# ONTOLOGY-BASED HEURISTIC ALGORITHM FOR TEXT CLASSIFICATION IN OIL INDUSTRY APPLICATION

Nayat Sanchez-Pi
*ILTC – Instituto de Lógica, Filosofia e Teoria da Ciência.*
*Rua São João, 119, sala 801. 24.020-042 - Niterói RJ. Brazil*

Luis Marti Orosa [1]
*Electrical Engeenering Department. Pontificia Universidade Catolica*
*RJ. Brazil **

Ana Cristina Bicharra Garcia
*ADDLabs. Instituto de Computação. Universidade Federal Fluminense*
*Rua General Milton Tavares de Souza, s/nº. 24210-340 / Boa Viagem - Niterói RJ. Brazil*

**ABSTRACT**

Today, with the advances of new technologies, accidents, incidents and occupational health records are stored in diverse repositories. The amount of information of HSE that is daily generated has become increasingly huge. Most of this information is stored as unstructured data. The challenge of unstructured data is a top priority for industries that are looking for ways to search, sort, analyze and extract knowledge from masses of data. In this paper we provide with a solution to oil and gas industry for accident investigation using information extraction techniques. The main objective is to propose and evaluate information extraction techniques in occupational health control process, particularly, for automatic detection of accidents from unstructured texts. Our proposal divides the problem in subtasks such as text analysis, recognition and classification of failed occupational health control, resolving accidents.

**KEYWORDS**

Artificial intelligence, information extraction, ontology, text classification, text recognition, text analysis, intelligent systems

## 1. INTRODUCTION

There is an important effort of oil and gas industry to reduce the number of accidents and incidents. There are standards to identify and record workplace accidents and incidents to provide guiding means on prevention efforts, indicating specific failures or reference, means of correction of conditions or circumstances that culminated in accident. Besides, oil and gas industry is increasingly concerned with achieving and demonstrating good performance of occupational health and safety (OHS), through the control of its OHS risks, consistent with its policy and objectives of OHS.

Health, Safety and Environment (HSE) continues to be a priority issue for the offshore oil and gas industry and a determining factor in its overall success. Years passed since community takes into account the implications of oil industry to Health, Safety and the Environment. Oil and gas industries are frequently in the news. Much of the time this news is related to changes in prices of oil and gas. Another less frequent subject of media attention is when disasters strike, as in the offshore oil drilling platform explosion and fire.

The development of automatic methods to produce structured information from unstructured text sources would be extremely valuable to the oil industry. A structured resource would allow researches and industry professionals to write a single query to retrieve all the transcription interactions of any accident. Instead of

[1] Part of this work is funded by CNPq BJT Project 407851/2012-7

the thousands of abstract provided by querying the unstructured corpus, the query on the structured corpus might result in a few hundred well-formed results; this would obviously save a tremendous amount of time and energy.

The main objective is to propose and evaluate information extraction techniques in occupational health control process, particularly, for automatic detection of accidents from unstructured texts. Our proposal divides the problem in subtasks such as text analysis, recognition and classification of failed occupational health control, resolving accidents. We present an ontology-based approach to the automatic text categorization. An important and novel aspect of this approach is that our categorization method does not require a training set, which is in contrast to the traditional statistical and probabilistic methods that require a set of pre- classified documents in order to train the classifier.

## 2. RELATED WORK

Automatic text categorization is a task of assigning one or more pre-specified categories to an electronic document, based on its content. Nowadays, text classification is extensively used in many contexts. One of the examples is the automatic classification of incoming electronic news into categories, such as entertainment, politics, business, sports, etc. Standard categorization approaches utilize statistical or machine learning methods to perform the task. Such methods include Naïve Bayes [Lewis, 1998], Support Vector Machines [Vapnik, 1995], Latent Semantic Analysis [Deerwester, S., et al., 1990] and many others. A good overview of the traditional text categorization methods is presented in [Sebastiani, 2002]. All of these methods require a training set of pre-classified documents that is used for classifier training; later, the classifier can correctly assign categories to other, previously unseen documents.

