World Academy of Science, Engineering and Technology
International Journal of Educational and Pedagogical Sciences
Vol:13, No:7, 2019

Using Textual Pre-Processing and Text Mining to Create Semantic Links

Ricardo Avila, Gabriel Lopes, Vania Vidal, Jose Macedo

***Abstract*—**This article offers a approach to the automatic discovery of semantic concepts and links in the domain of Oil Exploration and Production (E&P). Machine learning methods combined with textual pre-processing techniques were used to detect local patterns in texts and, thus, generate new concepts and new semantic links. Even using more specific vocabularies within the oil domain, our approach has achieved satisfactory results, suggesting that the proposal can be applied in other domains and languages, requiring only minor adjustments.

***Keywords*—**Semantic links, data mining, linked data, SKOS.

I. Introduction

THE Internet has afforded us extensive access to information in a highly democratic and diffuse way. As a technology that is in constant evolution, the Semantic Web has emerged with the purpose of ensuring that information found on the Web is promptly available, reusable and interoperable, thereby allowing for devices connected to the worldwide web to be truly omnipresent. The ubiquity of the Web is now part of the day-to-day of its users, providing new services that are enriched semantically through the collaboration between human beings and machines on a global scale.

Ontologies are conceptual models that allow us to share cohesive and unambiguous knowledge, providing the necessary tools for the integration of concepts. An ontology links conceptual labels to their interpretations, in other words, to the specifications of their meanings and this includes the definitions of concepts and their relationships with other concepts.

Text is the favoured means for exchanging information between specialists. With the enormous number of different databases already available to us and increasing daily, we are faced with the difficulty of locating, recovering and managing relevant information in an efficient and effective way. Furthermore, the manual construction of ontologies is a difficult, tedious, erroneous and lengthy process and one that invariably requires the collaboration of domain experts to validate and evaluate the data model developed to represent the set of concepts, and the relationships between these.

This article offers a new approach to the enrichment of ontologies using text data mining to enrich the semantic relationships between different terms, thereby generating, in a (semi)automatic way, new concepts and new links through existing data. It uses, as its test base, the vocabulary used in the domain of Oil Exploration and Production (E&P).

Ricardo Avila is with the Federal University of Ceara, Brazil (e-mail: ricardo.lims@gmail.com).

Considering that in the human world knowledge is expressed by using natural language in text, which is, generally, implicit and vague, the results for ontology that learn through text are significantly limited, thereby generating inexpressive ontologies and naked Taxonomies [26]. This article therefore seeks to use ontological learning methods as additional tools to help improve existing learning ontology techniques, thus easing the otherwise onerous task of constructing ontologies.

The remainder of this article is organised as follows: The section entitled "Terminology and Glossary Description" provides a brief summary of the most important concepts relating to semantic and ontological relationships, as well as a description of the glossaries used in the article. The section entitled "Proposed Methodology" focuses on the problem of linking text to ontologies. The section entitled "Text Mining" provides an introduction to the machine learning techniques used in this article. The section "Experiments and Results" describes the process of data collection used, the text pre-processing employed, the creation of machine learning models and an initial interpretation of the results obtained. The section "Related Works" provides a brief review of literature on the subject and finally, the section entitled "Conclusions and Future Work" completes the article.

II. Terminology and Glossary Description

An ontology O consists of a set of concepts C and their relationship R, where each $r \in R$ connects two concepts $c_1, c_2 \in C$. The correlation C between two ontologies O_1 and O_2 consists of a concept source $C_s \in O_1$, and a concept target $C_t \in O_2$. The relationship between two concepts c_1 e c_2 consists of a similarity measure of between 0 and 1, expressing the computed probability of correspondence.

A term is defined as a textual representation of a specialised concept, such as, for example, natural gas, rock, oil *etc*. The introduction of a new term implies the setting up of a new concept within a specific area of the domain of knowledge in question. The process of mapping a term for a concept within an ontology is essential to semantic interpretation.

One of the main problems of mapping is the fact that there is often no direct correlation between concepts and terms. In practice, machine learning techniques face the problem of variations in terms and their ambiguity, which makes it difficult to integrate available information into text and ontologies.

Variations in terms occur because natural language tends to express single concepts in a variety of different ways. For example, in the domain of Oil Exploration and Production

(E&P), there are many different synonyms for expressing rock, petroleum, natural gas *etc*. Having two or three different synonyms for one single concept is not uncommon in this field. The probability of two specialists in the domain of E&P using the same term when talking about the same concept is less than 15%.

