PatClust: une plateforme pour la classification sémantique des brevets

Abdoulaye Guissé, Khaled Khelif, Martine Collard

PatClust : une plateforme pour la classification sémantique des brevets

Abdoulaye Guisse, Khaled Khelif, Martine Collard

INRIA Sophia Antipolis - Méditerranée, Equipe-projet Edelweiss
{prenom.nom}@sophia.inria.fr

Résumé : Nous présentons ici une approche générique pour la classification des brevets fondée sur la sémantique contenue dans ces documents.

Mots-clés : Annotation sémantique, ontologie, classification

1 Introduction

Aujourd'hui, l'Internet est devenu une véritable infrastructure, qui englobe une grande diversité d'informations ; c'est un gigantesque milieu d'échanges de documents électroniques caractérisé par une facilité d'accès et la multiplicité des acteurs dans la mise en place et l'exploitation de ses ressources. A travers le Web, de nouvelles possibilités de recherche d'informations se présentent aux internautes leur offrant la liberté d'accéder à l'ensemble des connaissances enfouies dans une grande variété de sources dont : revues électroniques, bases de données, articles scientifiques, et en particulier, les brevets déposés par leur inventeur auprès d'organismes dédiés.

Récemment plusieurs travaux se sont orientés dans l'exploration du contenu des brevets. Un brevet est un document qui confère un droit exclusif sur une invention, qui est un produit ou procédé offrant, en règle générale, une nouvelle manière de procéder ou apportant une nouvelle solution technique à un problème. L'exploitation de ces types de documents offre aujourd'hui des opportunités de recherches scientifiques et apparait dans différents domaine tels que le web sémantique, l'ingénierie des connaissances et la fouille de données. La combinaison de ces trois axes constitue un des domaines le plus prometteurs dans l'exploitation des brevets.

Une des approches pour la classification des brevets consiste à combler les manquements des techniques standards en les associant aux technologies du Web Sémantique. L'enjeu dans ce présent travail consiste à utiliser les technologies du web sémantique afin d'étendre les techniques standards de classification des documents textuels. Pour résumer, nous pouvons dire que l'idée est d'apporter une couche de sémantique dans les résultats de classification dans le but de mieux faciliter leurs évaluations et leurs interprétations.

2 Les approches de classification sémantiques

L'objectif est d'utiliser les opportunités offertes par les technologies du web sémantique, en exploitant les liens sémantiques entre les concepts décrivant notre jeu de données. Plus précisément, il s'agit d'exploiter la hiérarchie et les relations entre les concepts de l'ontologie MeSH du domaine biomédical. Comme dans l'approche standard le corpus de brevets est représenté sous forme d'espace vectoriel où le poids d'un terme (instance d'un concept de l'ontologie) est soit sa fréquence (TF) soit sa fréquence pondérée par un coefficient inverse de sa fréquence dans le corpus entier (TF-IDF). Nous avons défini deux approches sémantiques décrites ci-dessous.

2.1 Approche par propagation des poids

L'objectif est de prendre en considération la hiérarchie des concepts dans l'ontologie afin d'élargir ou de propager notre espace de données en complétant les concepts décrivant les brevets du corpus avec leurs concepts pères (en propageant le poids). Ainsi, chaque brevet est décrit par ses concepts initiaux et par les concepts pères de ces derniers. Toutefois, le poids, d'un concept père, est défini en fonction du poids du concept fils, de sa profondeur et de la profondeur du concept fils. La profondeur d'un concept correspond à son niveau dans l'ontologie. Ainsi, le poids d'un concept père est défini comme suit :

$$P(c_p) = P(c_p) + \sum_{f \in F(Cp)} \frac{P(c_f)}{2^{depth(c_f) - depth(c_p)}}$$

$P(c_p)$: Le poids du concept père.
$P(c_f)$: Le poids du concept fils qui est soit sa fréquence soit son TF-IDF.
$depth(c)$: La profondeur d'un concept c dans l'ontologie.
$F(C_p)$: Ensemble des concepts pères de C

2.2 Approche par distance sémantique

Toute technique de classification non-supervisée définit une fonction de similarité pour évaluer la distance entre deux objets. Cette approche consiste à soumettre à ces algorithmes une fonction de similarité basée sur la similarité sémantique entre les textes des brevets. En effet, la similarité entre deux textes est définie en fonction des similarités de leurs concepts initiaux (c.à.d dont ils contiennent l'instance). La similarité entre deux concepts est calculée en fonction de leur distance sémantique dans l'ontologie. Dans cette étude, nous proposons d'utiliser la définition de la distance sémantique entre concepts d'ontologie du moteur sémantique CORESE (Corby et al, 2006), développé au sein de l'équipe Edelweiss. Cette distance est obtenue en calculant la longueur du plus court chemin entre les deux concepts dans l'ontologie. Les arcs de ce chemin sont pondérés par un poids calculé en prenant

compte la profondeur de chaque concept de l'ontologie. Ainsi la similarité sémantique entre deux concepts $sim(c_i, c_j)$ est une normalisation de cette distance.

