Semantic Patent Clustering for Biomedical Communities

Khaled Khelif  
Aroua Hedhili  
INRIA Sophia Antipolis Méditerranée  
2004 route des Lucioles  
BP 93, 06902 Sophia Antipolis, FRANCE  
khaled.khelif,aroua.hedhili,martine.collard@sophia.inria.fr

Abstract

In this paper, we present the PatClust clustering solution for textual documents based on semantic criteria. Our proposition is dedicated to patent documents of the biomedical domain. We present three different approaches and we show that semantic web techniques clearly allow to improve the quality of resulting clusters.

1 Introduction

Patents hold important information on intellectual property of inventions. A number of authors have reported on methods for automated tools facilitating the job of patent officers. In this paper, we propose a solution for users of biomedical communities that are searching for information stored inside patents. This work is part of the Sealife project\(^1\) which goal is the design and development of a semantic web browser for Life Sciences communities of users. We enhance our previous approach PatAnnot[2] in which the systems allowed the automatic generation of semantic annotations on this kind of documents. We propose patent clustering techniques driven by semantic annotations. Clusters are computed on the basis of semantic annotations on patents in order to provide a valuable support for personalization of the semantic browser according to user interestingness. The paper is organized in four sections. Section 2 present related works on document clustering methods and patent analysis. Section 3 presents the PatClust approach: the semantic representation of claims and three methods of claims clustering. Section 4 presents our first experiments. We conclude in Section 5.

2 Related works

Text mining techniques apply data mining solutions to structured representation of unstructured textual resources which are generally large set of documents. By applying tokenization processes, a stream of words is identified in a given document \(d\) which is then represented by a term vector \((f(d, t_1),..., f(d, t_m))\) where \(t_1, ..., t_m\) are words identified into the whole corpus of documents and each \(f(d, t_i)\) is the frequency of the \(t_i\) word into the document \(d\). We ran hierarchical methods\[^9\] and bi-secting \(K\)-means\[^4\] that are known to give good results for text clustering \[^8\] and are implemented in the CLUTO clustering toolkit \(^2\) on which we led our experiments. Classical text clustering uses sets of words identified into documents by tokenization processes. They ignore the semantic dimension of words and their real importance inside the document. Recent studies have enhanced text clustering by introducing background knowledge and conceptual features.

Due to the rapid development of technology, patent document have been identified for several years as a challenging source of information for numerous domains. Public institutions like USPTO\(^3\) or EPO\(^4\) have already made considerable effort for facilitating patent processing and classification. Most works have been dedicated to patent offices and intellectual property companies which need to efficiently assign a new received patent to appropriate categories according to patent classification schemes like the United States Patent Classification (USPC). Since manual categorization of patents is too much complex and time consuming, researches focused on automatic classification. Patent documents are composed of multiple items such as abstracts, description of invention, claims... that are differently written and may contain more or less useful knowledge. First attempts to automatic classification focused on abstract and summaries sections, but recent works [6, 7]

\(^{1}\)http://www.biotec.tu-dresden.de/sealife/  
\(^{2}\)http://glaros.dtc.umn.edu/gkhome/views/cluto  
\(^{3}\)http://www.uspto.gov/  
\(^{4}\)http://www.epo.org/
have shown interesting results on exploring patent claims.

3  PatClust approach

The PatClust approach presented in this paper: (i) enhances the previously designed PatAnnot approach, (ii) takes advantage of semantic concepts from the biomedical domain and semantic web techniques for annotating texts, (iii) is defined to operate on claims only, (iv) extends the notion of similarity to semantic similarity.

3.1 Semantic Representation of Patents

PatAnnot: Patent Semantic Annotations:

PatAnnot[2] is an approach which consists of generating semantic annotations by structuring and building a semantic representation of patent documents. A generated annotation comprises three parts: a structure annotation, a metadata annotation and a domain based annotation. To structure these annotations, PatAnnot relies on a modular ontology called PatOnto describing the different aspects of the patent: the structure and the content. The textual content is annotated w.r.t a domain vocabulary using the tool MeatAnnot [3] which analyses scientific texts through linguistic tools and generates automatically RDF. For patent annotation, we coupled PatAnnot with the UMLS metathesaurus[5] which covers the biomedical vocabulary.

