# Towards Automatic Building of Term Hierarchies from Large Patent Datasets

Maria Lucía Castro Jorge†, Roque Enrique López Condori†*, Gabriel Dieterich Cavalcante†,
Luiz Daniel Couto de Barros Lapolla†
†Elabora Consultoria, Campinas, Brazil
*San Agustin National University, Arequipa, Perú
{lucia, roque.condori, gabriel, lapolla}@elabsis.com

***Abstract*—Term Hierarchies are structures that represent semantic relations among terms, usually of the type hyperonym and hyponym (generality and specificity, respectively). There are many scenarios that may benefit from the knowledge within Term Hierarchies. Particularly, in the patents genre there is an important demand of knowledge representation for Information Retrieval purposes, and few works have approached this demand. In this work we proposed a three stage strategy for term hierarchy building from patents. In the first stage, terms were extracted through noun-phrase identification; in the second stage terms were organized hierarchically through n-gram decomposition; finally, in the third stage, the term hierarchy was enriched with term embeddings information, particularly from Word2Vec model. This strategy was applied over patents from the United States Patent and Trademark Office (USPTO) collection. Each term in the hierarchy generated a set of associated patents. For the evaluation task we applied two strategies over the sets of associated documents, one based on the clustering degree of the sets, and the other one based on the IPC (International Patent Classification) categories proportions within the sets. Results show that the produced term hierarchy efficiently captures generic and specific concepts.**

## I. Introduction

The increasing growth in the volume of online textual information has motivated the development diverse techniques for textual knowledge representation. In the last years, various approaches have been proposed in order to discover and organize knowledge, such as: semantic networks, ontologies, term hierarchies, among others.

Knowledge representation using Term Hierarchies is commonly used in Information Retrieval for organization and exploration of textual sets through terms that describe the main topics within them. An important characteristic of Term Hierarchies is that terms should be organized in levels, reflecting generality and specificity among them.

There are many scenarios that could benefit from Term Hierarchies. For patents' collections, there is an important demand of knowledge representation, since these textual domain contain novel research solutions that are relevant for analysis in the industry, business and law communities.

Patents have a particular writing style, characterized by technical vocabulary with an unusual distribution, since this vocabulary tends to be repeated in order to reinforce concepts and to avoid plagiarism. In order to project an appropriate strategy to build term hierarchies from patents, it is important to develop adequate techniques to extract relevant terms and to organize them in a hierarchical structure. Such techniques require an analysis of patents at a lexical, syntactic and semantic level, in order to achieve good results.

In this work we proposed an strategy for automatic Term Hierarchy building, from patents' sets, particularly from the United States Patent and Trademark Office (USPTO) collection. The term extraction process was approached using a noun phrase based criteria applied within the title, abstract and claims, since these sections tend to contain the main technical terms of the patent content [1]. The hierarchical organization of terms, reflecting generic and specific concepts, was performed using a document-frequency criteria related to the sets of terms' associated documents, and the n-gram cardinality of the terms. Additionally, we used word embeddings representations to enrich the hierarchy, preserving the criteria for generality and specificity. This approach was projected on a distributed environment, where the set of patents was divided for parallel processing.

For the evaluation task, we proposed two methodologies. The first one was based on the agglomeration degree within the document clusters associated to each term in the hierarchy. Particularly, we used Clustering Coefficient measure [2] for the calculation of intra-cluster agglomeration. The underlying assumption for this type of evaluation is that, more generic terms will tend to be associated to document sets with lower intra-cluster agglomeration degree, and vice versa. The results obtained with this methodology showed that the hierarchies efficiently captured generic and specific terms. The second methodology was based on the observation of the IPC categories proportions within document sets associated to the terms in the hierarchy. The underlying assumption for this criteria is that, document sets associated to more generic terms will manifest lower proportions for a higher number of IPC categories (meaning that these terms will appear in more categories of the IPC), while document sets associated to more specific terms will manifest higher proportions for a fewer number of IPC categories (meaning that these terms will appear more frequently in less categories of the IPC). Results followed this assumption.

The remaining of this paper is organized as follows: in Section II, we briefly introduced some main related works; the proposed method is described in Section III; in Section IV, we describe the data set and the results reported in the experiments; finally, in Section V, we present some final remarks.

ISBN 978-952-68397-8-3 (PDF)

