Classification of Knowledge Representations using an Ontology-based Approach

Ruben Costa[1], Paulo Figueiras[1], Pedro Maló[1] and Celson Lima[2]
[1]*CTS, Uninova, Dep.ª de Eng.ª Electrotécnica, Faculdade de Ciências e Tecnologia, FCT, Universidade Nova de Lisboa, 2829-516 Caparica, Portugal*
[2]*UFOPA / IEG / PSI, Federal University of Western Pará, Santarém, Brazil*

Keywords: Ontology Engineering, Unsupervised Document Classification, Vector Space Models, Semantic Vectors.

Abstract: One of the primary research challenges in the knowledge representation domain relates to the process of formalization of document contents using dependent metadata and in particular how the classifiers are derived. Most approaches to determining appropriate classifiers are limited and only take account of the explicit, word-based information in the document. The research described in this paper explores the potential classifier enrichment through incorporation of implicit information derived from the complex relationships (Semantic Associations) in domain ontologies with the addition of information presented in documents for unsupervised document classification. The paper introduces a novel conceptual framework for representation of knowledge sources, where each knowledge source is semantically represented (within its domain of use) by a Semantic Vector (SV), which is enriched using the classical vector space model approach extended with ontological support, employing ontology concepts and their relations in the enrichment process. The test domain for the assessment of the approach is Building and Construction, using an appropriate available Ontology. Preliminary results were collected using a clustering algorithm for document classification, which indicates that the proposed approach does improve the precision and recall of classifications. Future work and open issues are also discussed.

1 INTRODUCTION

The representation of knowledge has been an important human endeavor since the dawn of the human race. The creation of written and spoken languages is the best known example of the effort to represent knowledge in such ways as to preserve it and to guarantee that it will be transmitted to future generations.

The subject of knowledge representation gained a new dimension with the advent of the computer age. Particularly, with the creation of the World Wide Web, new forms of knowledge representation were needed in order to transmit data from source to recipient in common data formats, and to aid humans to find the information they want in an easily understandable manner.

With the evolution of the Semantic Web, knowledge representation techniques got into the spotlight, aiming at bringing human understanding of the meaning of data to the world of machines. Such techniques create knowledge representations of knowledge sources (KS), whether they are web pages or documents (Figueiras et al., 2012).

Most existing information retrieval techniques are based upon indexing keywords extracted from KS. Regrettably, keywords or index terms alone often cannot adequately capture the document contents, resulting in poor retrieval and indexation performances. Nevertheless, keyword indexing is widely used in commercial systems because it is still the most viable way by far to process large amounts of text.

This paper illustrates the development of a framework which supports the process of a representation of knowledge sources, using a vector space model (VSM) (Salton et al., 1975) approach and the enrichment of such representation using background knowledge available in a domain ontology. The proposed work will be assessed in the building and construction sector. The major steps of the work include the analysis of the relations between ontological concepts, and the KS they are representing as well as the enhancement of such

Costa R., Figueiras P., Maló P. and Lima C..
Classification of Knowledge Representations using an Ontology-based Approach.
DOI: 10.5220/0004541701840191
In *Proceedings of the International Conference on Knowledge Engineering and Ontology Development* (KEOD-2013), pages 184-191
ISBN: 978-989-8565-81-5

relations with semantic associations among concepts. Hence, the main contribution of this work is consequently not trying to develop new or improving any of the current classification algorithms but to affect the document term vectors in a way that we could and measure the effect of such semantic enrichment on existing classifiers.

This paper is structured as follows. Section 2 presents the related work. Section 3 illustrates the domain ontology used under this work. Section 4 describes the process of enrichment of KSs. Section 5 illustrates the empirical evidences of the work addressed so far. Finally, section 6 concludes the paper and points out the future work to be carried out.

2 RELATED WORK

The presented work is the continuation of the work presented in (Figueiras et al., 2012) and (Costa et al., 2012). In terms of the issue addressed here, Castells et al. (Castells et al., 2007) propose an approach based on an ontology and supported by an adaptation of the Vector Space Model, similarly to our approach. It uses the *tf-idf* (term frequency–inverse document frequency) algorithm, matches documents' keywords with ontology concepts, creates semantic vectors, and uses the cosine similarity to compare created vectors. A key difference between this approach and the presented work is that Castells' work does not consider semantic relations or the hierarchical relations between concepts (both taxonomic and/or ontological relations).