However, it is often the case that a suitable set of well categorized (typically by humans) training documents is not available. Even if one is available, the set may be too small, or a significant portion of the documents in the training set may not have been classified properly. This creates a serious limitation for the usefulness of the traditional text categorization methods.

As described by the World Wide Web Consortium (W3C), ontology defines the terms used to describe and represent an area of knowledge. Ontologies are used by people, databases, and applications that need to share domain information (a domain is just a specific subject area of knowledge, such as medicine, real estate, automobile repair, or financial management). More specifically, ontology is a data model that represents a set of concepts (entities) within a given domain and the relationships between those concepts. It is used to reason about the concepts within that domain.

In this paper, we introduce a novel text categorization method based on leveraging the existing knowledge represented in a domain ontology. The novelty of this approach is that it is not dependent on the existence of a training set, as it relies solely on the entities, their relationships, and the taxonomy of categories represented in the ontology.

In the proposed approach, the ontology effectively becomes the classifier. Consequently, classifier training with a set of pre-classified documents is not needed, as the ontology already includes all important facts. The proposed approach requires a transformation of the document text into a graph structure, which employs entity matching and relationship identification. The categorization is based on measuring the semantic similarity between the created graph and categories defined in the ontology require a training set of pre-classified documents that is used for classifier training; later, the classifier can correctly assign categories to other, previously unseen documents.

However, it is often the case that a suitable set of well-categorized (typically by humans) training documents is not available. Even if one is available, the set may be too small, or a significant portion of the documents in the training set may not have been classified properly. This creates a serious limitation for the usefulness of the traditional text categorization methods.

In this paper, we introduce a novel text categorization method based on leveraging the existing knowledge represented in domain ontology. The novelty of this approach is that it is not dependent on the existence of a training set, as it relies solely on the entities, their relationships, and the taxonomy of categories represented in the ontology. In the proposed approach, the ontology effectively becomes the classifier. Consequently, classifier training with a set of pre-classified text is not needed, as the ontology already includes all relevant facts. The proposed approach requires a transformation of the unstructured text

into a graph structure, which employs entity matching and relationship identification. The categorization is based on measuring the semantic similarity between the created graph and categories defined in the ontology.

# 3. PROPOSAL

Our proposal strategy is to use an ontology as the key component of our text classification heuristic algorithm. Besides the ontology itself, the algorithm is composed of a set of modules:

1. A lemmatization, stemming and stop-word removing preprocessing. In this work we applied for this task the functionality provided by the Apache Lucence framework (Gospodnetic, O. et at., 2009).
2. A thesaurus for locating words appearing in the text in the ontology. In our case we used a customized version of OpenOffice Brazilian Portuguese thesaurus (DicSin, 2013).
3. Set of ontology elements tagged with its corresponding classification label.
4. A thesaurus crawling algorithm that takes care of determining the matching degree of text words with a corresponding ontology term.

## 3.1 Ontology

Ontologies offer knowledge that is organized in a more structural and semantic way. Their use in text categorization and topic identification has lately become an intensive research topic. As ontologies provide named entities and relationship between them, an intermediate categorization step requires matching terms to ontological entities. Afterwards, an ontology can be successfully used for term disambiguating and vocabulary unification, as presented in [Bloehdorn, S., et al., 2004]. Another approach, presented in [Nagarajan, M., et al., 2006], reinforces co-occurrence of certain pairs of words or entities in the term vector that are related in the ontology. The use of descriptions of neighboring entities to enrich the information about a classified document is described in [Gabrilovich, et al., 2006]. Interesting approach, although very different, is presented in [Wu, S.et al., 2003]; where authors automatically build partial ontology from the training set to improve keyword-based categorization method. Other categorization approaches based on using recognized named entities are described in [Sheth, A.P., et al., 2002] and [Hammond, B., et al., 2002].

