The Role of Domain Ontology in Text Mining Applications: The ADDMiner Project

Ana Cristina B. Garcia, Inhaúma Neves Ferraz and Fernando Pinto
Universidade Federal Fluminense
bicharra@ic.uff.br; ferrazl@addlabs.uff.br; fernando@addlabs.uff.br

Abstract

Extracting insights from large text collections is an aspiration of any organization aiming to take advantage of their experience generally documented in textual documents. Textual documents, either digital or not, have been the most common form to register any organization transaction. Free text style is a very easy way to input data since it does not require users any special training. On the other hand, the text material easily collected becomes the major challenge for building automatic deciphering tools. In this paper we present ADDMiner, a text-mining model for extracting causality relationships from a large text collection of accident reports. Our model is based on using domain ontology as well as a corpus-based computational linguistics to guide the mining process. Examples from offshore oil platform accident reports illustrate the potential benefits of our approach.

1. Introduction

Offshore petroleum platform operation is a high-risk activity with an extremely high economic return. Accidents are frequently accidents due to the intrinsic danger of dealing with great amounts of combustive material in high pressure. In order to minimize the risks, accident reports, containing a description of the accident and the measures taken to solve it, help any petroleum organization to learn from history. Generally, the industry records the accident history in online textual documents. This rich material becomes almost worthless if not properly compiled. As the report collection grows, making sense of the information inside becomes an impossible task for human brains. It calls for an automatic answer. In this context, text mining represents a promising approach to deal with it. Although text-mining techniques have not yet provided conclusive results for general-purpose mining, a domain-specific application may have different results. The problem consists in extracting causal-effect relation in accident report documents.

This paper discusses the use of domain ontology to allow eliciting cause-effect relations in a large collection of accident report textual documents in oil offshore platforms.

2. Accident report domain

In Brazilian petroleum industry, offshore drilling and production processes are the predominant activity since most reservoirs lay in offshore areas. There are thirty-nine oil fields mostly located in Rio de Janeiro state. These oil fields are explored by sixty-four oil platforms, operated by forty thousand workers. The considerable high numbers of professionals involved allied to the nature of oil platform operation configure a high-risk operation economical, environmental and human-related.

One of the requirements to let a platform operate is the existence of a method to register accidents (or even incidents), including information describing people involved, consequences to the unity as well as the way it was solved and future actions to prevent recurrence. Generally, textual documents are created contemplating this requirement.

Although these reports are available electronically, very little can be done to consolidate all information included in them. The information in the anomaly treatment report is not structured. There is no database with clear attributes that would allow extract accident historic and analysis stored in. For making statistic analysis, figuring out the real cause of the accidents and correlation between platform measurements and accidents, the company need to hire experts that would careful read and make sense of each report and try to consolidate the information they found in those reports. If the number of reports was small, a human expert can take care of this job. However, since the amount of reports is huge and growing, automatic approach to this job seems to be the feasible approach. In this context, using ontology and text mining through the ADDMiner model presents as a promising approach to deal with our problem.

3. ADDMiner Model

ADDMiner, as illustrated in Figure 1, is divided in four main blocks:

• Natural Language Processing block: it represents the linguistic treatment to summarize each textual report into a set attribute-value pair. The text in each report is considered as a set of sentences. On the other hand, each sentence becomes a set of ordered words that will be identified using a lexicon indexed by stems and, syntactic and semantically classified. Finally, a parser syntactically analyses the sentences and builds a parsing tree for each sentence that will guide the semantic processing. As described, this is almost a classic natural language processing with some nuances. ·