The ambiguity of a term, meanwhile, occurs when it is used to refer to different concepts. Words generally have a number of different definitions in dictionaries and the meaning of a word can often vary depending on its context. For example, the term "exploration" in geology refers to the "method used as a supplement to conventional methods for exploring for oil or when these methods are not sufficiently effective", whereas in seismology, it denotes a "technique for seismic reflection processing used to determine source signatures based on registered data".

Since in natural languages one tends to use more than one word to express the same meaning, one of the goals of this present work was to identify the maximum possible number of synonyms for the concepts of the ontologies used. As this study was dealing with polysemic concepts and looking to obtain those that were most relevant within the domain under study, we decided to opt for the [16] approach.

A. Vocabulary Description

The ANP (Brazilian National Agency of Petroleum, Natural Gas and Biofuels) is responsible for regulating the activities of the oil, natural gas and biofuel sectors in Brazil. The ANP glossary includes labels in Portuguese containing 581 concepts that cover the main activities, processes, events, products, operations and other concepts relating to services linked to the extraction and production of oil natural gas and biofuels.

The DPLP (an oil sector dictionary in the Portuguese language) is available on the Web and offers an uniformization of different technical concepts relating to oil and gas, inherent not only to research and production but also to the many aspects of regulation and contracts pertinent to this sector. It offers a total of 17,168 different concepts, with labels in Portuguese, covering the areas of "Reservoir Technology", "Geology and Geophysics", "Production Technology", "Regulation and Contracts" and "Well Technology".

There are a total of 271 concepts listed in Wikipedia with labels in Portuguese, all relating to the production of oil and gas. In order to obtain the Wikipedia concepts, we carried out a process of crawling, database persistence and finally, triplication, converting the concepts obtained to an RDF format.

The three vocabularies use an SKOS data and vocab model in the representation. The glossaries are fully compatible with SKOS and consist of a set of *skos:Concept* where each one of the concepts is identified using its respective predicate *skos:prefLabel* and *skos:inScheme*, as well as having zero or more *skos:altLabels*, *skos:definition*, *skos:hiddenLabel*, *skos:subject* and *skos:scopeNote*.

These concepts can also be labelled with a zero or a *dc:creator*, *dc:date*, *skos:topConceptOf* and *skos:hasTopConceptOf* and may have zero or more *petro: urlImage*, *petro:legislation*, *petro:urlProvider* and *petro:referencedTerm*. The three glossaries differ in terms of size and granularity or coarseness, with the classes and relationships of the ontology used in this work shown in Fig. 1.

III. Proposed Methodology

The process of creating semantic links begins with the modelling of a Domain Ontology (DO). To validate the proposed concepts, we used as our baseline the predicate skos:defition of the Simple Knowledge Organization System (SKOS) [1], a data and vocabulary model that offers a formal representation of structures representing knowledge, such as catalogues, glossaries, thesauri, taxonomies, folksonomies and other types of controlled vocabulary [19]. Three vocabularies were used in this study, namely the Glossário da Agência Nacional do Petróleo, Gás Natural e Biocombustíveis (ANP) [2], the Dicionário do Petróleo em Língua Portuguesa (DPLP) [3] and a set of concepts in the area of oil and gas taken from the Wikipedia. The process of choosing these glossaries is described in the work of [5].

In the real world, such concepts contain information that can be used to identify semantic characteristics. The similarity or dissimilarity between these concepts can be measured if the links between them are represented by a set of resources, such as the SKOS vocabulary. The SKOS structure is based on the RDF standard, which allows one to represent information in a structured way and offers the possibility for integration between the different conceptual designs that use it as a baseline. In 2009, it was adopted as the official database model of the W3C for the availability and sharing of systems representing knowledge on the Web [18].

Ontologies offer clear descriptions per machine of concepts in the domain of oil and related areas. The link of specific terms in this field, in other words the textual representation of these concepts, as in, for example the descriptive fields, offers complementary resources that allow for semantic interpretations and for the discovery of new information. The use of machine learning techniques allows one to extract interpretable information on the concepts of oil, contrary to simple correlations discovered through text mining data using statistical information on occurrences between target classes of terms within this field. The knowledge extracted from text using text mining allows for the semantic enrichment of ontologies, thereby creating new components of an ontology. Fig. 2 shows the model of extracting knowledge from texts in order to enrich ontologies.