## II. Related works

For patent domain, most of the works related to term extraction have indentified relevant terms by considering noun phrases (NP) as term candidates. [3] identify terms using the NP present in all the sections of patents (title, abstract, claims, description, etc.) in English-Chinese comparable patents. To filter out the candidates terms, the authors only consider NP with less than 5 words in its structure trying to avoid possibles parsing errors. [4] uses a similar strategy filtering out shorter candidates that were present in longer candidates (e.g. "machine learning" could be filtered because of "supervised machine learning"). [5] proposed a method based on decision trees with boosting and bagging techniques to classify term candidates (NPs extracted from patent's text). The machine learning algorithm used three set of features: (i) a set of structural features characterizing the position of a term within the document structure, whether it is present in the title, abstract, introduction, etc., (ii) a set of content features which captured distributional properties of a term in relation to the overall textual content, e.g. TF-IDF a, and (iii) a set of lexical/semantic features which were produced using some terminological databases (e.g. Wikipedia). [6] used similar features to train a maximum entropy classifier. [7] considered only NP that co-occur in the abstract and first claim sections. The authors indicated that these sections described properly the invention of the patent and, thus, they contained valuable terms for analysis. Similarly, [8] only used the title section to extract terms for patent analysis. Recently, [9] proposed a graph based ranking algorithm that used vector representations of terms to improve the precision in the extraction of top-k terms. [10] presented a method to translate Chinese-Japanese patents, based on pre-building a parallel training corpus with bilingual terms using linguistic and statistical techniques.

Unfortunately, there are few approaches for automatically building Term Hierarchies from patents. [11] was one of the firsts works in this line. To create the hierarchy of terms, the authors identified hyponyms and hypernyms relations among terms using regular expressions from common writing style patterns of patents. For instance, in "A chair comprising a lag and a back", the word "comprising" which occurs very frequently in patents would be considered useful to identify hyponyms and hypernyms. [12] presented a tool to assist experts for building term-ontologies. In a use case for patents, this tool extracted terms of the title, abstract and claims sections and created a hierarchy of terms using some heuristics such as document frequency (generic term are more frequent) and term composition (e.g. "mature tree" is more generic than "mature avocado tree"). [13] use Wordnet to define relationships among frequent terms in order to organize and group them in an ontology.

## III. Proposal

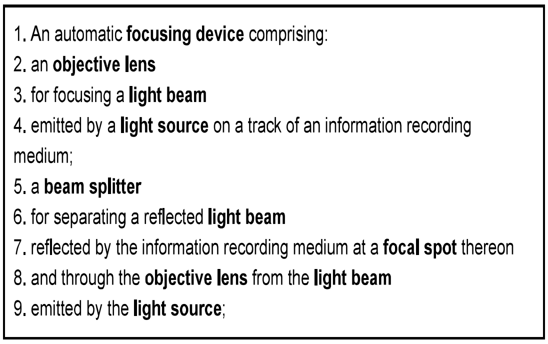
For the Term Hierarchy building, we projected a three-stage strategy: term extraction, hierarchy building and hierarchy enrichment. In the following sub-sections these three stages are described in detail.

### A. Term Extraction

Document keywords are terms or concepts that represent the main topics within a document. Keywords are usually manually annotated, however, in voluminous document sets there is a considerable number of documents which do not have associated keywords. Performing keywords human annotation in this scenario would be expensive or even impossible.

In this work, the term extraction task was approached based on the assumption that keywords (most relevant terms) in a patent are located within the title, abstract and claims sections, based on previous researches such as [1]. In order to extract term candidates (noun phrases) appropriately, we first segmented the title, abstract and claims. The title and abstract were segmented by sentences, while claims were segmented using [1] segmentation criteria. According to Ferraro, claims' segments were delimited by punctuation signs and other common markers in patents, which are terms such as: comprising, characterized, thereon, thereby, wherein, whereby, for, by, etc.

In Fig. 1 is shown an example extracted from the work of Ferraro, where it is illustrated how segmentation is performed.



1. An automatic **focusing device** comprising:
2. an **objective lens**
3. for focusing a **light beam**
4. emitted by a **light source** on a track of an information recording medium;
5. a **beam splitter**
6. for separating a reflected **light beam**
7. reflected by the information recording medium at a **focal spot** thereon
8. and through the **objective lens** from the **light beam**
9. emitted by the **light source**;

Fig. 1. Example of claims segmentation

It is observed from the figure above that NPs (in bold) occur within the identified segments.

Since the number of obtained NPs might be high, an important goal was to identify NPs that were more relevant to the patent. In other words, select from all the identified NPs the ones that best represent the main topics of the patent. An initial strategy would be select NPs with higher frequency throughout the title, abstract and claims. Another approach would be to count the number of sections (title, abstract, claims) in which the NP appears. Thus, NPs that appear in more sections should be considered more important.

Initially, the two approaches were implemented. However, there were some significant problems with the extracted terms. Terms with high specificity and terms which did not accurately describe the main content achieved high values. It was also observed that the terms that better described the main document topic were located commonly in the title, abstract and first claim of the patent. This was also supported by [7] and [8]. For this reason, we decided to consider as base terms all the NPs that occurred within these sections: title, abstract and first claim. After this selection, the NPs within the rest of the claims were selected if they were a lexical variation of any of the base terms. The purpose of this criteria was to ensured that only NPs related to the main topics were considered as terms.

### *B. Hierarchy Building*

A Term Hierarchy should reflect generality and specificity among the main topics in a set of documents. In this work, we assumed a term is more generic if its cardinality (number of lexicons composing the term) tends to be lower and the number of documents the term covers tends to be high. On the other hand, a term would be more specific if its cardinality tends to be higher and the number of documents the term covers tends to be lower. This assumption was inspired in the theory of Formal Concept Analysis [14], where each concept in the hierarchy represents the set of objects sharing the same properties and each sub-concept in the hierarchy is described by more properties and contains a subset of the objects in the concepts above it.