Ainsi, nous avons défini deux fonctions de similarité entre les brevets selon que le poids soit défini par la fréquence ou par le TF-IDF :

$$simsem\ (b_l, b_m) = \sum_{\substack{c_i \in C^l \\ c_j \in C^m}} \left(1 - \left|P_{c_i}^{b_l} - P_{c_j}^{b_m}\right|\right)\left[sim\ (c_i, c_j)\right]$$

$$simsem\ (b_l, b_m) = \sum_{\substack{c_i \in C^l \\ c_j \in C^m}} \frac{\left[sim\ (c_i, c_j)\right]}{1 + \left|P_{c_i}^{b_l} - P_{c_j}^{b_m}\right|}$$

C^x : L'ensemble des concepts décrivant le brevet b_x de poids différent de 0.
$P_{c_i}^a$: Le poids du concept c_i dans le brevet b_a.
Ces deux fonctions sont définies comme une somme pondérée des similarités entre concepts deux à deux décrivant chaque brevet.

3 Architecture de PatClust

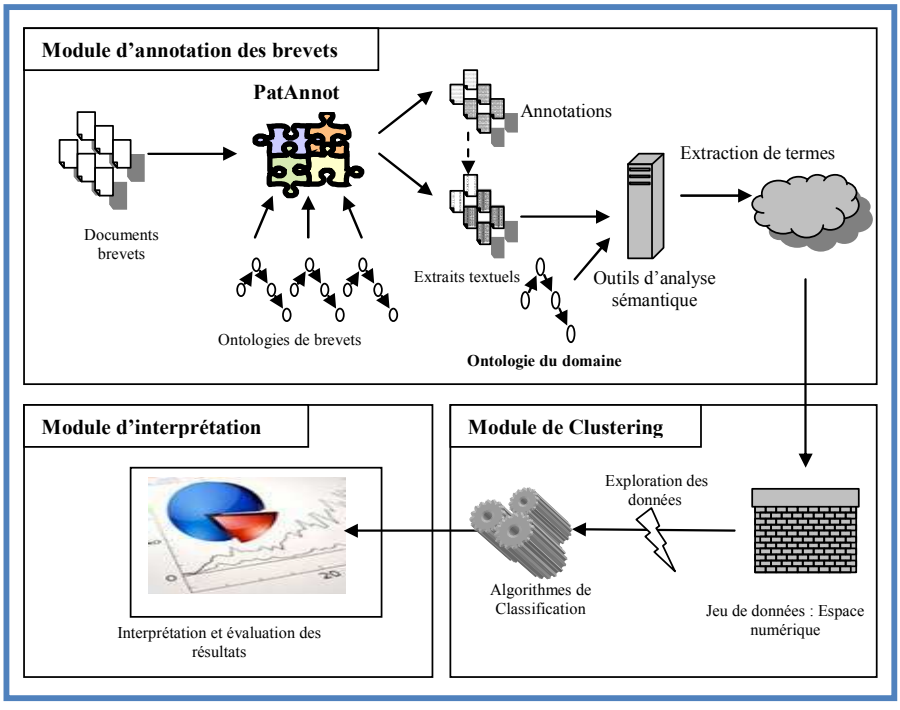

Fig. 1 - Architecture globale de la plateforme PatClust

PatClust est construit selon une architecture modulaire qui supporte le processus entier de classification sémantique d'un corpus de brevets allant du prétraitement des documents brevets à l'évaluation des résultats de classification. Ainsi, l'architecture repose sur trois grands modules (voir Figure 1) : le module d'**annotation des brevets**, le module de **clustering** (classification) et le module d'**interprétation des résultats**

3.1 Module d'annotation des brevets

Ce module est dédié au pré-traitement des documents sources. L'outil PatAnnot (Ghoula et al,. 2007)permet le formatage des documents brevets sources en de nouvelles structures de données plus faciles à explorer et permettre l'extraction des termes (variables descriptives) sur lesquels portera la classification. Le rôle de PatAnnot consiste à annoter les documents brevets suivant des ontologies construites via la structure d'un document brevets et à extraire les parties textuelles (« Abstract ») associées à chaque section d'un document brevet. Toutefois, l'extraction des termes ou concepts sémantiques n'est pas effectuée par cette approche mais plutôt par l'outil MeatAnnot (Khelif et al., 2007). Cet outil charge également l'ontologie du domaine MeSH afin d'extraire les concepts d'ontologies contenus dans ces textes.