In order to discover and create new concepts and new semantic links between the components of an ontology – classes, properties and individuals – and their correspondents in natural language, we used the *skos:definition* descriptive domain of the terms that make up each glossary. The descriptive fields were used to discover new correspondences through machine learning algorithms. In a second phase, the terms that have the predicate *skos:subject* were separated into files of text, one for each different area, with each file

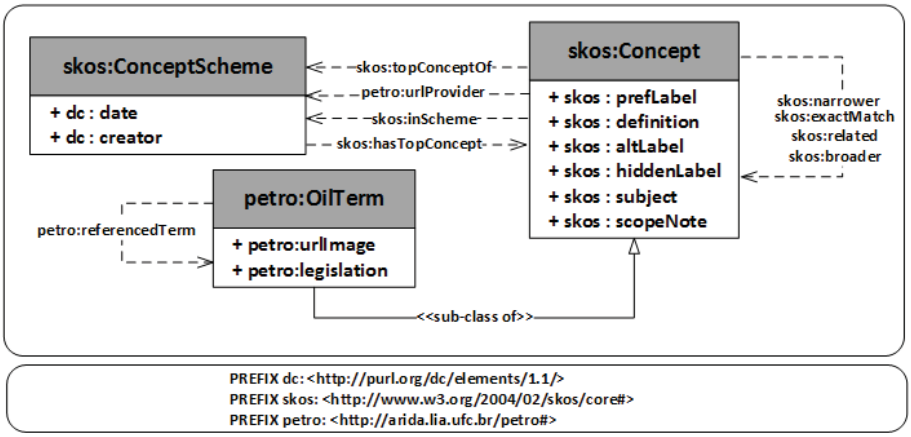

Fig. 1 Domain Ontology of the Glossary Mashu

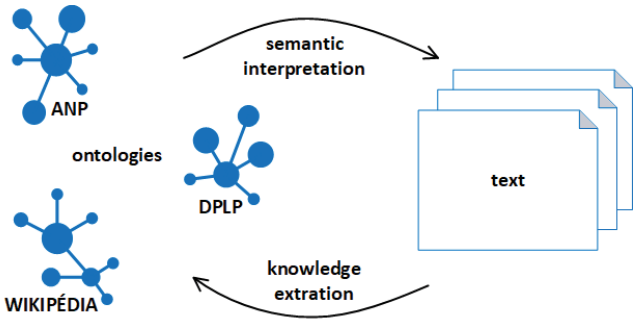

Fig. 2 Process of Creating Semantic Linksp

containing a definition relating to that area. Thus, 20 (twenty) different files were generated. We therefore used the predicate *skos:definition* in the discovery of classes, properties and semantic links, whilst the predicate *skos:subject* served as the basis for the classification of concepts in areas of knowledge within the specific domain of oil.

Stemming, lowercase and stopword removal text pre-processing techniques were used in order to improve the results of comparisons between different terms. The creation of a list of stopwords was done with the combination of two approaches, the first using a list of articles, prepositions, punctuation, conjunctions and pronouns, and the second using the identification of terms most frequently used in the context of oil.

Following pre-processing, the resulting text served as the basis for text mining through the application of machine learning techniques. Learning models were developed using algorithms of classification, regression and grouping. In order to be able to use the algorithms, when necessary, the texts were transformed into vectors with numerical characteristics.

In the context of this work, the techniques used in the transformation of the texts were able to identify the main characteristics that categorise the terms belonging to distinct classes and thereby carry out the correct process of classification.

Following the transformation of the text files into representative vectors, we were then able to classify them. A K-Fold cross-validation approach was used, with 10 folds and a calculation of the accuracy of the prediction was made using an accuracy cut-off for each dataset in each fold.

IV. Text Mining

Each mining technique needs different types of pre-processing [20]. The kind of pre-processing used varies in accordance with the characteristics of the dataset to which the mining technique is being applied. Given the large number of pre-processing methods available, many are possible combinations between methods. Definition of the method to be used may influence the performance of the result of the mining within the KDD process [20], [10], [11], [8].

After evaluating the initial results of different types of transformation, we decided to use only the stemming, lowercase and stopword removal techniques as the text pre-processing techniques most likely to improve the results of the comparisons between terms.

The sub-section below details the theories and techniques of mining used in this present work.

A. Naive Bayes

This linear classifier is a probabilistic generator for texts and one of the most widely used in machine learning thanks to its simplistic approach, which assumes that all the variables are independent. In other words, it assumes that the joint distribution of variables is equal to the product of the distribution of each one of these variables [9].