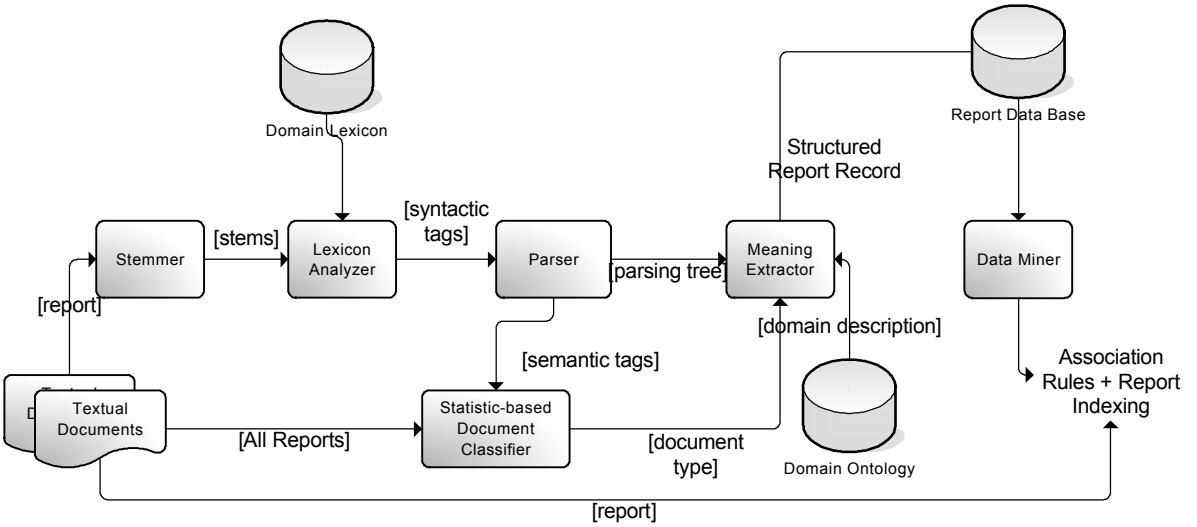


Figure 1: The ADDMiner Model.

The lexicon analysis uses a stemmer to preprocess the words and reduce the lexicon size. Furthermore, although it is desirable to have a previous syntactic processing to facilitate text understanding, the semantic extraction block can recover from parser failures.

• Statistic Classification block[1]: each report document was classified according to the type of accident its content reports. In order to build the classifier, we selected a set of reports and manually classified them in 15 different accident types according to our understanding from reading the report. The reports were statistically treated to remove worthless words and later to identify the words that represented each report.

• Meaning extractor block: instead of taking a general approach, we consider that meaning is context dependent. We developed an ontology for the domain of offshore oil platform accident report, as illustrated in Figure 2. The ontology works as a guide to search for content in a given accident report. ·

• Data Mining block: it represents the data mining process using association rules technique[2]

The use of a domain ontology is the key of our approach. As shown in Figure 2, an accident report contains: ·.