Claims analysis and annotation:

Patent claims is a set of sentences, following the description portion of the invention in a patent. The claims (i) define, in technical terms, the extent of the protection conferred by a patent, (ii) contain the most important information concerning the invention. For these reasons, we decided to rely on this section for patent clustering.

PatAnnot extracts automatically claims and it annotates them, the semantic annotation is then integrated in the whole annotation of the patent. For each term detected, we affected a weight proportional to its presence in the patent and in the whole of the corpus. This weight consists of the TFIDF[5] of the term. The use of TFIDF comes from the assumption that the importance of a term increases w.r.t the number of times the term appears in the document and it is offset by the frequency of the term in the corpus.

3.2 Ontology-based semantic clustering:

In this section, we present three approaches we experimented for semantic clustering of patents. In each of them: (i) we used the classical term vector representation for patent claims section with TF-IDF weights (ii) a patent claims section is considered as an unordered collection of instances of concepts in UMLS, and (iii) we used the same clustering algorithm bisecting k-means [4].

Standard approach:

In this approach we kept the simple TF-IDF vector for each patent claims section. Clusters were computed on the base of the standard similarity function cosine measuring the deviation of angles between the patents vectors. A cosine value of zero means that the two patents are orthogonal. In this approach the semantic touch is introduced by ontological concepts only, but we do not exploit relationships between them.

Hierarchical weight propagation approach:

In this approach, we introduced semantic concept relationship on weights. We assume for instance that: if a patent concerns microbiology with a weight n it also concerns biology with a weight which is lower to n. Therefore, we incremented the weight of concepts ancestors for each concept detected in the claims text. We divided the weight by 2 when passing from the concerned concept to its parent concepts. Then we spread a decreasing weight through the ontology hierarchy. The incremented weight for a given concept c which has a set of child concepts Chc is expressed as follows:

\[
weight(c) = \sum_{i \in Chc} \left( \frac{weight(i)}{2^{depth(c) - depth(i)}} \right)
\]

Where depth(x) is the depth of concept x in the ontology. To compute the similarity between patents we rely on the same function used in the standard approach.

Semantic distance approach:

In this approach, we introduced a semantic similarity function between patents without modifying basic TF-IDF weights. The idea is to use the conceptual distance (defined in [1]) between concepts annotating patents, this distance relies on the subsumption path in the UMLs metathesaurus. The semantic function is defined to reinforce the similarity between patent claims document which uses closed concepts. The length of a subsumption path between a concept c1 and one of its super concepts c2 in an inheritance hierarchy H is inductively defined by:

\[
\forall (c_1, c_2) \in H^2 : c_1 \leq c_2 : 
\]

\[
1_{H_\times}(c_1, c_2) = \sum_{\{c \in (c_1, c_2), c \neq c_1\}} \left[ \frac{1}{\text{depth}(c)} \right]
\]

(2)

We defined the semantic distance dist(c1, c2) between any two concepts of an inheritance hierarchy as the minimum of the sum of the lengths of the subsumption paths between each of them and a common super concept.
where $H_c$ is the hierarchy of concepts, $l_{H_c}$ the shortest path between $c_1$ and $c_2$, depth($c$) is the depth of the hierarchy. We relied on this distance to compute the semantic similarity between two concepts annotating patent claims. This similarity is defined by:

$$semSim(c_1, c_2) = \frac{1}{1 + dist(c_1, c_2)}$$

(3)

We then defined the semantic similarity $semSimPatent$ between two patents as the average weighted semantic similarity between their concepts 4.

$$semSimPatent(p_i, p_j) = \frac{1}{N_c} \sum_{c \in H_c} \sum_{p_i \in C(c)} \sum_{p_j \in C(c)} w_{ij} \cdot semSim(c_i, c_j)$$

(4)

where $w_{ij}$ is the weight of the concept $c_i$ in the patent $i$.