Li (Sheng, 2009) presents a way of mathematically quantifying such hierarchical or taxonomic relations between ontological concepts, based on relations' importance and on the co-occurrence of hierarchically related concepts, and reflects this quantification in documents' semantic vectors. Li's work aims at creating an Information Retrieval (IR) model based on semantic vectors to apply over personal desktop documents, and it has no relation to Web IR applications, as is the case of the presented work.

On the other hand, Nagarajan et al. (Nagarajan et al., 2007) propose a document indexation system based on the VSM and supported by Semantic Web technologies, just as we do here. They also propose ways of quantifying ontological relations between concepts, and represent that quantification in documents' semantic vectors. There are some differences between Nagarajan's work and our approach. For instance, Nagarajan et al. do not distinguish between taxonomic and ontological relations, also our work doesn't not include terms from documents within semantic vectors, such terms previously semantically mapped to ontology concepts.

Focusing on more recent works, Xia et al. (Xia and Du, 2011) propose a document classification mechanisms based on title vector based document representations, in which is assumed that terms in documents' titles represent main topics in those documents, and therefore the weights for title terms should be amplified.

Finally, the work of García et al. (García et al., 2010) aims to propose some new metrics to measure relationships among classes in an ontology. Relationships among classes in an OWL ontology are given by the object properties that are defined as a binary relation between classes in the domain with classes in the range. The proposal of García et al. is based on the coupling metric defined in the software engineering field, adapting it to the Semantic Web's needs.

3 THE ONTOLOGY

The domain-specific ontology used in this work was entirely developed using Protégé ontology editor (Stanford Center for Biomedical Informatics Research, s.d.), and it is written in OWL-DL language (Sean et al., s.d.). The ontology comprehends two major pillars, namely, concepts and relations. The first relates to specific elements (classes) of building and construction related areas which cover for example, type of project, project phase, and similar data. The other specifies how such concepts are related to each other.

Several levels of specificity are given for all concept families, as described for the 'Actor' concept. These specificity levels represent concepts hierarchies and, ultimately, taxonomic relations such as 'Architect' <is_a> 'Design Actor' and 'Design Actor' <is_a> 'Actor'. All classes, or concepts, have an instance, which corresponds to the class, and comprises the keywords or expressions gathered and related to each concept, through an ontological datatype property designated 'has Keyword'.

All concepts are themselves keywords, because they are expressions or terms that may occur in a knowledge source. In addition to themselves, concepts also possess *equivalent terms* that are terms or expressions relevant for capturing different semantic aspects of such concepts. For instance, the

'Learning_Facility' concept has a 'Higher_Education_Facility' individual, and this individual has several keywords designated as equivalent terms, such as 'university', 'science college', and 'professional college', meaning that each equivalent term belongs to some concept, as shown in Figure 1. Moreover, concepts are connected by ontological object properties called *ontological relations*. Ontological relations relate concepts among themselves and are described by a label (property) and the relevance (weight) of such relation in the context of the B&C domain ontology.

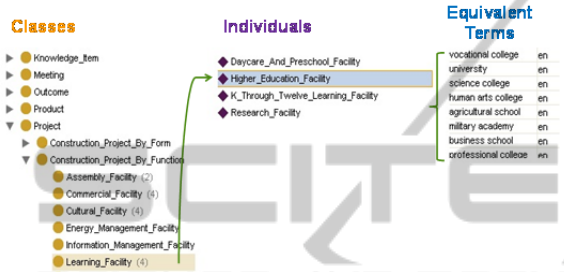

Figure 1: Domain Ontology elements.

4 THE PROCESS

In this section, we describe the justification behind our hypothesis that background knowledge available in domain ontologies can be used to enrich statistical term vectors representations. Our approach mainly focuses on knowledge representation of knowledge sources, but there are several steps that need to be performed before and after the knowledge representation itself. Figure 2 gives a general overview of our process, which consists of two main modules, namely *Document Analysis Module* and *Semantic Enrichment Module*.

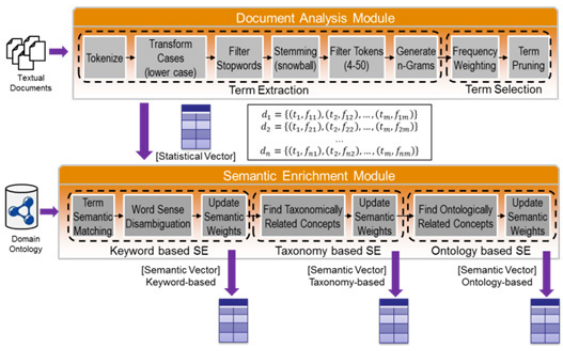

Figure 2: The process.