The joint distribution of probability of the classifier is given as,

$$P(A) = \frac{P(B)\,P(A)}{P(B)} \quad (1)$$

B. Support Vector Machine (SVM)

This is a linear classifier that uses the construction of a hyperplane based on the mapping of the entry vectors in a feature space that has a large number of dimensions [13]. Given that the training dataset is separable, the error rate of the SVM classifier is defined by the equation:

$$h = R^2|M^2 \tag{2}$$

where R is the radius of the shortest circumference that contains the training data and M is the margin that represents the distance between the hyperplane and the training vector closest to the feature space.

The training of the SVM classifier for problems of recognising patterns can be resolved through the optimisation of a quadratic function.

C. Linear Regression

This is a method that seeks to identify a linear function that allows for the accommodation of training data [30]. In other words, the linear regression creates a straight line through the set of points n in such a way as to make the sum of the squared residuals of the model as small as possible.

The Linear Least Square Fit (LLSF) is one of the most popular methods used to generate linear regression estimates, being influenced by the Gaussian noise. Its equation is defined as:

$$f(z) = w^R z \tag{3}$$

The LLSF algorithm calculates a vector of weight w based on the minimisation of the square loss between the model of exit $w^R z$ and *f(z)*.

D. Logistic Regression

This is a text classifier that is similar to other methods of linear classification, such as the SVM and linear regression, but that performs much better depending on the characteristics of the dataset [14].

Logistic regression seeks to model the conditional probability of p(u—z). In the case of the classification of a dataset that has two classes (binary), the probability can be modelled using the following equation:

$$p(u|z, w) = \frac{1}{1 + exp(-uw^R z)} \tag{4}$$

where p(u—z) is the conditional probability and $uw^R z$ is the function.

E. Linear Scoring Method (Scoring)

This method uses a linear function with weights to calculate the characteristics and a bias. It is used in problems of classification/categorisation where there is a need to identify the most relevant characteristics within an extremely large volume of data, calculating the weight of each one of them [29].

The scoring calculation can be done using the following equation:

$$Scoring(D) = \sum_j p_j l_j + b = pl + b \tag{5}$$

where D is the document, p_j is the weight of the j-ésima word in the dictionary of terms, b is a constant and l_j is given a value of 1 or 0, depending on whether the j-ésima word existed or not in the set of terms.

F. K-Nearest Neighbors (KNN)

This is a classification method that is considered relatively straightforward and is one of the first alternatives used when there is little or no previous knowledge available on the distribution of the data [23]. It uses a similarity measure (such as distance functions) that is usually based on the Euclidean distance between a test sample and the specified training samples.

Thus x_i is an entry sample with characteristics $p(x_{i1}, x_{i2}, \ldots, x_{ip})$, n is the total number of entry samples (i=1,2,...,n) and p is the total number of attributes (j=1,2,...,p). The Euclidean distance between the samples x_i and x_l (l=1,2,...,n) is defined as:

$$D(x_i, x_l) = (x_{i1}–x_{l1})^2 + (x_{i2}–x_{l2})^2 + ... + (x_{ip}–x_{lp})^2 \tag{6}$$

G. Decision Tree

This method of regression or classification divides the dataset up into smaller subsets and constructs a decision tree in an incremental way. The result is a tree with decision nodes and leaves. Decision trees can manipulate categoric and numerical data [7].

Thus n is a sample with two dimensions used to determine the categories of a dataset and one can use an entropy function such as:

$$i(n) = -\sum_j P(w_j) \log_2 P(w_j) \tag{7}$$

Decision trees use a distance metric function for calculating dimensions, like the Euclidean, the Manhattan or the Minkowski functions.

V. Experiments and Results

This section explains the metrics used to validate and evaluate the precision and effectiveness of the proposed method for ontology enrichment. The ontologies used were from the domain of Oil Exploration and Production (E&P).

Even a manual mapping, to discover new relationships, can achieve perfect or near perfect results. Therefore, we would like to determine the general quality of relationships, as well as the quality of the semantic links discovered. One should stress that the level of representational coarseness of the vocabularies used may generate distinctions within the context of a specific field. To mitigate this kind of problem, we have introduced different evaluation measures.

In the remainder of this section, we first discuss the measures used to evaluate the quality of the new concepts and new links (Section V.A) and then we evaluate the method

Fig. 3 Evaluation Metrics Strategy

Vocabulary	Definitions	New Concepts		New Links	
		A	M	A	M
ANP	581	88	24	45	19
DPLP	9.124	1.503	125	1.256	112
Wikipedia	274	33	13	22	10

Fig. 4 Number of New Concepts and New Links Discovered

of automatic and manual discovery (Section V.B). Finally, we apply the use of certain metrics, such as precision, coverage and the F-measure to evaluate the quality of the concepts discovered (Section V.C).