4 Experiments and Interpretation of resulting clusters

In order to validate and test the three proposed approaches, we constituted a patent corpus related to the biomedical domain. We downloaded this corpus (about 5000 patents) from the USPTO website and we chose 390 patents related to infectious diseases which is the principle use case of the Sealife project.

After this, we annotated the whole of the corpus using PatAnnot w.r.t the UMLS metathesaurus: (i) generation of annotations about metadata, and (ii) extraction of concept instances from claims text. Finally, we defined similarity measures to evaluate the quality and the consistency of resulting clusters (see section 4.1).

The three approaches are implemented using the framework CLUTO which offers a toolkit for data clustering and cluster analysis programs. CLUTO allows to compute clusters either from the term vector space or from a similarity matrix. For experimentation on approaches 1 and 2, the clustering algorithm operated on the vectorial space and for experimentation on approach 3, it operated on the similarity matrix.

4.1 Semantic similarity among clusters

CLUTO provides tools for evaluating resulting clusters and visualizing them. Clusters are evaluated by default according their internal similarity and their external similarity. Each computed cluster is described by its descriptive features with the percentage of the within cluster similarity that the feature can explain. In PatClust, since these features are concepts from an ontology, we were able to define semantic internal and external similarities too. Semantic similarities are computed on the basis of descriptive concepts of clusters. These descriptive concepts consist of a set of concepts that tend to to occur together in the patents of the same cluster.

Internal similarity (5): it is the average similarity between the documents (patents) of each cluster. It is expected to be maximized.

$$Isim(C_i) = \frac{1}{n_i} \sum_{p_i, p_k \in C_i} sim(p_i, p_k)$$

(5)

External similarity (6): it is the similarity of the documents (patents) of each cluster and the other documents. It is expected to be minimized.

$$Esim(C_i) = \frac{1}{\sqrt{\sum_{p_i, p_k \in C_i} sim(p_i, p_k)}}$$

(6)

Semantic internal similarity (7): it is the average semantic similarity between the most descriptive concepts of each cluster. It is expected to be maximized.

$$ISemsim(C_i) = \frac{1}{N_c} \sum_{k=1}^{N_c} \sum_{l=1}^{N_c} semSim(c_{ik}, c_{il})$$

(7)

Semantic external similarity (8): it is the average semantic similarity between the most descriptive concepts of a cluster and the most descriptive concepts of the other clusters. It is expected to be minimized.

$$ESemsim(C_i) = \max\{\frac{1}{N_c^2} \sum_{k=1}^{N_c} \sum_{l=1}^{N_c} semSim(c_{ik}, c_{il})\}$$

(8)

The notation in these equations are as follows: $C_i$ is the $i^{th}$ cluster, $n_i$ is the number of patents in the cluster $C_i$, $N_c$ is the total number of clusters, $S$ is the total patents to be clustered, $c_{ij}$ is the $j^{th}$ descriptive concept of the cluster $C_i$, $N$ is the number of descriptive concepts of the cluster, $p$ is a patent, $sim(p_i, p_j)$ is the similarity between two patents (i.e. cosine or $semSimPatent$) and $semSim(c_i, c_j)$ is the semantic similarity between two concepts.

4.2 Results and discussion

Table 1 shows standard similarities and semantic similarities we computed for resulting clusters. Only standard similarities were computed for resulting clusters in the third approach. Indeed in this semantic distance approach, clusters are computed on the basis of the semantic similarity matrix and they are not defined in terms of concepts.

We can observe either on Figure 1 or into Table 1 that:

(i) The hierarchical weight propagation approach allows to globally improve both standard and semantic internal similarities by comparison with the standard approach,
(ii) The hierarchical weight propagation approach provides bad results on external similarities by comparison with the standard approach. (iii) The semantic distance approach allows to globally improve both internal and external similarities by comparison with the standard approach.