In this work, the Term Hierarchy building was performed using the terms extracted in the previous phase. To reflect generic and specific topics, initially, terms were classified into three types: unigrams, bigrams and trigrams. The idea was to hierarchically organize terms according to the assumption described above. Thus, at the top of the hierarchy should be located unigrams and, gradually, in subsequent levels, bigrams and trigrams. Terms that were composed by more than three words were considered very specific topics and, therefore, excluded from the hierarchy, at this point. Since the NPs are usually represented by bigrams an trigramas, some unigrams were produced by splitting the bigrams and trigrams (e.g., since “metal oxide”, it was created two unigram terms “metal” and “oxide”). The goal for doing this splitting, was to expose generic concepts within the textual set (unigrams) that may have not be captured in the term extraction phase, due to NP identification strategy.

After unigrams, bigrams and trigrams were identified, the next step was to generate the hierarchical linking among them. This process was performed under the following criteria : (i) a bigram was connected to a unigram through a child relation if one of its words was the unigram; (ii) a trigram was connected to a bigram through a child relation if two elements (words) of the trigram were contained in the bigram term; (iii) if a trigram could not be linked to any bigram, then it was linked to the adequate unigram according to a criteria similar to (i); (iv) if a bigram or trigram could not be connected to any parent, then it was assumed a generic concept, and so, located at the top of the hierarchy.

### *C. Hierarchy Enrichment*

Although the hierarchy of terms allows to establish hierarchical levels, it may not guarantee an adequate coverage of the topics present in patents. Thus, in order to enrich the hierarchy and consequently improve the coverage, it was decided to use additional knowledge to enrich the hierarchy. This knowledge was given by the Word2Vec model [15].

Recently, [15] proposed the Word2Vec as a model for word embeddings representations, where words are represented as a vector of real values and not as discrete atomic symbols (which occur in traditional word representations, such as the Bag of Words model). Word2Vec is a two-layer neural network that is trained to learn the semantic contexts of words. For instance, if in documents are used the words “home” and “house” to express the same idea, but they not co-occur in the same sentences, Word2Vec model can learn its context and map them in close points in an n-dimensional space. Through vector arithmetic operations over this n-dimensional space, is possible to performer some analyses in the texts, such as searching of similar words, analogy detection and others.

The hierarchy enrichment could be performed through any other method or resource that models semantics among words, such as onthologies (e.g. Wordnet [16]), or statistical models like LDA [17], LSA [18] or other word embeddings. Though this variety of methods and resources, Word2Vec shows many benefits that these other methods may not have. For instance, Word2Vec relies in statistics and not in language dependant resources. Also, it provides statistics among terms, and not only statistic among terms and latent topics (as LDA does), which can lead to more specific knowledge among terms. Futhermore, in recent works ([19], [20], [21]), Word2Vec model had shown promising results in a variety of NLP tasks in comparison with other approaches.

Word2Vec can produce word embeddings vectors through two architectures: Continuous Bag Of Words (CBOW) or Skip Gram. CBOW predicts the target word from the contextual words that surround it, and the Skip Gram architecture predicts the contextual words given the target word. We performed experiments with both architectures, but the reported results in this paper correspond to the CBOW architecture because it achieved the best performance.

With the integration of Word2Vec model, we tried to cover more information at each level of the hierarchy. Thus, for each term of the hierarchy were associated new terms that shared the same semantic context according to the Word2Vec model. In general, the complete integration of Word2Vec model to the hierarchy of terms was carried out in two steps: building of word embeddings and searching of similar terms.

*1) Building of Word Embeddings:* This step aims to create mathematical models for each IPC (International Patent Classification) maingroup (a maingroup is a category in the hierarchical patent classification of IPC), i.e. create vector representations of words present in the patents of each maingroup. In this process, instead of using words, we used noun phrases (NP) extracted from title, abstract and claims of the patent. The choice of NPs was due to two reasons: (i) NPs provide more information than individual words and (ii) processing time for creating word embeddings is considerably less when not all words of patents are used.

Unlike term extraction phase, in the building of word vectors we considered all NPs, in order to complement the information provided by the terms of the hierarchy. To create word vectors (in our case NP vectors), a training process is needed. In this process, Word2Vec model learns the semantic context of each NP analyzed. It is important to indicate that Word2Vec model does not need labeled data for training, which is an important advantage, since most of the data of the patents used in this work does not contain labels.

The training of each maingroup was performed in a distributed environment using Spark framework and Gensim library. The word embeddings for each maingroup were produced using the CBOW architecture of Word2Vec model and it were used in the searching of similar terms.

*2) Searching of Similar Terms:* This step aims to associate to each level of the term hierarchy a set of new terms that can complement the information present in each level. This new set of terms was obtained using the word embeddings produced in the previous step.