4.1 Document Analysis Module

We start with a state-of-the art indexing tool, called RapidMiner (RapidMiner, 2012), to generate document term vectors (statistical vector) where terms are ordered by their level of importance within a document using a normalized *tf-idf* score.

There are two stages in the first module, namely *Term Extraction* and *Term Selection*, for reducing the dimensionality of the source document set. Both are described here.

4.1.1 Term Extraction

The whole extraction process is as follows:

a) First of all, each document is broken into sentences. Then, terms in each sentence are extracted as tokens (this process is called tokenization).
b) All tokens found in the document are transformed to lower case.
c) The terms belonging to a predefined stop word list are removed.
d) Remained terms are converted to their base forms by stemming, using the snowball method. The terms with the same stem are combined for frequency counting. In this paper, a term is regarded as the stem of a single word.
e) Tokens whose length is "< 4" or "> 50" characters are discarded.
f) The n-Grams generation is seen here as a creation of sequences of 1 to N words. For this case we are considering the generation of unigrams, bigrams (e.g. Waste Management) and trigrams (e.g. Electric Power Product).

4.1.2 Term Selection

We understand that terms of low frequencies are supposed as noise and useless, thus we apply the *tf–idf* (term frequency - inverse document frequency) method to choose the key terms for the document set. Equation 1, is used for the measurement of $tfidf_{ij}$ for the importance of a term t_j within a document d_i. The main limitation of *tf-idf* method is that long documents tend to have higher weights than short ones. It considers only the weighted frequency of the terms in a document, but neglects the length of the document. In Equation 2, tf_{ij} is the frequency of t_i in d_j, and the total number of occurrences in d_j is the maximum frequency of all terms in d_j used for normalization to prevent bias for long documents.

$$tfidf_{ij} = tf_{ij} * idf_i \quad (1)$$

$$tf_{ij} = \frac{number\ of\ occurrences\ of\ t_i\ in\ d_j}{total\ number\ of\ occurences\ in\ d_j} \quad (2)$$

$$idf_i = \log \frac{number\ of\ documents\ in\ D}{number\ of\ documents\ in\ D\ that\ contain\ t_i} \quad (3)$$

After calculating the weight of each term in each document, those which satisfy the pre-specified minimum *tf–idf* threshold γ are retained. For this work, we consider all terms where its *tf-idf* score was greater or equal than *0.001*. Subsequently, these retained terms form a set of key terms for the document set *D*.

A document, denoted d_i is a logical unit of text, characterised by a set of key terms t_j together with their corresponding frequency f_{ij}, and can be represented by $d_i = \{(t_1, f_{i1}), (t_2, f_{i2}), \dots, (t_j, f_{ij}), \dots, (t_m, f_{im})\}$. Such representation is entitled statistical vector, meaning that, for each document in *D* there is a resultant statistical vector. An example of a statistical vector is depicted in Table 1.

Table 1: Statistical Vector.

Key Term	Weight
sanitari	0,004101
water_suppli_drainag	0,003265
toilet	0,002482
personnel	0,002332

4.2 Semantic Enrichment Module

In this module we construct a new term vector, named Semantic Vector (SV) for all the documents in *D*. This vector comprises of ontology concepts that are on the domain ontology and whose equivalent terms semantically match terms which are present in the statistical vector, (Table 2).

Table 2: Ontological Equivalent Terms.

Ontological Concept	Equivalent Terms
Complete_Sanitary_Suite	complete sanitary suite, complete bathroom suite, bathroom, washroom,…
Plumbing_Fixture_And_ Sanitary_Washing_Unit	bathtub, shower, service sink, lavatory,…
Sanitary_Disposal_Unit	water closet, toilet, urinal,…

A semantic vector is represented by two columns: the first column contains the concepts that build up the knowledge representation of the KS, i.e., the most relevant concepts for contextualizing the information within the KS; the second column keeps the degree of relevance, or weight, that each term has on the knowledge description of the KS (Costa et al., 2012).

