# INFORMATION EXTRACTION TECHNIQUES FOR HEALTH, SAFETY AND ENVIRONMENT APPLICATIONS IN OIL INDUSTRY

Nayat Sanchez-Pi
*ADDLabs. Computer Science Department. Universidade Federal Fluminense*
*Rua General Milton Tavares de Souza, s/nº. 24210-340 / Boa Viagem - Niterói RJ. Brasil*

Luis Marti Orosa
*Electrical Engeenering Department. Pontificia Universidade Catolica*
*RJ. Brasil **

Ana Cristina Bicharra Garcia
*ADDLabs. Computer Science Department. Universidade Federal Fluminense*
*Rua General Milton Tavares de Souza, s/nº. 24210-340 / Boa Viagem - Niterói RJ. Brasil*

## ABSTRACT

The accident investigation process in oil industry is a critical activity. In accident investigation, the volume of information collected and analyzed, makes analysis of its causes a challenging task. Making sense of its large volume of information is also a challenging task because of the diverse interpretations made by experts and auditors. With the advances of new technologies, data of these systems have become increasingly huge. 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

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 OH & S risks, consistent with its policy and objectives of OHS. The development of automatic methods to produce structured information from unstructured text sources would be extremely valuable to the oil industry.

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.

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: i)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), ii)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), iii)Set of ontology elements tagged with its corresponding classification label. iv)A thesaurus crawling algorithm that takes care of determining the matching degree of text words with a corresponding ontology term.

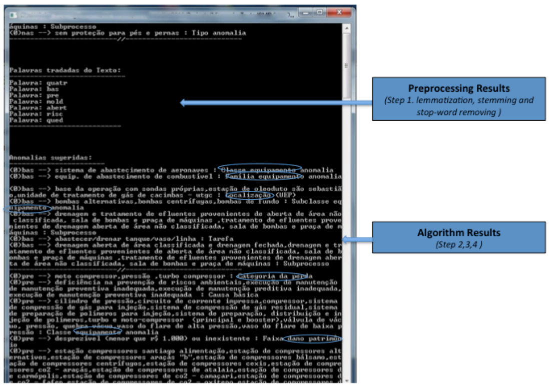
We built domain ontology for the Health, Safety and Environment (HSE) of oil and gas domain. We also obtain the inferences that describe the dynamic side and finally we group the inferences sequentially to form tasks. As mentioned before, the classification algorithm proposed in this work relies on the 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 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() . 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.

# 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 1.



```
áquinas : Subprocesso
<0>nas --> sem proteção para pés e pernas : Tipo anomalia
--------------------------//------------------------------



Palavras tradadas do Texto:
----------------------------
Palavra: quatr
Palavra: bas
Palavra: pre
Palavra: mold
Palavra: abert
Palavra: risc
Palavra: qued
---------------------------



Anomalias sugeridas:
------------------------
<0>bas --> sistema de abastecimento de aeronaves : Classe equipamento anomalia
<0>bas --> equip. de abastecimento de combustível : Família equipamento anomalia

<0>bas --> base da operação com sondas próprias,estação de oleoduto são sebastiã
o,unidade de tratamento de gás de cacimbas - utgc : Localização (UEP)
<0>bas --> bombas alternativas,bombas centrífugas,bombas de fundo : Subclasse eq
uipamento anomalia
<0>bas --> drenagem e tratamento de efluentes provenientes de aberta de área não
 classificada, sala de bombas e praça de máquinas ,tratamento de efluentes prove
nientes de drenagem aberta de área não classificada, sala de bombas e praça de m
áquinas : Subprocesso
<0>bas --> abastecer/drenar tanque/vaso/linha : Tarefa
<0>bas --> drenagem aberta de área classificada e drenagen fechada,drenagem e tr
atamento de efluentes provenientes de aberta de área não classificada, sala de b
ombas e praça de máquinas ,tratamento de efluentes provenientes de drenagem aber
ta de área não classificada, sala de bombas e praça de máquinas : Subprocesso
--------------------------//------------------------------
<0>pre --> moto compressor,pressão ,turbo compressor : Categoria da perda
<0>pre --> deficiência na prevenção de riscos ambientais,execução de manutenção
de manutenção preventiva inadequada,execução de manutenção preditiva inadequada,
execução de manutenção preventiva inadequada  : Causa básica
<0>pre --> cilindro de pressão,circuito de corrente impressa,compressor,sistema
de compressão de gás para injeção,sistema de compressão de gás residual,sistema
de preparação de polímeros para injeção,sistema de preparação, distribuição e in
jeção de polímeros,turbo e moto-compressor  (principal e booster),válvula de vác
uo, pressão, quebra vácuo,vaso do flare de alta pressão,vaso do flare de baixa p
ressão : Classe equipamento anomalia
<0>pre --> desprezível (menor que r$ 1.000) ou inexistente : Faixa dano patrimôn
io
<0>pre --> estação compressores santiago alimentação,estação de compressores alt
ernativos,estação de compressores araçás "b",estação de compressores bálsamo,est
ação de compressores centrífugos,estação de compressores cexis,estação de compre
ssores co2 - araçás,estação de compressores de atalaia,estação de compressores d
e carmópolis,estação de compressores de co2 - camaçari,estação de compressores d
```

Preprocessing Results
*(Step 1. lemmatization, stemming and stop-word removing )*

Algorithm Results
*(Step 2,3,4 )*

**Figure 1.** 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

# ACKNOWLEDGEMENT

Part of this work is funded by CNPq BJT Project 407851/2012-7.

# 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

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)