In general, for each term of the hierarchy were associated 5 new terms, i.e., terms that did not have relations with terms within the hierarchy. These new terms were ones that share (with high probability) the same semantic context with the original terms (unigrams, bigrams or trigrams). Experiments were performed using 5, 10 and 15 new terms, but the best results observed in experiments were with 5 terms, thus, in subsequent experiments, we used this parameter.

To select new terms, we used the word vectors created in the previous step. The search of new terms was performed with the function $most\_similar$ of Gensim package.

With new terms associated to each term of the hierarchy, the next step consists in identifying the generality and specificity of the new terms with respect to the original terms within the hierarchy. For instance, given the term "computer" and its new associated terms "technology", "touchscreen" and "mouse", it must be determined whether "computer" is more specific or general than "technology", "touchscreen" and "mouse".

To identify the generality and specificity between the terms, we used the criterion of term frequency, which has been used in various works of literature ([22], [23], [12], [24], etc.). This criterion indicates that uncommon terms are considered more specific and frequent terms more general. Thus, we calculated the term frequency in patents for the terms of the hierarchy and their new associated terms. If a term of the hierarchy was more frequent than a new term, it was considered as its parent within the hierarchy, in the other hand, it was considered as its child. This procedure was performed for each new associated term.

## IV. Experiments, evaluation and results

### A. Data and experimental methodology

The goal of the experimental task was to produce Term Hierarchies with the methodology proposed in this paper, in order to evaluate the capability of this hierarchies to properly capture the notions of generality and specificity. For this aim, we used a collection of 1987208 patents obtained from the United States Patent and Trademark Office (USPTO). This collection was divided into various groups, with the purpose of projecting a parallel processing schema. Each group division was generated based on the International Patent Classification (IPC) structure. The IPC provides a hierarchical semantic structure of categories for patents organization, which represent traditional areas of innovation, e.g. Physics, Biology, Medicine, etc. Particularly, we used the maingroup category of the IPC. For each maingroup of patents it was build a hierarchy of terms according to the strategy proposed in this paper.

### B. Evaluation methodology

As mentioned before, the main goal of the evaluation task is to measure the degree in which the term hierarchies capture generality and specificity (hyperonym and hyponym). In order to achieve this goal, two evaluation criteria were proposed in this work, (i) one guided by IPC structural information (IPC-guided) and (ii) the other one based on Clustering Coefficient measure. The IPC-guided evaluation relies on the assumption that documents that are associated to more generic terms will tend to occur in smaller proportions in many IPC categories (e.g. sections), while documents related to more specific terms will tend to occur in higher proportions in few IPC categories. The Clustering Coefficient measure relies on the assumption that more specific terms tend to cluster more similar documents, according to a similarity measure. In the following subsections, both approaches are explained in detail.

*1) IPC-Guided Evaluation:* For the IPC-guided evaluation, we considered two categories of the IPC in order to perform the evaluation: sections and subgroups. The aim for choosing these two categories is to observe the proportions in which more generic categories (sections) and more specific categories (subgroups) occur within the document cluster associated to a particular term in the hierarchy. It is important to highlight that each patent is already labeled under a set of ICP categories (sections, classes, groups, subgroups, etc.).

Two approaches were performed for this type of evaluation. Under the first approach, we calculated the proportion of IPC categories within a cluster in relation to all categories of the IPC, this proportion was defined as the ratio of the number of unique sections and subgroups within a cluster of documents $c_i$, and the total number of unique sections and subgroups in the IPC, respectively. This is formalized in equations (1) and (2).

$$Sec(c_i) = \frac{UniqueSections(c_i)}{TotalUniqueSectionsIPC} \tag{1}$$

$$Sub(c_i) = \frac{UniqueSubgroups(c_i)}{TotalUniqueSubgroupsIPC} \tag{2}$$

In the second approach we calculated the proportion of each IPC category within a cluster of documents $c_i$ in relation to the other IPC categories within the same cluster of documents. This proportion was defined as the ratio of the number of documents classified under a given section $s_j$ and subgroup $sg_j$ (respectively), and the total number of documents within the cluster. This is formalized in equations (3) and (4)

$$Sec(c_i, s_j) = \frac{NumberDocuments(s_j)}{TotalDocuments(c_i)} \tag{3}$$

$$Sub(c_i, sg_j) = \frac{NumberDocuments(sg_j)}{TotalDocuments(c_i)} \tag{4}$$

*2) Clustering Coefficient Measure:* In traditional clustering/categorization scenarios the intra-cluster evaluation task is performed using class labels. In this case, for the generality-specificity evaluation goals, we proposed two strategies that are non-dependent of class labels, by using the Weighted Global Clustering Coefficient measure (WCC) [2], which is a traditional graph measure that provides the degree in which the nodes of a graph tend to agglomerate together. In our scenario, each cluster of documents grouped under a term in the hierarchy can be seen as a graph, where nodes represent the documents and the edges represent a dissimilarity degree between each pair of documents. In order to apply this, we used the Python package NetworkX, where WCC is defined according to (5), (6) and (7).

$$WCC = \frac{1}{n} \sum_{u \in G} C_u \tag{5}$$

where:

$$C_u = \frac{1}{deg(u)(deg(u)-1)} \sum_{uv} (\hat{w}_{uv}\hat{w}_{uw}\hat{w}_{vw})^{\frac{1}{3}} \tag{6}$$

and:

$$\hat{w}_{uv} = \frac{w_{uv}}{max(w)} \tag{7}$$

Equation (5) describes the global WCC, which is given by the sum of the WCC of each node of the graph. Equation (6) describes the WCC of a single node which is defined as the geometric average of the subgraph edge weights ($\hat{w}$), which in turn are normalized by the maximum weight in the network (Equation (7)).

