An Effective TF/IDF-Based Text-to-Text Semantic Similarity Measure for Text Classification

Shereen Albitar, Sébastien Fournier, and Bernard Espinasse

Aix-Marseille University, LSIS UMR CNRS 7296
Domaineuniversitaire de St. Jerome, 13397 Marseille Cedex 20, France
first_name.last_name@lsis.org

Abstract. The use of semantics in tasks related to information retrieval has become, in recent years, a vast field of research. Considering supervised text classification, which is the main interest of this work, semantics can be involved at different steps of text processing: during indexing step, during training step and during class prediction step. As for class prediction step, new text-to-text semantic similarity measures can replace classical similarity measures that are traditionally used by some classification methods for decision-making. In this paper we propose a new measure for assessing semantic similarity between texts based on TF/IDF with a new function that aggregates semantic similarities between concepts representing the compared text documents pair-to-pair. Experimental results demonstrate that our measure outperforms other semantic and classical measures with significant improvements.

Keywords: Classification, Semantics, Text-to-Text Semantic Similarity.

1 Introduction

Supervised text classification is currently a challenging research topic, particularly in areas such as information retrieval, recommendation, personalization, user profiles etc. Generally, supervised text classification methods use syntactical and statistical models for text document representation. This applies to the most popular text classification methods such as: Naïve Bayes Classifier (NB), Support Vector Machines (SVMs), Rocchio, and k Nearest Neighbors (kNN). These representation models ignore all semantics that reside in the original text that can help in text classification.

However, it is possible to use semantic resources to take into account meaning of the words in text representation in order to improve classification effectiveness. Thus, resulting text representation models can take into account synonyms, relations between words and also can resolve some ambiguities. Many researchers reported that using semantics in text classification improves its effectiveness in specific domains especially by deploying domain specific semantic resources [1].

There are several possibilities for involving semantics during the process of supervised texts classification. In this work, we are interested in involving semantics in class prediction step using, text-to-text semantic similarity measures. Hence, we propose a new text-to-text semantic similarity measure (TF/IDF based), called in this

B. Benatallah et al. (Eds.): WISE 2014, Part I, LNCS 8786, pp. 105–114, 2014.

article SemTFIDF, and we present an experimental study to evaluate it in the context of text classification. In addition, we compare it with another text-to-text semantic similarity measure proposed in the literature (IDF based) called semIDF in this article, and also with the well-known classical similarity measure Cosine that is usually deployed in the Vector Space Model. These experiments are carried out in the biomedical domain using the Ohsumed corpus and domain specific knowledge base Unified Medical Language System (UMLS®) and Rocchio with Cosine as the baseline [2].

Second section reviews state of the art methods deploying semantics in classification or other tasks related to information retrieval or data mining. Third section focuses on the use of semantics during class prediction step and presents our new measure (SemTFIDF) based on TF/IDF and suitable for supervised text classification. Fourth section presents experimental setup that we used to evaluate our new measure. Fifth section analyses the experimental results obtained with Cosine classical similarity measure and these two text-to-text similarity measures (SemIDF and SemTFIDF). Finally, we conclude and present our perspectives for future works.

2 Involving Semantics in Supervised Text Classification

Typically, most of supervised text classification techniques are based on statistical and probabilistic hypothesis in both training and classification procedures. As for text representation or indexing, the importance of a term to a document is assessed using the frequency of its occurrences in the document. So far, the intended meaning of terms and the relations among them are not treated or used in text classification. In other words, semantics and relatedness behind literally occurring words are missing in classical text classification techniques. However, last few years have seen different approaches seeking to introduce semantics during indexing, training and prediction.

Involving Semantics in Indexing. Semantics can be used during indexing for a semantic text representation. Indeed, vector-based (binary or TF/IDF) representations, used by these classical supervised classification methods, enable semantic integration or "conceptualization" that enriches document representation model using background knowledge bases [1, 3].To involve semantic features in indexing, state of the art approaches used either implicit semantics through topic modeling[4] or explicit semantics derived from structured resources and used as new features for text representation[1, 6]. Other approaches use either type in semantic kernels to support some supervised classification techniques [5].

Involving Semantics in Training. In these approaches, concepts replace words in text representation. In addition, the hierarchy and the relations among the added concepts are taken into consideration in the training step which affects the learned model, so the classification model is either the entire ontology or part(s) of its hierarchy. Both works [7, 8] used the hierarchical structure of semantic resources to involve related concepts in text representation. Authors in [8] used propagation algorithm to propagate the weights of identified concepts in patents to their superconcepts. Furthermore, authors in [9] used similar concepts in order to enriched text representation and proposed the approach Enriching vectors. Similarities among concepts are assessed using relations between concepts in the semantic resource. Both Generalization [7, 8] and Enriching vectors [9] involve semantics in the classification model implicitly.