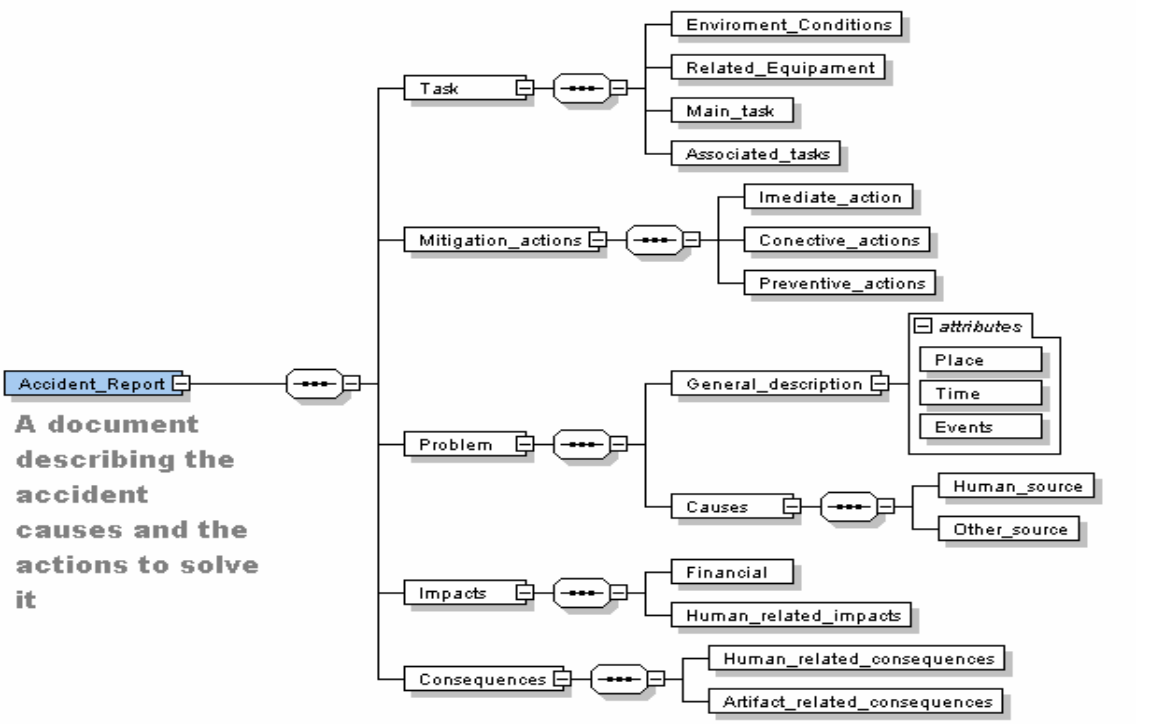


Figure 2: A sample of ADDMiner´s accident report ontology.

Information on the task during which the accident happened including the environment conditions, list of equipment involved, the main task as well as the associated tasks; ·information on the actions taken to mitigate the problem including immediate actions as well as definitive corrective action and related preventive actions to prevent recurrence; · information on the problem itself including a general problem description of when, where and how the accident happened[3], as well as information on the sources of the causes; · Information on the impacts both financial and human-related; and finally Information on the consequences brought[4] by the accident both to the artifact (the oil platform) and human-related (people that works in the oil platform)

4. An Example of using ADDMiner in the Petroleum Accident Report Domain

As an example, we present a typical case of text mining in our accident report data set domain.

Both the algorithm and the software are still under development. The data input is a collection of flat text, one for every Anomaly Report

Each report contains a set of sentences and each sentence a set of words. Our lexicon is concise, for this reason we used a stemming processor to reduce each word into a stem (token) that can be found in the lexicon. For example the word accidentado becomes accident + ado (token + suffix). Token accident is used as an index to recuperate syntactic and semantic information stored in the lexicon. Syntactic information is used to help the tokenizer/stemmer process.

All, tokens and semantic information are processed under the statistic analyzer which recognize the anomaly type of each report. This statistic analyzer was previously trained with a sample of the domain. Such process is illustrated in Figure 3 and it was classified/typified as Accident with Injury Report Type.

Figure 3: Anomaly type process recognition.

As each anomaly (accident) type are recognized for every report, we switch to the corresponding ontology to execute the adequate ontologic information extraction, for our sample it was automatically choosed the Accident with Injury Ontology.

We have developed a domain ontology describing all 15 types of anomalies/accidents that may occur operational fault, accident with injury and machine break.

Once statistic characterization of the Anomaly Type is done, the text is passed to the Information Analyzer block which uses techniques of information extraction over the selected ontology, this block is the KEY of our process and it is shown on Figure 5.

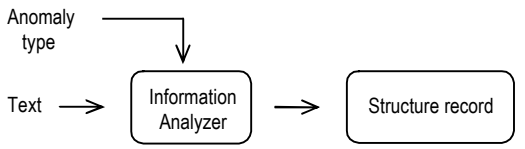


Figure 5: Information extraction.

The information extracted on this phase guides the database attributes filling. Database modeling was based on ontologies considered. This information extraction uses some sort of grammatical and semantic treatment for finding the corresponding concepts described in the ontology.

Next, as illustrated in Figure 6, database is processed to find Association Rules guiding the rules only with the fact that our objective is to find Source of Anomalies. This is done by filtering rules with possible reasons of faults, injuries accordingly with ontology.