3.2 Module de Clustering

Ce module permet de construire l'espace numérique et de lancer des tâches de classification. Le module implémente deux approches de classification : une approche standard basée sur la mesure de distance standard (cosinus) et une approche sémantique basée d'une part sur les liens hiérarchiques entre les entités sémantiques dans l'ontologie du domaine et d'autre par sur la similarité sémantique définie dans Section 2. Il repose, dans la version actuelle, sur deux types d'algorithmes, les algorithmes par partitionnement de type KMeans et les algorithmes hybrides de type ***Bisecting K-Means***.

3.3 Module d'interprétation des résultats

Ce module consiste à évaluer et interpréter l'ensemble des résultats de classification. L'**évaluation** s'effectue selon des indicateurs statistiques de similarité interne et externes des clusters extraits. L**'interprétation** consiste à décrire chaque cluster à l'aide de ses concepts les plus descriptifs, c.à.d. les concepts apparaissant le plus fréquemment dans les brevets classifiés dans le cluster. PatClust implémente les cinq indicateurs statistiques suivants:

Isim qui définit la similarité interne de chaque cluster de la classification. Nous jugerons de la cohérence de la classification que si cette similarité est élevée.

Esim qui définit la similarité externe de chaque cluster de la classification. Cette similarité permet de juger de la cohérence de la classification que lors qu'elle présente une valeur faible. Ainsi, plus ces similarités sont petites meilleure est la classification. Elles sont définies selon les formules suivantes :

$$Isim(S_i) = \frac{1}{|S_i|^2} \sum_{\substack{d_m \in S_i \\ d_n \in S_i}} sim(d_m, d_n) \qquad Esim(S_i) = \frac{\sum_{\substack{d_i \in S_i \\ d_j \in S}} sim(d_i, d_j)}{\sum_{\substack{d_i \in S_i \\ d \in S_i}} sim(d_i, d)}$$

Où S_i représente l'ensemble des textes d'un cluster et S l'ensemble initial des textes.

EsimCentroides définit la similarité externe du centre de chaque cluster avec les autres les centres des clusters restants. Elle est définie selon la formule suivante :

$$EsimCentroides(S_i) = sim(C_i, C)$$

Où :

S_i Correspond au Cluster, C_i Centres de S_i et C Ensemble initial des textes.

Isimsem qui définit la similarité interne sémantique de chaque cluster. Elle est définie sous la base de similarité Isim tout en considérant les similarités entre les objets d'un même cluster.

Esimsem qui définit la similarité externe sémantique de chaque cluster. Elle est définie sous la base de similarité Esim tout en considérant les similarités entre les objets d'un même cluster et les similarités entre ces mêmes objets et les objets appartenant aux autres clusters.

Elles sont définies selon les formules suivantes :

$$Isimsem(S_i) = \frac{1}{|S_i|^2} \sum_{\substack{d_m \in S_i \\ d_n \in S_i}} simsem(d_m, d_n) \qquad Esimsem(S_i) = \frac{\sum_{\substack{d_i \in S_i \\ d_j \in S}} simsem(d_i, d_j)}{\sum_{\substack{d_i \in S_i \\ d \in S_i}} simsem(d_i, d)}$$

4 Evaluation et interprétation des résultats

Nous présentons ici les résultats obtenus avec les par un algorithme de classification hybride sur 7 clusters. La Table 1 donne la moyenne des valeurs des cinq indicateurs statistiques de similarité obtenus selon chacune des trois approches. Ces résultats montrent que les approches sémantiques par propagation des poids et par distance sémantique améliorent très clairement la cohérence des clusters non seulement du point de vue de la similarité sémantique, mais également du point de vue de la similarité standard. Les figures qui suivent, obtenus avec le logiciel CLUTO[1], permettent de donner une vue globale sur la qualité de la classification. Chaque cluster est représenté par un pic dont la hauteur est proportionnelle à la similarité interne dans le cluster et dont la taille est proportionnelle au nombre d'éléments qu'il contient. Cette représentation confirme nos hypothèses de départ à savoir que les approches sémantiques fournissent des clusters plus uniformes et mieux distribués.