Our approach takes into account three different but complementary procedures for building up the semantic vector, where each iteration is expected to add new semantic enrichment of the KS representation: keyword-based, taxonomy-based, and ontology-based semantic vectors. The first step is related with the definition of a keyword-based semantic vector.

4.2.1 Keyword-based Semantic Vector

The keyword-based semantic vector takes into consideration only the relation between terms existing in the statistical vector and ontology concepts presented on the domain ontology.

In this module, we use semantic background knowledge from ontologies as a way to augment traditional syntactic term vectors. A fundamental drawback behind Vector Space Model is that it treats a document as a bag of words and ignores the dependence between terms, i.e., it assumes that terms in a document occur independent of each other. Capturing dependency between key terms within syntactic term vectors in terms of co-occurrences has been successfully attempted by the use of statistical techniques (Nagarajan et al., 2007). However there are cases when terms do not co-occur very often and are also not related in a way that such techniques can help. For example, if terms “bathtub” and “shower” in Table 2 do not co-occur frequently, statistical techniques will fail to identify a possible correlation between them.

The next iteration deals with finding similarities between the statistical vector’s keywords and equivalent terms which are linked to ontological concepts from the domain ontology. The matching process between equivalent terms presented on the domain ontology and the keywords within the statistical vector is done by using a similarity measure between words (cosine similarity).

The keyword-based semantic vector is then stored in the database in the form $\left[\sum_{i=1}^{n} x_i\ ;\ \sum_{i=1}^{n} w_{x_i}\right]$, where *n* is the number of concepts in the vector, x_i is the syntactical representation of the concept and w_{x_i} is the semantic weight corresponding to the concept.

Table 3 depicts the weight of every ontology concept associated to each key term within the

statistical vector, where the first column corresponds to the ontology concepts that were matched to describe the most relevant terms extracted from the statistical vector, the second column indicates the most relevant terms that were matched to ontology equivalent terms, and the third column indicates the semantic weight for each ontology concept matched.

Table 3: Keyword-based semantic vector.

Concept	Key Term	Weight
Sanitary_Disposal_Unit	toilet, urin, water_closet	0,149514
Sanitary_Laundry_and_Cleaning_Equipment_Product	sanitari	0,132629
Team	person, personnel	0,104497
Commitee	subcommitte	0,067880

4.2.2 Taxonomy-based Semantic Vector

Taxonomy-based vectors push one step further in the representation of KSs by adjusting the weights between expressions according to the taxonomic relation among them, i.e., expressions that are related with each other with the 'is_a' type relation. If two or more concepts that are taxonomically related appear in a keyword-based vector, the existing relation can boost the relevance of the expressions within the KS representation.

Definition 1: In the hierarchical tree structure of the ontology, concept A and concept B are homologous concepts if the node of concept A is an ancestor node of concept B. Hence, A is considered the nearest root concept of B, R(A,B). The taxonomical distance between A and B is given by:

$$d(A,B) = |depth(B) - depth(A)| = |depth(A) - depth(B)| \quad (4)$$

In Equation 4, depth (X) is the depth of node X in the hierarchical tree structure, with the ontological root concept's depth being zero (0).

Definition 2: In the hierarchical tree structure of the ontology, concept A and concept B are non-homologous concepts if concept A is neither the ancestor node nor the descendant node of concept B, even though both concepts are related by kin; If R is the nearest ancestor of both A and B, then R is considered the nearest ancestor concept for both A and B concepts, R(A,B); The taxonomical distance between A and B is expressed as:

$$d(A,B) = d(R,A) + d(R,B) \quad (5)$$

Figure 3 depicts the difference between homologous and non-homologous concepts.

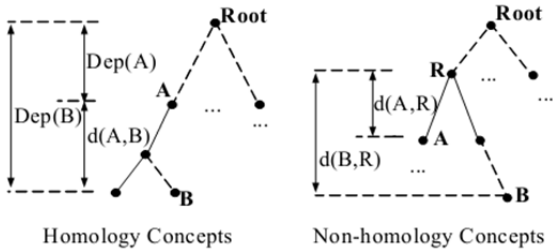

Figure 3: Homologous and non-homologous concepts (Sheng, 2009).