<table>
<thead>
<tr>
<th>Standard approach</th>
<th>$C_{11}$</th>
<th>$C_{21}$</th>
<th>$C_{31}$</th>
<th>$C_{41}$</th>
<th>$C_{51}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$I_{sim}$</td>
<td>0.223</td>
<td>0.219</td>
<td>0.176</td>
<td>0.134</td>
<td>0.129</td>
</tr>
<tr>
<td>$E_{sim}$</td>
<td>0.007</td>
<td>0.02</td>
<td>0.017</td>
<td>0.017</td>
<td>0.014</td>
</tr>
<tr>
<td>$I_{Semsim}$</td>
<td>0.05</td>
<td>0.019</td>
<td>0.83</td>
<td>0.023</td>
<td>0.015</td>
</tr>
<tr>
<td>$E_{Semsim}$</td>
<td>0.147</td>
<td>0.194</td>
<td>0.353</td>
<td>0.414</td>
<td>0.441</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Hierarchical weight propagation approach</th>
<th>$C_{12}$</th>
<th>$C_{22}$</th>
<th>$C_{32}$</th>
<th>$C_{42}$</th>
<th>$C_{52}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$I_{sim}$</td>
<td>0.96</td>
<td>0.663</td>
<td>0.876</td>
<td>0.678</td>
<td>0.808</td>
</tr>
<tr>
<td>$E_{sim}$</td>
<td>0.448</td>
<td>0.193</td>
<td>0.712</td>
<td>0.546</td>
<td>0.696</td>
</tr>
<tr>
<td>$I_{Semsim}$</td>
<td>0.075</td>
<td>0.225</td>
<td>0.22</td>
<td>0.106</td>
<td>0.128</td>
</tr>
<tr>
<td>$E_{Semsim}$</td>
<td>0.179</td>
<td>0.207</td>
<td>0.415</td>
<td>0.631</td>
<td>0.869</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Semantic distance approach</th>
<th>$C_{13}$</th>
<th>$C_{23}$</th>
<th>$C_{33}$</th>
<th>$C_{43}$</th>
<th>$C_{53}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$I_{sim}$</td>
<td>0.373</td>
<td>0.16</td>
<td>0.206</td>
<td>0.236</td>
<td>0.356</td>
</tr>
<tr>
<td>$E_{sim}$</td>
<td>0.017</td>
<td>0.016</td>
<td>0.016</td>
<td>0.014</td>
<td>0.013</td>
</tr>
</tbody>
</table>

Table 1. Results of the different approaches

![Figure 1. Patents distribution over clusters](image)

We can observe too that the internal similarities of clusters obtained by the hierarchical weight propagation approach are higher than those provided by the third approach. Indeed the weight propagation tends to reinforce the similarity between two clusters which descriptive concepts share common ancestor concepts. The 3D mountain representations shown in Figure 1 are provided by CLUTO. This representation allows to estimate the relative similarity of clusters as well as their size, internal and external similarity. Each cluster is represented as a peak. The height of each peak is proportional to the cluster’s internal similarity. The volume of a peak is proportional to the number of elements contained within the cluster. This figure confirms our previous observations and demonstrates that the third approach using semantic distance, provide more uniformly distributed clusters.

5 Conclusions and Future Work

In this paper we presented three approaches for semantic patent clustering for biomedical communities. We assumed that semantics can improve text clustering which is confirmed by the obtained results. The approaches rely on standard semantic web technologies (RDF, SPARQL, etc.) and they are generic and domain-independent. As further work, we envisage to: (i) Use descriptive concepts for automatic cluster characterization (ii) Propose a predictive model for each approach (iii) Perform an expert-centered evaluation to validate the coherency of our clusters (iv) Link clustering techniques with profile detection techniques in order to propose personalized recommendations to the Sealife browser users (v) Test the approaches on other scientific corpus such as scientific papers.

Acknowledgements: Funding by the Sealife project (IST-2006-027269) is kindly acknowledged.

References