[1]http://glaros.dtc.umn.edu/gkhome/cluto/cluto/overview

	Isim	Esim	EsimCentroides	Isimsem	Esimsem
Approche standard	0.309	0.201	0.749	0.162	0.015
Propagation des poids	0.810	0.160	0.937	0.500	0.017
Distance sémantique	0.794	0.151	0.999	0.515	0.016

Tab. 1 : Comparaison des trois approches sur les TF-IDF

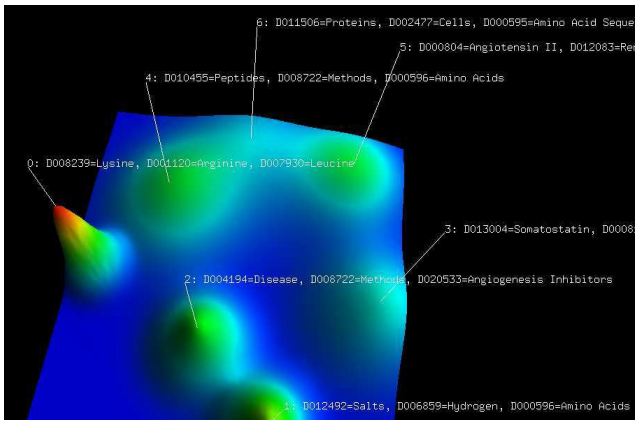

Fig. 2 : Les clusters dans l'approche Standard sur les TF-IDF

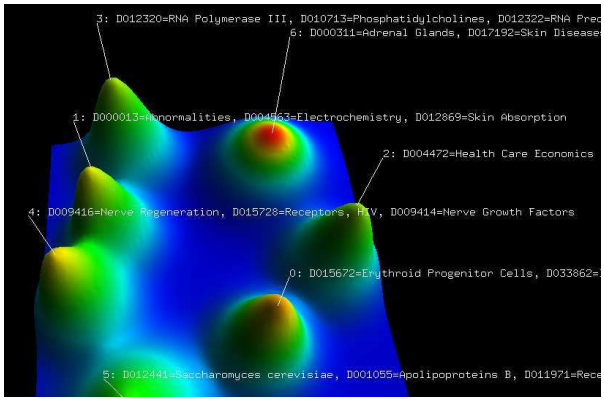

Fig. 3 : Les concepts les plus descriptifs pour chaque cluster dans l'approche par propagation sur les TF-IDF

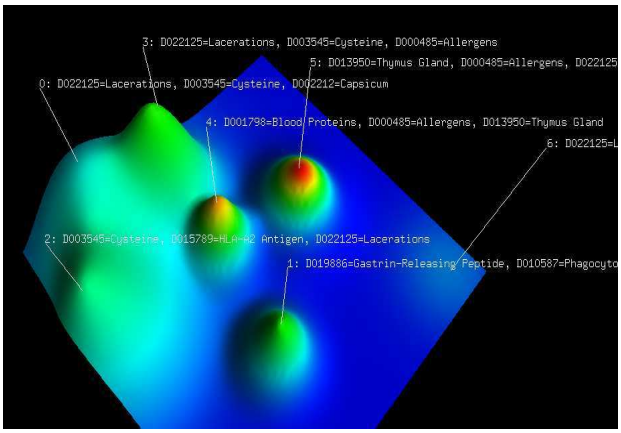

Fig. 4 : Les concepts les plus descriptifs pour chaque cluster dans l'approche par distance sémantique sur les TF-IDF

5 Conclusion

Ce travail s'intègre dans le projet européen Sealife (IST-2006-027269) dont l'objectif est de proposer un navigateur sémantique pour des communautés d'utilisateurs du domaine de la biomédecine. Ce navigateur doit faciliter, entre autre, la tâche de l'utilisateur en lui proposant des sources d'informations d'intérêt en fonction de son profil. Un des éléments de ce projet concerne la fouille et la recherche intelligente des brevets basées sur des techniques du web sémantique.

Références

CORBY O. & DIENG-KUNTZ R. & FARON-ZUCKER C. & GANDON F. (2006). Searching the Semantic Web: Approximate Query Processing based on Ontologies. In IEEE Intelligent Systems & their Applications, vol. 21, no 1, p. 20-27, 2006.

KHELIF K. & DIENG-KUNTZ R. & BARBRY P. (2007). An Ontology-based Approach to Support Text Mining and Information Retrieval in the Biological Domain. In Journal of Universal Computer Science (JUCS), Special Issue on Ontologies and their Applications, Vol. 13, No. 12, pp. 1881-1907, 2007.

GHOULA N. & KHELIF K. & DIENG-KUNTZ R. (2007), Supporting Patent Mining by using Ontology-based Semantic Annotations, In Proc. of Web Intelligence'07, USA, 2007.