Particularly, edges' weights in the graph were given by dissimilarity measure proposed by [25], called Word Mover's Distance (WMD), where the distance between two documents is measured as the minimum distance of their words in word embedding space using the Euclidean distance among them. We chose WMD due to its outperforming performance in comparison to traditional dissimilarity measures.

The underlying assumption in our evaluation methodology is that more generic terms will form clusters of documents with lower values of WCC, since documents within the cluster will tend to have higher dissimilarity among them. As terms in the hierarchy are more specific, the WCC value will tend to increase.

### C. Experiments and results

In order to apply IPC-guided and WCC evaluations, we randomly selected 10 paths of the hierarchy, corresponding to the IPC maingroups: "H01J17" and "G06N3". Each path is formed by an unigram and its corresponding bigram and trigram descendants and its associated new terms calculated with Word2Vec model. Each unigram, bigram, trigram and Word2Vec term in a path is associated to a group/cluster of patents. The patents associated to a given term in the hierarchy are the ones that have the term within the title, abstract or claims sections.

In Fig. 2 we illustrate these hierarchy paths, where terms located more to the left are considered more generic and terms located more to the right are considered more specific. The edges in dashed lines indicate that the node child was calculated using Word2Vec model.

In Table I, we show the results of WCC evaluation. In the first column there are shown the maingroups and their terms corresponding to the paths illustrated in Fig. 2. In the second column of the table it is shown the WCC value for the document clusters associated to each term and the Median Absolute Deviation (MAD). MAD was used to measure the variability of results, reducing the impact of outliers.

It may be observed from Table I, that for most of the paths of at least a parent and a child, the results satisfy the assumption for generality and specificity described previously. Also the MAD value shows that results are quite stable.

TABLE I. RESULTS FOR WCC EVALUATION WITH WORD MOVER'S DISTANCE

| Terms | | WCC | | WMD | |
|---|---|---|---|---|---|
| | | Value | MAD | Value | MAD |
| H01J17 | metal | 0.690 | 0.018 | 1.634 | 0.147 |
| | metal halide | 0.740 | 0.027 | 1.440 | 0.150 |
| | high-intensity discharge lamp | 0.991 | 0.000 | 1.395 | 0.011 |
| | metal halide lamp | 0.736 | 0.022 | 1.418 | 0.144 |
| | discharge chamber | 0.765 | 0.014 | 1.413 | 0.130 |
| | phosphor | 0.680 | 0.027 | 1.557 | 0.169 |
| | phosphor layer | 0.694 | 0.022 | 1.513 | 0.162 |
| | side thicknes | 0.761 | 0.000 | 0.705 | 0.072 |
| | green phosphor layer | 0.726 | 0.029 | 1.400 | 0.136 |
| | blue phosphor layer | 0.767 | 0.026 | 1.471 | 0.129 |
| | substrate | 0.662 | 0.022 | 1.529 | 0.154 |
| | front substrate | 0.683 | 0.022 | 1.427 | 0.136 |
| | front glass substrate | 0.745 | 0.020 | 1.367 | 0.117 |
| | dielectric | 0.661 | 0.019 | 1.521 | 0.134 |
| | dielectric layer | 0.649 | 0.017 | 1.495 | 0.127 |
| | dielectric layer anda | 0.629 | 0.039 | 1.078 | 0.332 |
| G06N3 | neural | 0.681 | 0.022 | 1.396 | 0.115 |
| | neural network | 0.709 | 0.021 | 1.323 | 0.106 |
| | biological sample | 0.839 | 0.000 | 0.559 | 0.013 |
| | artificial neural network | 0.794 | 0.021 | 1.340 | 0.112 |
| | probability distribution relationship | 0.839 | 0.000 | 0.559 | 0.013 |
| | experimental data | 0.718 | 0.018 | 0.789 | 0.198 |

It might seem obvious for the paths composed by unigrams, bigrams and trigrams that results would meet the assumption, since they are compositional terms and, at each descendant level of the hierarchy, they will necessarily cover less documents that would lead to a higher WCC value. But also, the hierarchical relations produced through Word2Vec enrichment meet the assumption in most of the paths (e.g. metal halide → high-intensity discharge lamp), showing that our strategy for identifying generality and specificity is promising.