The taxonomy-based semantic vector is calculated using the keyword-based vector as input, where taxonomical relations are used to boost the relevance of the concepts already present within the vector or to add new concepts. The weight of the concepts is boosted when two concepts found in the keyword-based vector are highly relevant, with the degree of relevance being defined by a given threshold. If the relevance of the taxonomical relation between two concepts is higher than the predefined threshold, then the semantic weight of such concepts is boosted in the taxonomy-based vector. If a concept already present in the keyword-based vector is taxonomically related to a concept than is not present in the vector, then the related concept is added into the taxonomy-based vector.

An example of a taxonomy-based semantic vector is depicted in Table 4. The taxonomical similarity is calculated differently for both homologous and non-homologous taxonomical relations defined previously:

$$Sim(A,B) = \left(1 - \frac{\alpha}{depth(A)+1}\right)\frac{\beta}{d(A,B)}\frac{son(B)}{son(A)} \quad (6)$$

If $d(A,B) \neq 0$ and A and B are homologous.

$$Sim(A,B) = \left(1 - \frac{\alpha}{depth(R)+1}\right)\frac{\beta}{d(A,B)}\frac{son(A)+son(B)}{son(R)} \quad (7)$$

If $d(A,B) \neq 0$ and A and B are non-homologous.

$$Sim(A,B) = 1 \quad (8)$$

If $d(A,B) = 0$.

Table 4: Taxonomy-based semantic vector.

Concept	Weight
Sanitary_Disposal_Unit	0,107615
Sanitary_Laundry_and_Cleaning_Equipment_Product	0,092500
Team	0,075767
Plumbing_Fixture_and_Sanitary_Washing_Unit	0,057912

The concept 'Plumbing_Fixture_and_Sanitary _Washing_Unit' weight was boosted within the Taxonomy-based semantic vector because it is highly related with the concepts 'Sanitary_Disposal_Unit' and 'Sanitary_Laundry_and_Cleaning_Equipment_prud ct'.

4.2.3 Ontology-based Semantic Vector

The third iteration in the semantic vector creation process is the definition of the semantic vector based on the ontological relations defined in the domain ontology. Our system uses human input (knowledge experts in the building and construction domain) to establish the final numerical weights on each ontological relationship.

The first step is to analyse the ontological relations among concepts found in the input semantic vector. The taxonomy-based semantic vector is used as input for this analysis. The creation of the ontological-based semantic vector is a two-step process: the first step boosts weights of concepts already present in the taxonomy-based vector, depending on the relevance of the ontology associations among them; the second step adds new concepts that are not present in the input vector, according to ontological relations they might have with concepts belonging to the taxonomy-based vector (Costa et al., 2012).

Analogously to the creation of a taxonomy-based semantic vector, the new concept is added to the semantic vector only if the importance of an ontological relation exceeds a pre-defined threshold, for the same constraint purposes. The ontological relation's significance, or relevance, is not automatically computed; rather, as explain before, it is calculated by knowledge experts in the building and construction domain, and is defined by a vector comprising a pair of concepts and the weight associated to the pair relation, as shown in Table 5.

Table 5: Ontological Relations.

Property	Subject	Object	Weight
is_part_of	Complete _Sanitary _Suite	Sanitary_Laun dry_and_Clea ning_Equipme nt_Product	0,07
is_part_of	Sanitary_ Disposal_ Unit	Sanitary_Laun dry_and_Clea ning_Equipme nt_Product	0,07

The equation 9 describes the process of boosting of concepts or addition of new ones. Where Ow_{C_y}, is the new weight of the ontological concept, Tw_{C_y} is the taxonomy weight of the concept to be boosted, if the concept is added then Tw_{C_y} should be zero. Tw_{C_x} is the taxonomical weight of the concept related to C_y and $TI_{C_xC_y}$ is the weight of the relation between C_y and C_x.

$$Ow_{C_y} = Tw_{C_y} + \sum(all\ related\ C_xs)\left[Tw_{C_x} * \left(TI_{C_xC_y}\right)\right] \quad (9)$$

An example of an ontology-based semantic vector is depicted in Table 6.

Table 6: Ontology-based semantic Vector.

Concept	Weight
Sanitary_Disposal_Unit	0,111718
Sanitary_Laundry_and_Cleaning_Equi pment_Product	0,099504
Team	0,074115
Plumbing_Fixture_and_Sanitary_Was hing_Unit	0,056649

In this example, the concepts 'Sanitary_Disposal_Unit' and 'Sanitary_Laundry_and_Cleaning_Equipment_Prod uct' where boosted because they are already present in the taxonomy-based vector and are related by the ontological relation '<is_part_of>'.