A. Evaluation Metrics

We use precision, coverage and the F-measure to evaluate the new concepts and new links, for each ontology used in the experiments. We can thus prove the quality of the vocabularies regardless of the representational coarseness of each one.

Fig. 3 shows two sets of relationships, with A generated automatically and M mapped manually by a domain expert. We determined the precision (p), coverage (r) and the F-measure (f), based on the number of relationships identified for each concept:

$$p = \frac{|A \cap M|}{M}, r = \frac{|M \cap A|}{A}, f = \frac{2pr}{p+r} \quad (8)$$

In order to validate the quality, we selected 0.10% of the total number of new concepts and new links. The selection of the test data followed an arbitrated formula with the establishment of a base 2 and an index 2 that resulted in the selection of the lines of paired entries in the files. The sample population parameters were calculated using a confidence level of 95% and a confidence interval of 5 [6].

B. Evaluation of the Automatic Discovery and the Manually Mapped Discovery

At this stage of the evaluation, we used the vocabularies listed in Section II.A for both the automatic discovery and the manual discovery. A process to discover new concepts and new semantic links was applied to each vocabulary using the proposed approach, thereby generating two sets, with A generated automatically and M mapped manually by a domain expert. Fig. 4 provides relevant information on the tasks implemented to discover the new concepts and the new links between the vocabularies. Both approaches use the descriptive domain skos:definition of the concepts in the process of text mining.

In the case of the ANP vocabulary, and using the automatic strategy, a total of 88 (15.15%) new concepts and 45 (7.75%) new semantic links were discovered. When manual mapping was used, 24 (4.13%) new concepts and 19 (3.27%) new semantic links were discovered. There is a certain dependence on the part of the number of new links generated, since these are created after the discovery of new concepts. We can thus affirm that such new links occur with both old concepts as well new ones discovered during the proposed strategy of semantic enrichment.

	r	p	f
ANP	.76	.78	.77
DPLP	.91	.90	.90
Wikipedia	.96	.95	.95

(a) New Concepts

	r	p	f
ANP	.85	.80	.82
DPLP	.90	.87	.84
Wikipedia	.93	.91	.92

(b) New Links

Fig. 5 Evaluation of Quality of the New Concepts and New Links Discovered

In the case of the DPLP vocabulary, 1,503 (16.47%) new concepts and 1,256 (13.77%) new links were discovered automatically. This proved to be the best scenario of the automatic discovery of new concepts and links, influenced by the total number of concepts and indicative of the quality of both the vocabulary and the proposed strategy, seen that the manual mapping approach only produced 125 (1.37%) concepts and 112 (1.23%) links in all.

Finally, in the case of the Wikipedia vocabulary, the automatic approach produced 33 (12.04%) new concepts and 22 (8.03%) new links. The manual approach, meanwhile produced 13 (4.74%) new concepts and 10 (3.65%) new links. Even with fewer discoveries in absolute terms, the percentages of manual discoveries in this scenario were higher when compared to the other vocabularies used in the present work. This is probably due to the quality of the definitions available.

One should point out that our approach validated the new concepts and new links generated automatically and manually in an overlapping way. The provenance of the concepts is extremely important in terms of using the proposed approach, since its respective definitions are the main resource used by the text mining techniques applied.

We consider the result of the automatic mapping scenarios of some relevance, since they deal with the discovery of new concepts and new semantic links using text pre-processing techniques together with classification, regression and clustering algorithms. The three ontologies cover the same domain, namely that of Oil Exploration and Production (E&P).

C. Evaluation of the Quality of the Concepts and Semantic Links

We can only consider the proposed approach as being significant if the true positives are considered relevant. Fig. 5 shows the results of precision (p), coverage (r) and the F-measure (f) following an evaluation by a domain expert. One should stress that only the results of automatic discoveries are shown.