Additionally, in Table I is showed the results for Word Mover's Distance (WMD) and its MAD values. The WMD-value column indicates the average WMD distance values among all documents associated to the correspondent term. As it can seen from Table I, the WMD values for more generic terms tend to be higher than the WMD values for specific terms. These values reflect that the hierarchical organization of the terms (n-grams and Word2Vec terms) are promising at capturing notions of generality and specificity.

In Table II, we show the results for the IPC-guided evaluation. The first column of the table corresponds to the hierarchy terms from the paths illustrated in Fig. 2. The second and third columns correspond to the evaluation results according to the first approach of the IPC-guided evaluation, calculated by equations (1) and (2).

It can be seen from the results above that generic terms (e.g. metal, phosphor, substrate) tend to manifest higher values for subgroups and sections, while more specific terms tend to manifest lower values for the same calculation. For instance, let's observe IPC evaluation (subgroup and section) for "metal halide" with respect to "metal", or "high-intensity discharge lamp" with respect to "metal halide", which manifest lower values as the specificity of the term increases according to

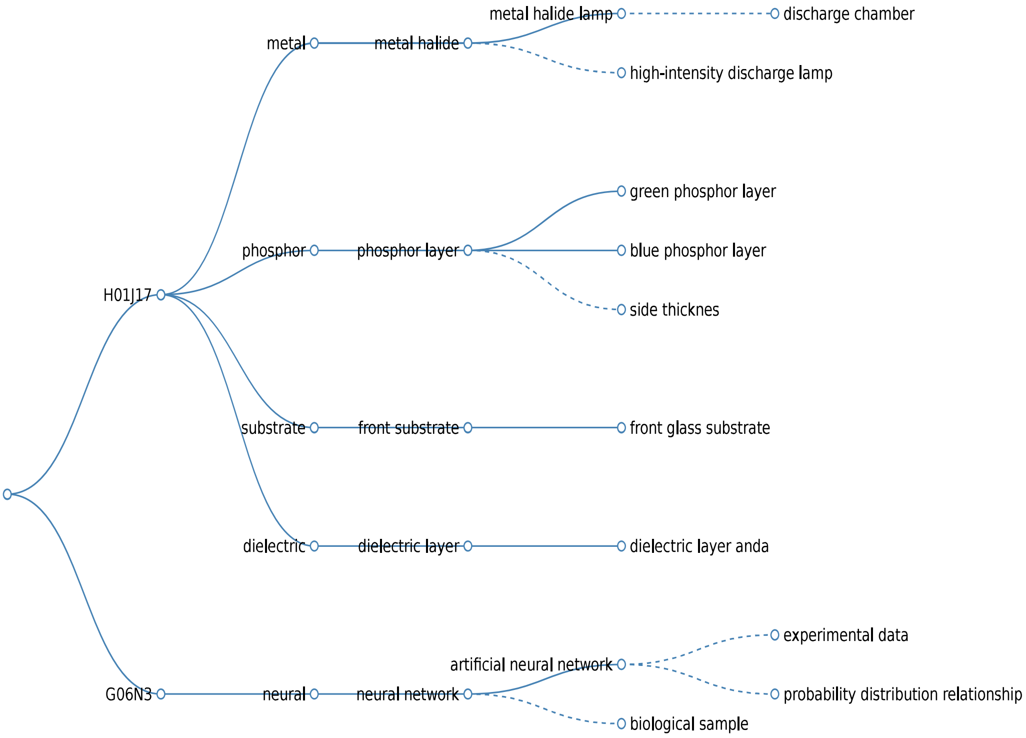


Fig. 2. Example of hierarchy of terms

TABLE II. RESULTS FOR IPC-GUIDED EVALUATION FOR SUBGROUPS AND SECTIONS

| Terms | | Subgroup | Section |
|---|---|---|---|
| H01J17 | metal | 0.382 | 0.375 |
| | metal halide | 0.118 | 0.125 |
| | high-intensity discharge lamp | 0.088 | 0.125 |
| | metal halide lamp | 0.147 | 0.125 |
| | discharge chamber | 0.265 | 0.125 |
| | phosphor | 0.324 | 0.375 |
| | phosphor layer | 0.147 | 0.250 |
| | side thicknes | 0.029 | 0.125 |
| | green phosphor layer | 0.029 | 0.250 |
| | blue phosphor layer | 0.029 | 0.250 |
| | substrate | 0.441 | 0.500 |
| | front substrate | 0.235 | 0.250 |
| | front glass substrate | 0.088 | 0.250 |
| | dielectric | 0.324 | 0.375 |
| | dielectric layer | 0.206 | 0.205 |
| | dielectric layer anda | 0.029 | 0.125 |
| G06N3 | neural | 1.000 | 0.375 |
| | neural network | 0.889 | 0.250 |
| | biological sample | 0.111 | 0.125 |
| | artificial neural network | 0.778 | 0.250 |
| | probability distribution relationship | 0.111 | 0.125 |
| | experimental data | 0.222 | 0.125 |

the hierarchy (see Fig. 2). This behavior may arise due to the assumption that more generic terms occur in more categories of the IPC, while more specific terms tend to occur in a lower number of IPC categories.