5 ASSESSMENT OF THE PRESENTED WORK

Our dataset for evaluation in this paper is primarily focused in related products used in building and construction. Figure 4 shows part of the taxonomy that we classified the documents into. Although the taxonomy related with product contains 16 sub-categories, we chose a small subset (5 categories as shown in Figure 4).

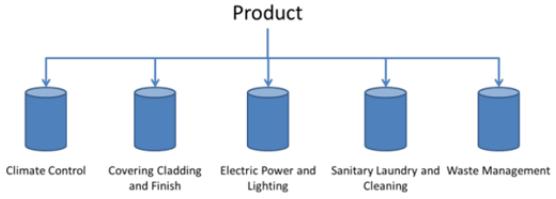

Figure 4: Categories used for evaluation.

We tested our approach with 20 scientific publications containing on average 3.500 words each. The reason for choosing scientific publications was the significant amount of words in each document, which makes the scattering of each

document in terms of key terms much higher when compared to simple webpages or news headlines, making the precise classification a challenge.

All our test documents were manually pre-labeled with the support of ICONDA search engine (IRB, 1986) and a close human evaluation.

The final goal of the assessment is to measure into what extent, a document altered term vector using the proposed approach, implies a more meaningful representation of its contents. In other words, can we affirm that, adding new concepts, boosting the important ones and removing the less important ones from a semantic vector leads to a truly enrichment of KS representations? In order to answer such, we must first verify, if classifiers can perform better clustering analysis, by grouping documents which are more similar within the same category, using the semantic vectors

Our system uses the altered term vectors as inputs to various classification algorithms - specifically, we used an unsupervised classification algorithm for the evaluations (K-Means clustering (MacQueen, 1967)).

In the following sub-section, we present the results of our approach and give details on the kinds of classification patterns we have observed.

5.1 Results

Our metrics for evaluation of our approach are based on the traditional notions of precision and recall. Nevertheless, the precision of such classification tends to be a subjective issue. As an example, the way how ontology relations between concepts were evaluated will deeply affect such classification. As stated before, our system uses human input (knowledge experts in the building and construction domain) to establish the final numerical weights on each ontological relationship. The importance of relationships between ontological concepts is by its nature, an independent and customizable component that affects classification.

The figures below present the classification statistics. According to such results, we will explain in detail why some documents have been successfully classified and why others didn't. Average recall and precision values for 5 categories using all four vectors (see Figure 5 and Figure 6).

When analyzing in more detail the categories into which KSs have been assigned to, it was interesting to conclude that in some cases the proposed approach brought an added value and in other situations such added value was not so evident.

Considering the 'Sanitary Laundry and Cleaning'

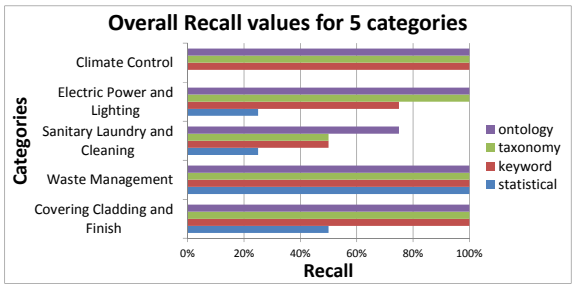

Figure 5: Overall Recall Values for 5 Categories.

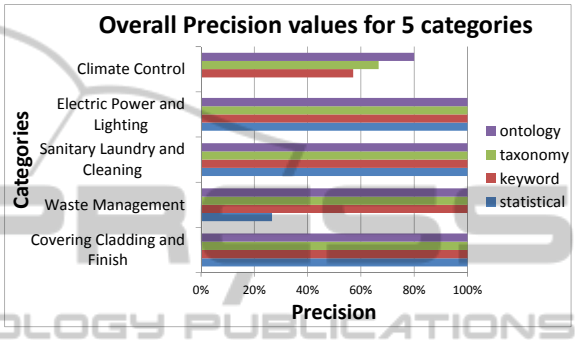

Figure 6: Overall Precision Values for 5 Categories.

category, we can conclude that using our approach there was a substantial improvement in terms of recall metric, from 25% using the statistical-based approach to 75% using the ontology-based approach. In this case, the usage of ontological relations presented in the domain ontology (as shown in Table 5), improved the recall metric from 50% to 75%.