The discovery of new concepts obtained an F-measure of between 77 and 95%, suggesting a good level of effectiveness on the part of the proposed approach. Precision was very good (78 to 95%), similar to the results obtained for coverage (76 to 96%). The ANP vocabulary was the one that had the

		Naïve Bayes	SVM	Linear Regression	Logistic Regression	Scoring	KNN	Decision Tree
ANP	*p*	.77	.87	.89	.89	.84	.89	.79
	r	.76	.84	.84	.85	.80	.83	.76
	f	.76	.85	.86	.86	.82	.85	.77
DPLP	*p*	.66	.78	.77	.86	.78	.78	.82
	r	.74	.74	.70	.84	.72	.73	.80
	f	.70	.76	.74	.85	.74	.75	.81
Wikipédia	*p*	.72	.89	.89	.92	.89	.86	.88
	r	.70	.86	.85	.89	.84	.84	.87
	f	.71	.87	.87	.90	.86	.84	.87

Fig. 6 Evaluation of Quality of the New Concepts and New Links Discovered by Algorithm

lowest results, 76% for coverage, 78% for precision and 77% for the F-measure. We consider the results relevant, since this approach depends on the quality of the sources used to determine the discovery of new concepts.

Thanks to the quality of their sources, the DPLP vocabulary achieved 91% in coverage, 90% in precision and 90% in the F-measure, whilst the Wikipedia was even better, at 96% in coverage, 95% in precision and 95% in the F-measure. Even though it had fewer concepts discovered in absolute terms, the Wikipedia vocabulary stood out in terms of the quality evaluation criteria proposed.

Finally, new links, which are discovered following the discovery of new concepts, obtained an F-measure of between 82 and 92%, precision of between 80 and 91% and coverage of between 85 and 93%. The discovery of new links per vocabulary was in line with the performance of the new concepts, with the ANP vocabulary coming in slightly lower with 85% in coverage, 80% in precision and 82% in the F-measure. The DPLP vocabulary produced 90% in coverage, 87% in precision and 84% in the F-measure, whilst once again, the Wikipedia vocabulary produced the best results, with 93% in coverage, 91% in precision and 92% in the F-measure.

In general, the approach proposed here works very well if we ignore the false positives. In order to identify the strong points of each machine learning technique used in this work, Fig. 6 sets out the results individually for each algorithm according to the vocabulary analysed. Once again, the table only shows the results of the automatic discoveries of new concepts.

Looking solely at the performance of the learning methods in the case of the ANP vocabulary, we can see that the results of the classifications vary from 9 to 12%, with precision of between 77 and 89%, coverage of between 76% and 85% and the F-measure of between 76 and 86%, with the Naïve Bayes classifier coming in with the poorest performance and Logistic Regression with the best performance.

The DPLP vocabulary produced the highest rates of variation between the different learning methods, with values ranging from 14 to 20% and precision of between 66 and 86%, coverage of between 70 and 84% and the F-measure of between 70 and 85%. Once again, the Naïve Bayes classifier came in with the poorest performance and Logistic Regression with the best. In our view, these variations occurred as a result of the characteristics of the sources used, which involved changes in the writing style in the production of the definitions of the concepts and, primarily, because of the large number of different areas there are within the domain of oil.

Finally, in the case of the Wikipedia vocabulary, there were variations of between 19 and 20% in the methods of learning, with precision coming in at between 72 and 92%, coverage at between 70 and 89% and the F-measure at between 71 and 90%. Once again, the Naïve Bayes classifier came in with the poorest performance and Logistic Regression with the best and again the considerable number of different writing styles contributing to the variations in the results.

One can identify Logistic Regression as the best classifier, with coverage at 92%, precision at 89% and the F-measure at 90% in the case of the Wikipedia vocabulary, considerably better than the Naïve Bayes classifier, which had the worst performance in all the scenarios, at 66% in precision, 74% in coverage and 70% in the F-measure in the case of the DPLP vocabulary. An individual analysis of the learning methods showed that the sources used had the necessary quality for an automatic identification of new concepts using the approach proposed in the present work. Even taking into account the sizable variations seen in performance between logarithms, we consider the results acceptable.

VI. Related Works

The discovery of new links between ontologies and concepts using machine learning techniques has been the focus of attention of a large number of researchers [32], [4], [12], [17], [21]. A number of different machine learning techniques can be applied as part of the approach suggested here with the most commonly used with Naive Bayes [31], including Linear Regression [30] and Logistic Regression [15], SVM [25] and Decision Trees [24]. In this present work we chose to apply other machine learning techniques for the enrichment of ontologies using text mining in order to enrich the semantic relationships between terms. In addition to the aforementioned techniques, we also included KNN and the Linear Scoring Method, comparing the results and carrying out an analysis of the performance observed through the discovery of new concepts and new links between vocabularies in the domain of oil, the focus of this research work.