In Fig. 3, we illustrate the results for the second approach of the IPC-guided evaluation. It is important to highlight that, due to space limitations, we only show a sample corresponding to subgroups of the maingroup “G06N3”, though the evaluation task was performed in a complete IPC scenario.

We can see in the results that, given by the calculus of equation (4), more generic terms (e.g. “neural”) tend to have lower values throughout various subgroups, while more specific terms tend to have higher values in a few number of subgroups (e.g. “biological sample”).

Despite our evaluation methods (IPC-guided and WCC) may not be highly precise and so results may not be conclusive, all of them indicate that our strategy for Term Hierarchy building shows promising at capturing the notions of generality and specificity.

## V. FINAL REMARKS

In this paper we proposed an automatic strategy for Term Hierarchy building for patents’ genre. The strategy consisted of three stages: (i) term extraction through NPs identification, (ii) term hierarchical organization through term cardinality, and (iii) hierarchy enrichment through Word2Vec model. Evaluation results reveal that our strategy has a promising performance at identifying generic and specific terms for patents.

This work presents a significant contribution since few investigations have approached the patent scenario for automatically extracting terms and subsequently building a hierarchy of the extracted terms.

Besides this, it is important to mention some of the limitations faced throughout this work. One highlighting limitation was the absence of a human evaluation of term hierarchies. Another limitation is the few availability of linguistic studies on the patent genre, which limits the creation of adequate strategies for term extraction and hierarchy building. In future works these limitations may be addressed. Additionally, we plan to evaluate some of the most relevant state of the art works using our proposed evaluation methods, in order to provide a

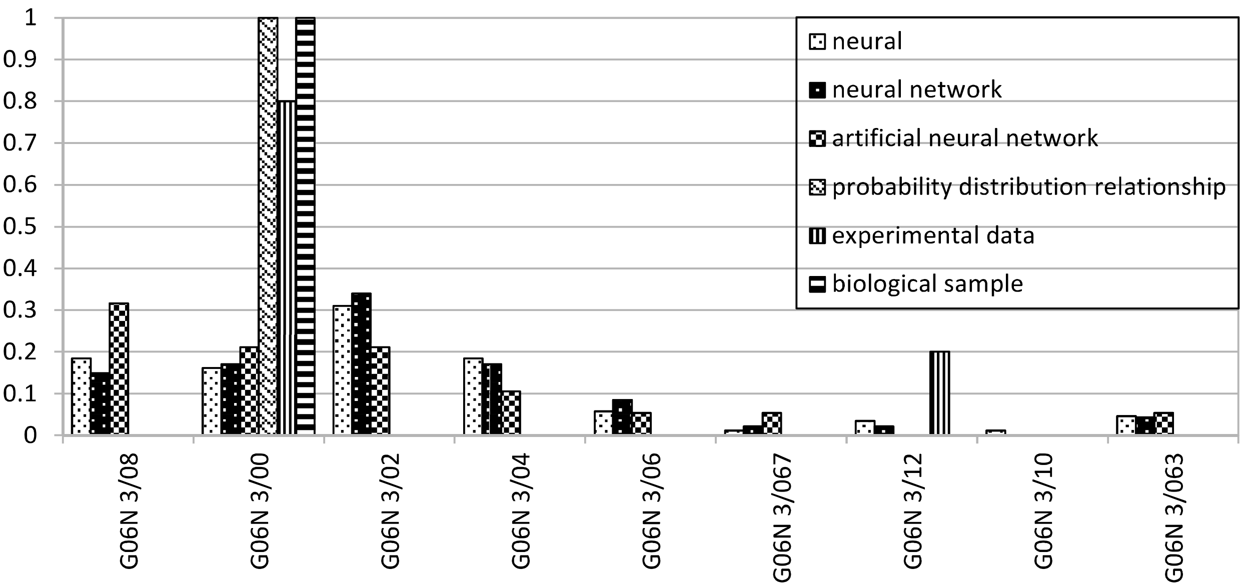


Fig. 3. Proportion of IPC subgroups within terms' document clusters

more accurate analysis on the contributions of our strategies for patent Term Hierarchy building.

## ACKNOWLEDGEMENTS

We would like to thank FAPESP and Elabora Consultoria for supporting this work.

## REFERENCES

[1] G. Ferraro, "Towards Deep Content Extraction from Specialized Discourse: The Case of Verbal Relations in Patent Claims," Master's thesis, Universitat Pompeu Fabra, Barcelona, 2012.

[2] D. J. Watts and S. H. Strogatz, "Collective Dynamics of Small-World' Networks," *Nature*, vol. 393, no. 6684, pp. 409–10, 1998.

[3] B. Lu and B. K. Tsou, "Towards Bilingual Term Extraction in Comparable Patents," in *Proceedings of the 23rd Pacific Asia Conference on Language, Information and Computation (PACLIC)*, 2009, pp. 755–762.