Our results also have shown that quite a few key document terms had no direct matching with ontology equivalent terms instances, the reason for that is related with the use of an incomplete domain model (further work in extending the Ontology knowledge base can help to solve this issue to some extent) and also related with the lack of a proper method for performing word sense disambiguation during the matching process (as explained before).

It is possible for a domain Ontology to have nothing to do with the classification. The goal is to do no worse than the statistical-based approach when the Ontology is relevant or irrelevant.

Our document dataset for evaluation took into account several categories that had some similarities among key terms present in such documents. For example, contents in 'Climate Control' and 'Electric Power and Lighting' categories have a lot of similar terms that make such document classification between the categories a non-trivial task. Statistical term vectors that rely solely on document contents have shown to be poor representations, when

compared to vectors which take into account the ontology concepts and their relationships.

6 CONCLUSIONS AND FUTURE WORK

This paper's contribution targets a novel approach for the representation of unstructured information (described here as KSs) which can be applied in various areas for information retrieval, including, importantly, the semantic web. The knowledge representations enrichment process is supported using a semantic vector holding a classification based on ontological concepts. Illustrative examples showing the process are part of this paper.

The main objective behind our approach was to alter documents term vectors by relating them with domain ontology concepts, turning a term vector into a semantic vector (vector formed by ontology concepts). The way how ontology concepts are related (relatedness includes all possible relationships modeled in an Ontology), enables boosting the discriminative power of the most important concepts within KSs. A consequence of this process was the weakening (and sometimes the removal) of the less important concepts.

The results achieved so far are part of an on-going work that will evolve and mature over time and do not reflect the final conclusion of the proposed approach. Nevertheless preliminary results indicate that the inclusion of additional information available in domain ontologies in the process of representing knowledge sources can augment such knowledge representations. More extensive evaluation needs to be undertaken to reach more formal conclusions including additional metrics (rather than the classic precision and recall) for assessing the performance of the proposed method. However we can conclude that Ontologies help improve the precision of a classification.

The domain ontology itself is seen as something that is static and not evolving over time with organizational knowledge. One possible approach being considered is to extract new knowledge coming from KSs (new concepts and new semantic relations) and to reflect such new knowledge in the domain ontology. One possibility for accomplishing this may be the adoption of association rules learning algorithms, correlating the co-occurrence of terms within the document corpus. Such measures can be considered as an estimation of the probability of terms being semantically related. The weights of such semantic relations should also be updated every time new KSs are introduced into the knowledge base.

REFERENCES

Castells, P., Fernandez, M. & Vallet, D., 2007. An Adaptation of the Vector-Space Model for Ontology-Based Information Retrieval. *IEEE Transactions on Knowledge and Data Engineering,* 19(2), pp. 261-272.

Costa, R. et al., 2012. *Capturing Knowledge Representations Using Semantic Relationships.* Barcelona, Spain, IARIA.

Figueiras, P. et al., 2012. *Information Retrieval in Collaborative Engineering Projects – A Vector Space Model Approach.* Barcelona, Spain, INSTICC, pp. 233-238.

García, J., García, F. & Therón, R., 2010. *Defining coupling metrics among classes in an OWL ontology.* Cordoba, Spain, Springer-Verlag, pp. 12-17.

IRB, F., 1986. *ICONDA®Bibliographic.* s.l.:s.n.

MacQueen, J., 1967. *Some methods for classification and analysis of multivariate observations.* Berkeley, University of California Press.

Nagarajan, M. et al., 2007. *Altering Document Term Vectors for Classification - Ontologies as Expectations of Co-occurrence.* Alberta, ACM, pp. 1225-1226.

RapidMiner, 2012. *Rapid-I GmBH.* s.l.:s.n.

Salton, G., Wong, A. & Yang, C. S., 1975. A Vector Space Model for Automatic Indexing. *Communications of the ACM,* November, 18(11), pp. 613-620.

Sean, B. et al., n.d. *OWL Web Ontology Language Reference.* s.l.:W3C Proposed Recommendation.

Sheng, L., 2009. A Semantic Vector Retrieval Model for Desktop Documents. *Journal of Software Engineering and Applications,* 2(1), pp. 55-59.

Stanford Center for Biomedical Informatics Research, n.d. *Stanford's Protégé Home Page.* s.l.:s.n.

Xia, T. & Du, Y., 2011. *Improve VSM Text Classification by Title Vector Based Document Representation Method.* Singapore, IEEE.