An ontology is defined as "an explicit specification of a conceptualization" [Gruber, 1993]. An ontology created for a given domain includes a set of concepts as well as relationships connecting them within the domain. Collectively, the concepts and the relationships form a foundation for reasoning about the domain.

A comprehensive, well-populated ontology with classes and relationships closely modeling a specific domain represents a vast compendium of knowledge in the domain. It is only natural to expect that having such a comprehensive knowledge about the domain, one should be well equipped to create software systems implementing a variety of tasks concerning the domain of the ontology. Recently, ontologies have been used in various semantic applications, ranging from business analytics to semantic data integration [Buccella, A., et al., 2005].

We believe that the knowledge represented in such a comprehensive ontology can be used to identify topics (concepts) in a text document, provided the document thematically belongs to the domain represented in the ontology. Furthermore, if the concepts in the ontology are organized into hierarchies of higher-level categories, it should be possible to identify the category (or a few categories) that best classify the content of the document.

Within the area of computing, the ontological concepts are frequently regarded as classes that are organized into hierarchies. The classes define the types of attributes, or properties common to individual objects within the class. Moreover, classes are interconnected by relationships, indicating their semantic interdependence (relationships are also regarded as attributes) [Sheth, A.P., 2003]. Class hierarchies and class relationships form the schema level of the ontology, while the individuals (object instances or just instances) and links among them (relationship instances) form the so-called ground level of the ontology.

We built domain ontology for the Health, Safety and Environment (HSE) of oil and gas domain, see Figure 1. We also obtain the inferences that describe the dynamic side and finally we group the inferences sequentially to form tasks. Principal concepts of the ontology, see Figure 2, are the following:

- **Anomaly:** Undesirable event or situation which results or may result in damage or faults that affect people, the environment, equity (own or third party), the image of the Petrobras System, products or production processes. This concept includes accidents, illnesses, incidents, deviations and non-conformances.
    - **Neglect**: Any action or condition that has the potential to lead to, directly or indirectly, damage to people, to property (own or third party) or environmental impact, which is inconsistent with labor standards, procedures, legal or regulatory requirements, requirements management system or practice.
        - **Behavioral neglect:** Act or omission which, contrary provision of security, may cause or contribute to the occurrence of accidents.
        - **Non-behavioral neglect:** Environmental condition that can cause an accident or contribute to its occurrence. The environment includes adjective here, everything that relates to the environment, from the atmosphere of the workplace to the facilities, equipment, materials used and methods of working employees who is inconsistent with labor standards, procedures, legal requirements or normative requirements of the management system or practice.
    - **Incident:** Any evidence, personal occurrence or condition that relates to the environment and / or working conditions, can lead to damage to physical and / or mental.
    - **Accident:** Occurrence of unexpected and unwelcome, instant or otherwise, related to the exercise of the job, which results or may result in personal injury. The accident includes both events that may be identified in relation to a particular time or occurrences as continuous or intermittent exposure, which can only be identified in terms of time period probable. A personal injury includes both traumatic injuries and illnesses, as damaging effects mental, neurological or systemic, resulting from exposures or circumstances prevailing at the year's work force. In the period for meal or rest, or upon satisfaction of other physiological needs at the workplace or during this, the employee is considered in carrying out the work.
        - **Accident with injury:** It's all an accident in which the employee suffers some kind of injury.
            - **Injury**: Any damage suffered by a part of the human organism as a consequence of an accident at work.
                - **With leave:** Personal injury that prevents the injured from returning to work the day after the accident or resulting in permanent disability. This injury can cause total permanent disability, permanent partial disability, total temporary disability or death.
                - **Without leave:** Personal injury that does not prevent the injured to return to work the day after the accident, since there is no permanent disability. This injury, not resulting in death, permanent total or partial disability or total temporary disability, requires, however, first aid or emergency medical aid. Expressions should be avoided "lost-time accident" and "accident without leave", used improperly to mean, respectively, "with leave injury" and "injury without leave."
            - **Without Injury:** Accident causes no personal injury.