[4] S. Sheremetyeva, "An Efficient Patent Keyword Extractor as Translation Resource," in *Proceedings of the 3rd Workshop on Patent Translation (in Conjunction with MT-Summit XII)*, 2009, pp. 25–32.

[5] P. Lopez and L. Romary, "Experiments with Citation Mining and Key-Term Extraction for Prior Art Search," in *Proceedings of the Conference on Multilingual and Multimodal Information Access Evaluation (CLEF)*, 2010.

[6] P. Anick, M. Verhagen, and J. Pustejovsky, "Identification of Technology Terms in Patents," in *Proceedings of the 9th International Conference on Language Resources and Evaluation (LREC)*. European Language Resources Association (ELRA), 2014, pp. 2008–2014.

[7] P. Qu, J. Zhang, Y. He, W. Zeng, and H. Xu, "Term Extraction Using Co-Occurrence in Abstract and First Claim for Patent Analysis," in *Proceedings of the International Conference on Identification, Information and Knowledge in the Internet of Things (IIKI)*. IEEE, 2014, pp. 60–63.

[8] J. Kim, J. Kang, J. Lee, S. Jun, S. Park, and D. Jang, "Technology Roadmap using Patent Keyword," in *Proceedings of the International Conference on Economics and Business Management (EBM)*, 2015, pp. 141–145.

[9] M. T. Khan, Y. Ma, and J. Jae Kim, "Term Ranker: A Graph-Based Re-Ranking Approach," in *Proceedings of the 29th International Florida Artificial Intelligence Research Society Conference (FLAIRS)*. AAAI Press, 2016, pp. 310–315.

[10] W. Yang, J. Yan, and Y. Lepage, "Extraction of Bilingual Technical Terms for Chinese–Japanese Patent Translation," in *Proceedings of the NAACL Student Research Workshop*. Association for Computational Linguistics, 2016, pp. 81–87.

[11] F. Lin and F. Huang, "The Study of Patent Prior Art Retrieval Using Claim Structure and Link Analysis," in *Proceedings of the Pacific Asia Conference on Information Systems (PACIS)*. AISeL, 2010, pp. 1953–1962.

[12] C. Nédellec, W. Golik, S. Aubin, and R. Bossy, "Building Large Lexicalized Ontologies from Text: A Use Case in Automatic Indexing of Biotechnology Patents," in *Proceedings of the 17th International Conference in Knowledge Engineering and Management by the Masses (EKAW)*. Springer Berlin Heidelberg, 2010, pp. 514–523.

[13] C. Trappey, T. M. Wang, S. Hoang, and A. J. Trappey, "Constructing a Dental Implant Ontology for Domain Specific Clustering and Life Span Analysis," *Advanced Engineering Informatics*, vol. 27, no. 3, pp. 346–357, 2013.

[14] B. Ganter and R. Wille, *Formal Concept Analysis: Mathematical Foundations*, 1st ed. Springer-Verlag New York, Inc., 1997.

[15] T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Efficient Estimation of Word Representations in Vector Space," *arXiv preprint arXiv:1301.3781*, 2013.

[16] G. A. Miller, "WordNet: A Lexical Database for English," *Communications of the ACM*, vol. 38, no. 11, pp. 39–41, 1995.

[17] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent Dirichlet Allocation," *Journal of Machine Learning Research*, vol. 3, pp. 993–1022, 2003.

[18] "A Solution to Plato's Problem: The Latent Semantic Analysis Theory of Acquisition, Induction, and Representation of Knowledge," *Psychological Review*, vol. 104, pp. 211–240, 1997.

[19] M. Campr and K. Ježek, "Comparing Semantic Models for Evaluating Automatic Document Summarization," in *Proceedings of the 18th International Conference on Text, Speech and Dialogue (TSD)*. Springer International Publishing, 2015, pp. 252–260.

[20] L. Niu, X. Dai, J. Zhang, and J. Chen, "Topic2Vec: Learning Distributed Representations of Topics," in *Proceedings of the International Conference on Asian Language Processing (IALP)*, 2015, pp. 193–196.

[21] M. Seok, H. J. Song, C. Y. Park, J. D. Kim, and Y. S. Kim, "Comparison of NER Performance Using Word Embeddings," in *Proceedings of the 4th International Conference on Artificial Intelligence and Application*, 2015, pp. 16–19.

[22] G. Salton, *Automatic Information Organization and Retrieval*. McGraw Hill Text, 1968.

[23] R. Pum and C. Key, “Measuring the Specificity of Terms for Automatic Hierarchy Construction,” in *Proceedings of the 16th European Conference on Artificial Intelligence, Workshop on Ontology Learning and Population*, 2004.

[24] G. Grefenstette, “INRIASAC: Simple Hypernym Extraction Methods,” in *Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval).* Association for Computational Linguistics, 2015, pp. 911–914.

[25] M. J. Kusner, Y. Sun, N. I. Kolkin, and K. Q. Weinberger, “From Word Embeddings to Document Distances,” in *Proceedings of the 32nd International Conference on Machine Learning (ICML)*, 2015, pp. 957–966.