As an example of a rule, extracted in this phase, we have

$$\text{stairs} + \text{steel} + \text{making hole} \Rightarrow \text{injury}$$

with a support x and confidence y.

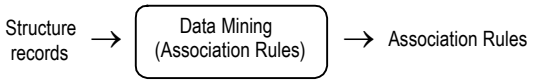


Figure 6: Association Rules Generation.

The final phase is to show rules to user as structured and graphical information. Again, this is based on the selected ontology.

In figure 7 it its shown rule number 31, and some of its properties, for specific document. In central bottom part, it is shown all documents that meets rule 31 (7 documents). Second one was selected and visualized on bottom right part. At right top is shown the rule with all the corresponding attributes adequately presented under the current ontology. The objective of showing bold colored tags is to indicate that they were filled automatically by the program and that it is going to color, the corresponding text in document, the same as tags; right now, it is shown (text in document), when double clicking each property. In this case, it was selected description of activity property.

This way it was found the following properties (ontology components) already on block processing in Figure 5.

Patient (paciente) = Nilceu Mario Moro
Company (Empresa) = ATM
Injury (lesão) = corte contuso / escoriação
Body part (parte do corpo) = punho da mão esquerda
Activity (atividade) = furava una antepara de aço para fixação de um suporte de ferramentas no local
Injury reason (razão da lesão) = Ao vazar a parede, a broca quebrou, o empregado desequilibrou-se
Activity type (classe de atividade) = Parte de atividade
Activity schedule (horário de atividade) = Durante o trabalho
Immediate action (ação imediata) = Enfermaria

Also it was also found as the probably causes for this accident as lack of victim attention and environment imperfection.

5. Discussion

Most researches on text mining focus on developing broad general-purpose technologies to improve web text document retrieval. Since our objective is to answer a well defined question: "What is causing accidents?," we could take a domain-dependent approach when developing a tool to process the domain data source. This paper presented an approach to reveal cause-effect information buried in textual accident report document files.

The text mining question can be understood as three sub-questions[5]:

