Journal of Applied Logic 17 (2016) 48–58

Contents lists available at ScienceDirect

Journal of Applied Logic

www.elsevier.com/locate/jal

Improving ontology-based text classification: An occupational health and security application

Nayat Sanchez-Pi [a,*], Luis Martí [b], Ana Cristina Bicharra Garcia [b]

[a] *Institute of Mathematics and Statistics, Rio de Janeiro State University, Rio de Janeiro, RJ, Brazil*
[b] *Institute of Computing, Fluminense Federal University, Niterói, RJ, Brazil*

ARTICLE INFO

Article history:
Available online 28 September 2015

Keywords:
Text classification
Ontology
Oil and gas industry

ABSTRACT

Information retrieval has been widely studied due to the growing amounts of textual information available electronically. Nowadays organizations and industries are facing the challenge of organizing, analyzing and extracting knowledge from masses of unstructured information for decision making process. The development of automatic methods to produce usable structured information from unstructured text sources is extremely valuable to them. Opposed to the traditional text classification methods that need a set of well-classified trained *corpus* to perform efficient classification; the ontology-based classifier benefits from the domain knowledge and provides more accuracy. In a previous work we proposed and evaluated an ontology-based heuristic algorithm [28] for occupational health control process, particularly, for the case of automatic detection of accidents from unstructured texts. Our extended proposal is more domain dependent because it uses technical terms and contrast the relevance of these technical terms into the text, so the heuristic is more accurate. It divides the problem in subtasks such as: (i) text analysis, (ii) recognition and (iii) classification of failed occupational health control, resolving accidents as text analysis, recognition and classification of failed occupational health control, resolving accidents.

1. Introduction

The need for effective methods of automated Information Retrieval has grown during years because of the amount of unstructured data in natural language form generated in modern organizations [4]. There is a need of performing analysis, decision-making, and knowledge management tasks using this unstructured information.

* Corresponding author.
E-mail address: nayat@ime.uerj.br (N. Sanchez-Pi).

http://dx.doi.org/10.1016/j.jal.2015.09.008
1570-8683/

Nontraditional Information Retrieval strategies are: *text mining* that uncovers previously invisible patterns in existing resources and *text classifications* that is a subfield of data mining which refers generally to the process of deriving high quality of information from a text [7].

Automatic text classification is a task of assigning one or more pre-specified classes to a text, based on its content. Text classification techniques are used in many applications, including e-mail filtering, mail routing, spam filtering, news monitoring, sorting through digitized paper archives, automated indexing of scientific articles, classification of news stories and searching for interesting information on the Web, biomedical applications [8], etc. A good survey of hybrid classifiers systems can be found at [33].

However, it is often the case that a suitable set of well classified trained corpus is not available. Even if one is available, the set may be too small, or a significant portion of the corpus in the training set may not have been classified properly. This creates a serious limitation for the usefulness of the traditional text classification methods.

Our proposal is to use the background knowledge represented by means of an ontology. In the area of computing, the ontological concepts are frequently regarded as classes which 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 [30].

In previous work we propose and evaluate an ontology-based heuristic algorithm [28] for occupational health control process, particularly, for the case of automatic detection of accidents from unstructured texts. Our extended proposal is more domain dependent because it uses technical terms and contrast the relevance of these technical terms into the text, so the heuristic is more accurate. It divides the problem in subtasks such as: (i) text analysis, (ii) recognition and (iii) classification of failed occupational health control, resolving accidents.

The rest of this manuscript goes on by describing the theoretical foundations that support it. After that, in Section 3, we describe the elements that are involved in our proposal: (i) the elaboration of the ontology, (ii) the use of a thesaurus as a crawling tool, (iii) the use of the ontology as a classifier, (iv) the compensated classifier using techniques terms. Section 4 proposes an oil and gas industry application scenario: occupational health and security where some comparative experiment are presented. Finally, Section 5 presents some final remarks.

2. Foundations

Due to the ever growing amounts of textual information available electronically, organizations are facing the challenge of organizing, analyzing and extract knowledge from masses of unstructured information for decision making process.

Traditional classification approaches use statistical or machine learning methods to perform the task. Such methods include Naïve Bayes [22], Support Vector Machines [32], Latent Semantic Analysis [9] and many others. A good overview of the traditional text classification methods is presented in [29]. 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.

During the last decades, a large number of machine learning algorithms have been proposed for supervised and unsupervised text categorization. So far, however, existing text categorization systems have typically used the Bag-of-Words model where single words or word stems are used as features for representing document content [26].

However, the work on integrating semantic background knowledge into text categorization is still quite scattered. Early works are: [3,13,6,18]. They use WordNet [11] to improve the text clustering task. WordNet is a network of related words, organized into synonym sets, where each of the sets represents one lexical underlying concept. WordNet has been successfully used both in text categorization and clustering [25].