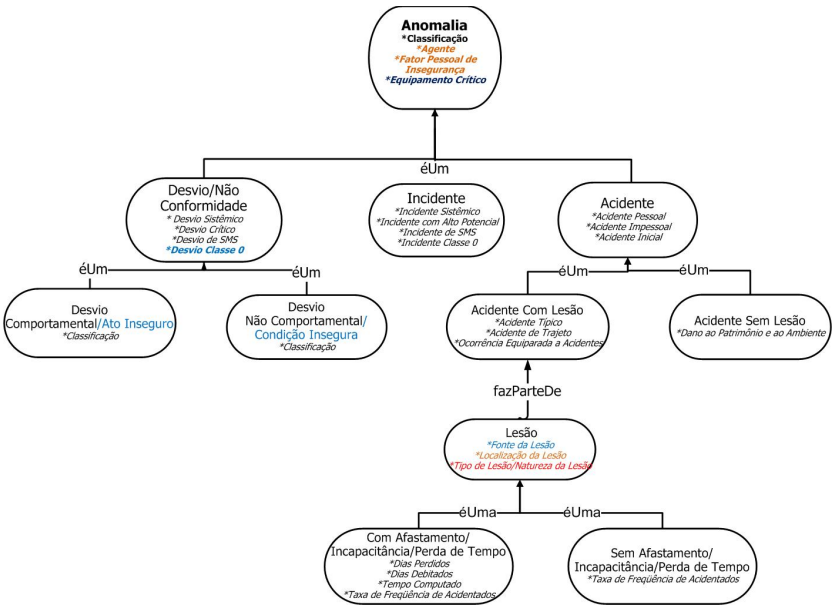


**Figure 2.** Ontology for Health, Safety and Environment in Oil Industry. *Anomaly* Concept.

## 3.2 Classification algorithm

As mentioned before, the classification algorithm proposed in this work relies on the previous ontology, a thesaurus to establish the degree of matching between a given sentence and some ontology terms of interest. The algorithm is presented as pseudo-code in Figure 1 as function ClassifyText(). It proceeds by first filtering and rearranging the input sentence in order to render it in a format suitable for processing (PreprocessText() method in step 1 of Figure 3). We have employed Apache Lucene text processing tools for stemming, lemmatization and stop-word removal.

Having the filtered text represented as a set of words, the algorithm proceeds to identify which terms of the ontology are most closely related to that set. It carries that out by invoking for each word the function ComputeSimilarityLevels(). This function —which is described in Figure 4— returns the set of ontology terms that are related with a given word by recursively traversing a thesaurus up to a given number of levels. If a connection between a word and a term is established that term is included, along with its level of similarity in the set of related terms Θ. The level of similarity is defined as the number of jumps needed to get from to word to the term using the thesaurus. A lower level implies higher similarity.

The result of the classification is one or more ontology terms that are most closely related to the text, or, posed in other words, the terms with minimal level of similarity. It should be beard in mind that the two functions presented here have been simplified for didactical reasons, and in practice some a harder to read but more efficient option is used.

| | |
|---|---|
| **function:** | ClassifyText($t$): $\langle \Gamma, l \rangle$ |
| **input:** | $t$: **string** – the text to be classified. |
| **output:** | $\Gamma := \{w_0, \dots, w_n\}$ – set of ontology terms that can classify the current text.<br>$l$: **integer** – level of similarity of the terms. |
| 1 | $\Omega =$ PreprocessText($t$) – yields a set of words after lemmatization, stemming and stop-word removal. |

```
2   l = ∞ – best similarity level.
3   Γ = ∅.
4   for each w_k ∈ Ω do
5       Θ_k = ComputeSimilarityLevels(w_k, 0).
6       for each ⟨w_j, l_j⟩ ∈ Θ_k
7           if l_j < l then
8               Γ = {w_j}.
9               l = l_j.
10          else if l_j = l then
11              Γ = Γ ∪ {w_j}.
12          end-if
13      end-for each
14  end-for each
15  return ⟨Γ, l⟩.
```

