant qui veut simplement copier/coller un texte dans un masque de recherche sans aller plus loin.

Point important : il n'a pas été vérifié si parmi les 50 réponses proposées par IPRally, certaines non repérées par l'examinateur méritaient un code X. Une telle vérification aurait peut-être permis d'améliorer la performance constatée.

IA et recherche brevet : quel futur ?

Ces techniques vont-elles donc rendre obsolète la recherche traditionnelle ?

Les tests rapportés ici ne permettent pas d'annoncer une telle révolution, tout au moins sur le court terme. Ces produits se positionnent actuellement plus comme des compléments que comme des substituts aux techniques classiques.

Attention, on aurait tort de tirer argument de résultats parfois mitigés pour condamner ces produits.

En premier lieu, car ces techniques sont jeunes, et ont encore une marge de progression importante.

En second lieu, car ces technologies répondent à une nécessité. À côté de l'intérêt scientifique qu'il y a à développer ces nouveaux outils, la motivation est en effet très pragmatique : face à la masse croissante d'information – constituant l'art antérieur – les techniques traditionnelles de recherche atteignent leurs limites, les temps de recherche risquent de devenir de plus en plus longs, ou, si on les maintient pour des raisons économiques dans une limite tolérable, c'est la qualité elle-même des recherches qui pourra s'en trouver affectée, et donc en particulier la sécurité juridique des titres délivrés par les offices. Il faut donc développer de nouvelles méthodes aptes à traiter la masse croissante d'information à laquelle est confronté le documentaliste brevet, et l'IA offre une opportunité à saisir.

En l'état actuel, il convient donc de suivre de près ces produits, voire de prendre un abonnement en complément à un accès plus classique, si bien entendu le budget le permet.

Mais on peut parier qu'il sera d'ici peu même inutile de prendre un abonnement complémentaire, l'intérêt pour ces nouvelles technologies nous promettant de probables mouvements de concentration. Octimine a par exemple déjà été racheté, en 2018, mais pas par un serveur traditionnel, l'acquéreur étant Dennemeyer, acteur qui combine une activité traditionnelle dans le conseil à la conception d'une offre logiciel en matière de gestion de brevets et de marques.

Les tarifs d'IPRally

5000 €/an pour un accès, ce qui revient à sensiblement doubler le budget base de données brevet si on opte pour IPRally en plus d'un serveur traditionnel

10 000 €/an pour 3 accès