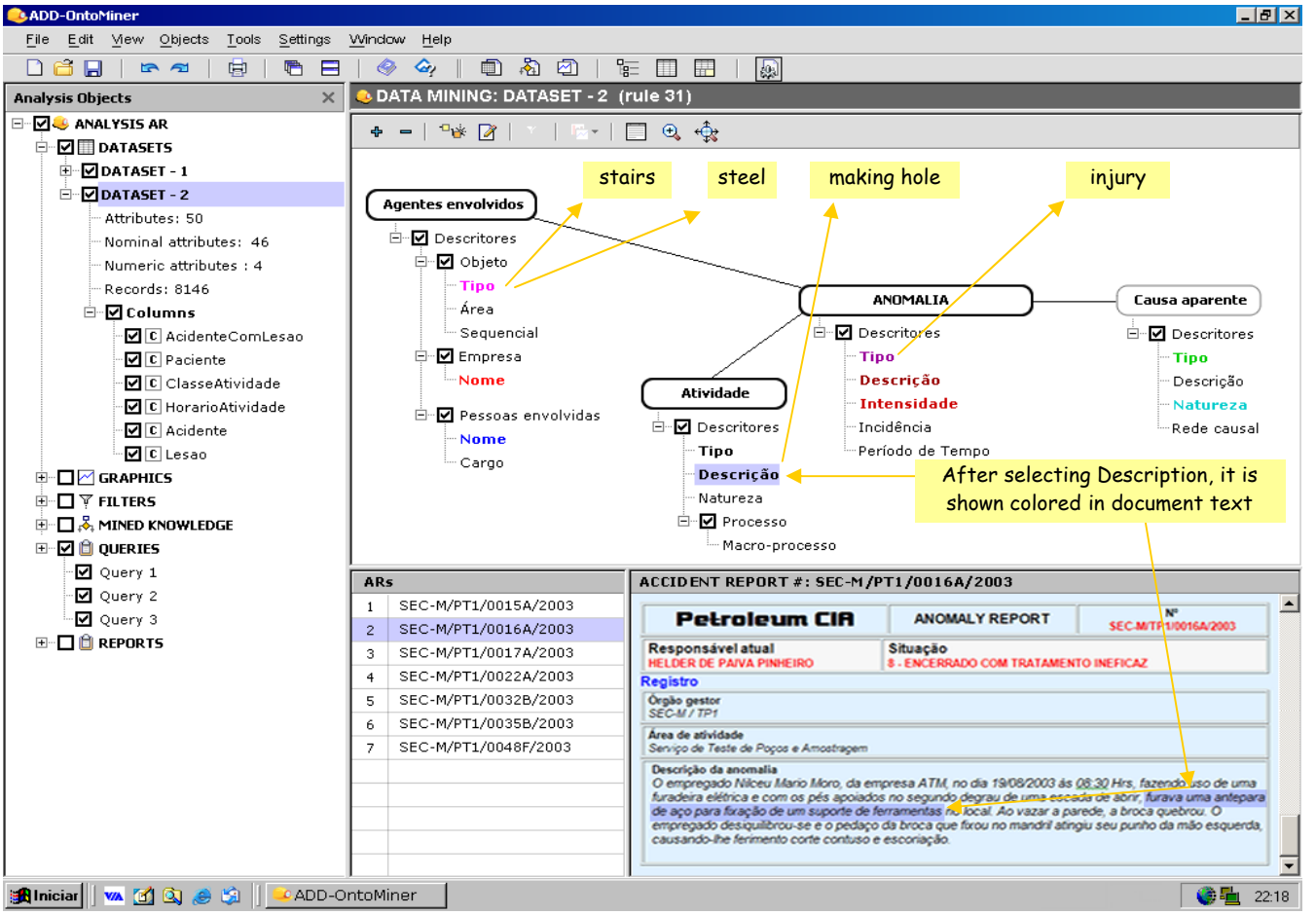
- What is written in a accident report? Is there any structured in the storytelling style that can guide a report understanding?
- What information is expected to be provided when describing an accident?
- Is it possible to draw cause-effect inferences from the reported accidents? Is each case unique?

The first question was addressed by using a natural language processor that combines a stemmer to reduce the size of the domain lexicon, combined with a parser that deal with incomplete information.

The second question was addressed by including a domain ontology[6] describing what should be in an accident report (the touch of domain-dependency approach). The ontology guided the semantic processing by providing an expectation and guidance of what should be looked for in the text.

The third question was addressed by an association rule data miner with pos-processing to prune the output. Rule visualization and the ability to retrieve accident report sample that complies with the rule are the most effective pos-processing technique considered here.

We developed a tool according to ADDMiner model that has been implemented in C++ showing the feasibility of our approach.



6. References

[1] Usama Fayyad and Ramasamy Uthurusamy. Data mining and knowledge discovery in databases: Introduction to the special issue. Communications of the ACM, 39(11), November, 1999.

[2] Marti Hearst. Untangling text data mining. Proceedings of ACL'99: the 37th Annual Meeting of the Association for Computational Linguistics. 1999.

[3] Ricardo Baeza-Yates and Berthier Ribeiro-Neto. Modern Information Retrieval. Addison-Wesley Longman Publishing Company. 1999.

[4] Ouzounis, C. A., TEXTQUEST: Document clustering of MEDLINE
abstracts for concept discovery in molecular biology. PSB 2001, pp. 384–395, 2001.

[5] Glenisson, P., Antal, P., Mathys, J., Moreau, Y. & De Moor, B., Evaluation of the Vector
Space Representation in Text-Based Gene Clustering. PSB 2003, pp. 391-402, 2003.

[6] H. D. White and K. W. McCain. Bibliometrics. Annual Review of Information Science and Technology, 24:119–186, 1989.