**Figure 3.** Pseudo-code description of the algorithm used to compute the levels of similarity between a given word found in text and ontology terms.

```
function:  ComputeSimilarityLevels(w_0, l_0): Θ
input:     w_0: string – the word to be processed.
           l_0: integer – initial level.
global:    l_max: integer – maximum number of levels to be reached.
output:    Θ := {⟨w_1, l_1⟩, ..., ⟨w_n, l_n⟩} – set of pairs of ontology term and level of similarity.
           Each pair represents how "close" is word w_n to w.

1   if l_0 = l_max then
2       return Θ = ∅ – max. number of levels reached.
3   end-if
4   if OntologyContains(w_0) then
5       return Θ = {⟨w_0, l_0⟩} – the word is a term of the ontology, no further search
        is necessary.
6   end-if
7   Υ = Thesaurus(w_0) – determine the synonyms of w_0.
8   Θ = ∅.
9   for each w_k ∈ Υ do
10      Θ_k = ComputeSimilarityLevels(w_k, l_0 + 1).
11      Θ = Θ ∪ Θ_k.
12  end-for each
13  return Θ.
```

**Figure 4.** Pseudo-code description of the algorithm used to compute the levels of similarity between a given word found in text and ontology terms.

# 4. CONCLUSION

In this paper, we introduce a novel text categorization method based on leveraging the existing knowledge represented in domain ontology. The novelty of this approach is that it is not dependent on the existence of a training set, as it relies solely on the entities, their relationships, and the taxonomy of categories represented in the ontology. The classification algorithm proposed herein has an adequate computational performance. However it has some clear drawbacks when confronted to complex and contradictory texts. This is not an issue for our application domain. In spite of the texts are written in a natural language, for this particular

domain, unstructured texts are written in a very direct discourse and there was no a large variation in the amount of information in each text, issues that were good for the step 1. See Figure 5.

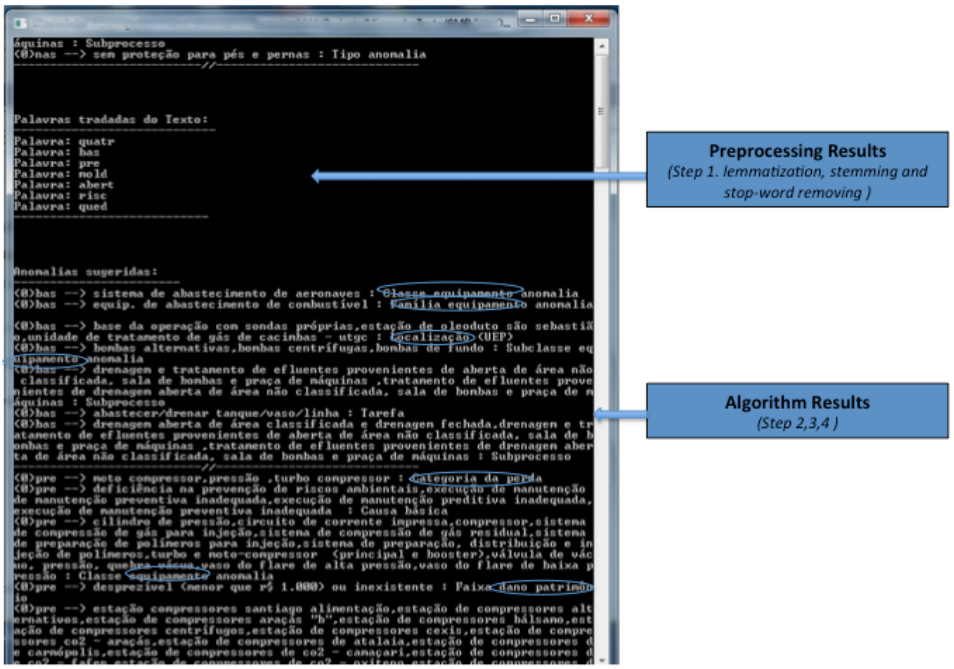


**Figure 5.** Preprocessing and Algorithm results.

In the proposed approach, the ontology effectively becomes the classifier. Consequently, classifier training with a set of pre-classified documents is not needed, as the ontology already includes all important facts. The proposed approach requires a transformation of the document text into a graph structure, which employs entity matching and relationship identification. The categorization is based on measuring the semantic similarity between the created graph and categories defined in the ontology require a training set of pre-classified documents that is used for classifier training; later, the classifier can correctly assign categories to other, previously unseen texts.