The work by [22] uses a series of text processing procedures to obtain representative keywords within the specific domain under study. This kind of approach is extremely important to improving the performance of algorithm processing, considerably reducing the number of comparisons between characters. In general, text pre-processing is used in different computational approaches to help remove special symbols and other kinds of noise that might otherwise have an impact on the final result of similarity or dissimilarity between these entities.

The discovery and maintenance of links on the data web using the SILK tool was proposed by [27]. SILK uses the language of declarative specification (SILK-LSL) which allows one to specify what kinds of RDF links should be revealed from among different data sources, as well as what conditions should be fulfilled to ensure they are interlinked. It offers various kinds of similarity methods between chains of characters, such as Jaccard's similarity coefficient, the Jaro distance or the Levenshtein distance. Even though this tool is not used in the present research work, these strategies for generating links are similar to those we propose.

Other kinds of similarity methods can be used to measure the distance between chains of characters. In [28] different approaches are employed, including the edit distance based on similarity, similarity based on vectors, a strategy based on names and a strategy based on instances, among others. In the validation process we used series of data from both the OAEI and the LOD. The results showed that the methods employed offered highly precise correspondence results and that the semantic structure of the concept bases improved the results.

Most of the approaches used to discover semantic links follow the strategy of generating candidate predicates using the similarity between entities as a base, rules of inference and other types of models that largely depend on domain experts to validate the results obtained. Our approach assumes the use of a source of knowledge validated by experts from within the domain in question. The glossaries used in this present body of work offer the quality and relevance required for text mining. We would like to reaffirm that the results obtained, in accordance with the chosen methodology, were fully validated and evaluated by experts within the oil domain in order to properly verify the metrics used.

VII. Conclusions and Future Work

This article offers a new approach to the automatic discovery of semantic concepts and links in the domain of Oil Exploration and Production (E&P). It uses different learning methods combined with textual pre-processing techniques to identify the local characteristics of texts and thus generate new concepts and new links. Even using more specific vocabularies that cover different areas of knowledge within the domain of oil, our approach achieved satisfactory results, which at best obtained 92% in terms of coverage, 89% of precision and 90% of the F-measure. The proposed approach can be used in other domains and idioms, requiring only minor adjustments.

As a suggestion for future work in this area, we would like to recommend the use of vocabularies in different languages and areas of knowledge, providing that the sources used are reliable. We would also suggest the use of other machine learning methods to help identify new semantic relationships, predicates, classes and priorities.

We plan to continue our research, considering that there is still room for improvement in the results of other languages and that we can further develop strategies to mitigate the problem of the style of writing within the different repositories of knowledge available, including dictionaries, thesauri and other reliable sources. The results obtained here show that when definitions offer many different grammatical and syntactic variants, the discovery of standards and the identification of new concepts, within a domain with many areas of knowledge, tends to suffer from these noises, impairing the process of learning these algorithms.

We believe that the approach proposed here can be safely adopted, even with the variants of quantitative results, since the methodology applied works independently from the source used and from the way in which the machine learning methods are computed.

References

[1] Miles, A. & Brickley, D. (2009, August 18). *SKOS Simple Knowledge Organization System Primer.* Retrieved from https://www.w3.org/TR/skos-primer/.

[2] Agência Nacional do Petróleo, Gás Natural e Biocombustíveis (2016, August 19). *Glossário* Retrieved from http://www.anp.gov.br/glossario.

[3] Fernández, E. F., Pedrosa Junior, O., Pinho , A. C. (2015, January 7). *Dicionário do Petróleo* Retrieved from http://dicionariodopetroleo.com.br.

[4] Anthonysamy, P., Edwards, M. J., Weichel, C. & Rashid, A. (2016). *Inferring Semantic Mapping Between Policies and Code: The Clue is in the Language.* In: ESSoS (p./pp. 233-250): Springer.

[5] Avila, Ricardo, Santos, Salomao, Araujo, David, Vidal, Vania Maria Ponte and de Macedo, Jose Antonio Fernandes. *Semantic Links Using SKOS Predicates.* Paper presented at the meeting of the KES, 2017.

[6] Bland, J. M. and D. G. Altman (1996). *Transformations, means, and confidence intervals.* 312(7038), 1079.

[7] Bot, M. C. J. (2000). *Improving Induction of Linear Classification Trees with Genetic Programming.* In: Proc. of the Genetic and Evolutionary Computation Conference (GECCO-2000). Las Vegas,Nevada,USA, pp. 403–410.

[8] Brown, M. L. and J. F. Kros (2009). *Imprecise Data and the Data Mining Process.* In: Encyclopedia of Data Warehousing and Mining. IGI Global, pp. 999–1005.

[9] Chakrabarti, S. (2002). *Mining the Web: Discovering Knowledge from Hypertext Data.* Morgan-Kauffman.

[10] Engels, R., G. Lindner, and R. Studer (1997). *A Guided Tour through the Data Mining Jungle.* In: KDD. AAAI Press, pp. 163–166.

[11] Engels, R. and C. Theusinger (1998). *Using a Data Metric for Preprocessing Advice for Data Mining Applications.* In: ECAI, pp. 430–434.

[12] Hasan, M. A., V. Chaoji, S. Salem, and M. Zaki (2006). *Link Prediction Using Supervised Learning.* In: Proc. of SDM 06 workshop on Link Analysis, Counterterrorism and Security.

[13] Joachims, T. (2002). *Learning to Classify Text Using Support Vector Machines.* Kluwer Academic Publishers.

[14] Kuhn, M. and K. Johnson (2013). *Applied predictive modeling.* Vol. 26. Springer.

[15] Lampos, V., B. Zou, and I. J. Cox (2017). *Enhancing Feature Selection Using Word Embeddings: The Case of Flu Surveillance.* In: WWW. ACM, pp. 695–704.

[16] Lesk, M. (1986). *Automatic sense disambiguation using machine readable dictionaries: how to tell apine cone from an ice cream cone.* In: Proceedings of ACM SIGDOC Conference, pp. 24–26.

[17] Lichtenwalter, R., J. T. Lussier, and N. V. Chawla (2010). *New perspectives and methods in link prediction.* In: KDD, pp. 243–252.

[18] Miles, A. and S. Bechhofer (2008). *SKOS Simple Knowledge Organization System Reference.* W3C. URL: http://www.w3.org/TR/skos-reference.

[19] Miles, A., B. Matthews, M. Wilson, and D. Brickley (2005). *SKOS core: Simple Knowledge Organisation for the Web.* In: Proc. of international conference on DC and metadata applications. DC Metadata Initiative, pp. 1–9.

[20] Morik, K. (2000). *The Representation Race - Preprocessing for Handling Time Phenomena.* In: ECML. Vol. 1810. Lecture Notes in Computer Science. Springer, pp. 4–19.

[21] Muller, P., C. Fabre, and C. Adam (2014). *Predicting the relevance of distributional semantic similarity with contextual information.* In: Proc. of the 52nd Annual Meeting of the Association for Computational Linguistics. Volume 1, pp. 479–488.

[22] Mustapha, S. M. F. D. S. (2018). *Case-based reasoning for identifying knowledge leader within online community.* Expert Syst. Appl. 97, 244–252.

[23] Su, Y. and S.-U. Guan (2016). *Density and Distance Based KNN Approach to Classification.* IJAEC7(2), 45–60.

[24] Sun, S., D. Liu, G. Li, W. Yu, and L. Pang (2010). *Combination of Ontology Model and Semantic Link Network in Web Resource Retrieval.* In: SKG. IEEE Computer Society, pp. 285–288.

[25] Ukey, K. and A. Alvi (2012). *Text Classification using Support Vector Machine.* In: International Journal of Engineering and Technology (IJERT).

[26] Volker, J., P. Haase, and P. Hitzler (2009). *Learning expressive ontologies.* IOS Press.

[27] Volz, J., C. Bizer, M. Gaedke, and G. Kobilarov (2009). *Discovering and Maintaining Links on the Web of Data.* In: International Semantic Web Conference. Vol. 5823. Springer, pp. 650–665.

[28] Wang, Z., J. Li, Y. Zhao, R. Setchi, and J. Tang (2013). *A unified approach to matching semantic data onthe Web.* Knowl.-Based Syst.39, 173–184.
[29] Weiss, S. M., N. Indurkhya, T. Zhang, and F. Damerau (2005). *Text Mining: Predictive Methods for Analyzing Unstructured Information.* Springer.
[30] Zhang, C., G.-R. Xue, Y. Yu, and H. Zha (2009). *Web-scale classification with naive bayes.* In: WWW.ACM, pp. 1083–1084.
[31] Zhang, J. and Y. Yang (2003). *Robustness of regularized linear classification methods in text categorization.* In: SIGIR. ACM, pp. 190–197.
[32] Zhuge, H. (2009). *Communities and Emerging Semantics in Semantic Link Network: Discovery and Learning.* IEEE Trans. Knowl. Data Eng.21 (6), 785–799.