In subsequent steps we intend to combine this algorithm with other machine learning approaches like Naïve Bayes, Support Vector Machines, etc. This combination will be capable of yielding a more general solution that is able to cope with situations that fall outside the scope covered by the ontology while preserving the accuracy of the ontology based classification.

## REFERENCES


Lewis, D.D.: Naive (Bayes) at forty: The independence assumption in information retrieval. ECML-98, 10th European Conference on Machine Learning, Chemnitz, DE (1998)

Vapnik, V.: The nature of statistical learning theory. Springer Verlag (1995)

Deerwester, S., Dumais, S., Furnas, G.W., Landauer, T.K., Harshman, R.: Indexing by Latent Semantic Analysis. Journal of the Society for Information Science (1990) 41 (1990) 391-407

Sebastiani, F.: Machine learning in automated text categorization. ACM Computing Surveys (CSUR) 34 (2002) 1 – 47

Bloehdorn, S., Hotho,A.: Text Classification by Boosting Weak Learners based on Terms and Concepts. 4th IEEE International Conference on Data Mining (ICDM'04) (2004)

Nagarajan, M., Sheth, A.P., Aguilera, M., Keeton, K., Merchant, A., Uysal, M.: Altering Document Term Vectors for Classification - Ontologies as Expectations of Co-occurrence. LSDIS Technical Report (November, 2006)

Gabrilovich, E., Markovitch, S.: Overcomingthe Brittleness Bottleneck using Wikipedia: Enhancing Text Categorization with Encyclopedic Knowledge. 21th National Conference on Artificial Intelligence, Boston, MA, USA (2006)

Wu, S.-H., Tsai, T.-H., Hsu, W.-L.: Text categorization using automatically acquired domain ontology. 6th international workshop on Information retrieval with Asian languages - Volume 11, Sapporo, Japan (2003)

Sheth, A.P., Bertram, C., Avant, D., Hammond, B., Kochut, K.J., Warke, Y.: Semantic Content Management for Enterprises and the Web. IEEE Internet Computing July/August 2002 (2002)

Hammond, B., Sheth, A.P., Kochut, K.J.: Semantic Enhancement Engine: A Modular Document Enhancement Platform for Semantic Applications over Heterogeneous Content. Real World Semantic Web Applications, IOS Press, 2002 (2002)

Gruber, T.: A Translation Approach to Portable Ontology Specifications. Knowledge Acquisition 5 (1993) 199-220, 1993

Buccella, A., Cechich, A., Brisaboa, N. R.: Ontology-Based Data Integration. In: Rivero, L.C., Doorn, J.H., Ferraggine, V.E. (eds.): Encyclopedia of Database Technologies and Applications. Information Science Reference (2005)

Sheth, A.P., Arpinar, I.B., Kashyap, V.: Relationships at the Heart of Semantic Web: Modeling, Discovering, and Exploiting Complex Semantic Relationships. In: Nikravesh, M., Azvin, B., Yager, R., Zadeh, L. (eds.): Enhancing the Power of the Internet: Studies in Fuzziness and Soft Computing. Springer Verlag (2003)

Gospodnetic, O.; Hatcher, E., McCandless M.: Lucene in Action (2nd ed.). Manning Publications. ISBN 1-9339-8817-7 (2009).

-- DicSin: Dicionário de Sinônimos Protuguês Brasil. Apache OpenOffice.org http://extensions.openoffice.org/en/project/DicSin-Brasil (2013)