Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 112 (2017) 822–832

Procedia Computer Science

www.elsevier.com/locate/procedia

International Conference on Knowledge Based and Intelligent Information and Engineering Systems, KES2017, 6-8 September 2017, Marseille, France

Crowd-based Feature Selection for Document Retrieval in Highly Demanding Decision-making Scenarios

Julliano Trindade Pintas[a], Luís Correia[b], Ana Cristina Bicharra Garcia[a,*]

[a] *Universidade Federal Fluminense, Rua Passos da Patria, 156 Niteroi, RJ, Brazil*
[b] *BioISI, Universidade de Lisboa, Lisbon, Portugal*

Abstract

Automatic dimensionality reduction in text classification requires large training data sets due to the high dimensionality of the native feature space. However, in several real world multi-label problems, such as highly demanding decision-making scenarios, to manually classify and select features in large document sets is usually unfeasible even by specialist teams. This paper presents CrowdFS a first approach on using collective intelligence techniques to select label specific relevant features from a large document set. An experiment in the context of competitive intelligence for a multinational energy company showed CrowdFS producing better results than an automatic state of the art technique.

Keywords: dimensionality reduction; crowd; collective intelligence; document retrieval; business intelligence

1. Introduction

Companies, which operate in highly competitive and dynamic markets, need to continuously search for information, inside and outside the company. Information, such as the competitors moves or market regulation's changes, leads to insights and perceptions of opportunities, guiding decision-makers in their choices. Large companies have a team of experts with the task of searching the Internet and other media, for news that might impact a specified decision setting. It is an overwhelming task, since new information keeps continuously feeding the web. To aggravate this scenario, the same piece of information contained in a document might serve to different decision settings (multi-labeling). So, tagging a document once does not mean tagging it forever. To improve the information retrieval, each news article could be classified simultaneously on several aspects or categories, such as the relevance, the level and type of impact for the company, according to related rival companies and the areas of the company that can be affected. However, the task of manually analyzing and classifying each available information item according to all different perspectives is usually not feasibly. For this reason, there is a call for effective automatic multi-label text classifiers to support the process of retrieving relevant information for leveraging decision-making scenarios.

[*] Corresponding author. Tel.: +5521-2629-5861
E-mail address: bicharra@ic.uff.br

1877-0509 © 2017 The Authors. Published by Elsevier B.V.
Peer-review under responsibility of KES International
10.1016/j.procs.2017.08.074

A known big challenge for text classification problems is how to deal with the high dimensionality of the feature space[1]. The native feature space of a textual document consists of the unique terms (words or expressions) that compose the document. Tens or hundreds of thousands of terms can be extracted even in medium text datasets[1]. Therefore, the selection of a reduced set of native features is highly desirable to improve any classifiers (a) efficiency, by decreasing the feature input space, and (b) effectiveness (precision and recall), by eliminating noisy features[2]. The mainstream automatic feature selection approaches are based on statistical tests, correlational coefficients and cross-validation using a significant training dataset[3]. Their performance depends on the size and quality of the training set. However in highly demanding decision-making scenarios, the search space (all available documents) is very large and significant pre-labeled training datasets are not commonly available.

A number of studies have already proven that aggregating the judgment of several individuals may result in estimates that are close to the real value in different domains, a phenomenon of collective intelligence known as wisdom of the crowds (WoC)[4]. The popularization of crowd-sourcing platforms and initiatives such as Amazon Mechanical Turk allow for an easy access to WoC. This paper presents a new approach called CrowdFS that applies collective intelligence techniques to support the selection of label specific features for multi-label text classification. Section 2 presents background information, followed by Section 3 that presents CrowdFS. Section 4 presents a quantitative experiment conducted on a multinational energy company to evaluate the proposed approach. The results analysis demonstrates, in section 5, the feasibility of this approach and its usefulness mainly in small labeled training sets scenarios. The challenges, issues and future work related to this new term selection approach are raised and discussed in section 6.

2. Background and Related Work

Our research is mainly related to three well known areas:

- feature selection studies,
- multi-label text learning studies and
- collective intelligence studies.

2.1. Feature Selection

In some machine learning related problems whose native feature space can be considerable large, such as text classification problems, the dimensionality reduction of the feature space is desired to improve model performance, facilitate the data understanding and avoid model from overfitting[3,5]. Feature extraction and feature selection are frequently used techniques to reduce the feature space[5]. Feature extraction is the process of generating a reduced new set of features that map to the the original set of features[6]. Feature subset selection technique selects, from the original set, the features that better represent the document[7,8]. The approach presented in this paper is a based on the second technique, feature subset selection.

There are three main approaches for feature subset selection[9](some authors[10] only consider the first two approaches):

- Filter methods use specific metrics, such as information gain, chi-square and the correlation coefficient, to evaluate each feature for further selecting the best-scored ones;
- The wrapped approach iteratively trains and evaluates classifier models with different sets of features, optimizing some objective function, such as coverage, precision, or f-measure.
- Embedded methods include variable selection as part of the training process without splitting the data into training and testing sets.

Typical filter, wrapper and embedded methods use pre-classified documents to evaluate the relevance of each feature. Also, wrapper and embedded methods use classifiers whose performance is highly dependent on the amount and quality of the training set. Hence the aforementioned feature selection approaches are dependent on the same factors and as a consequence feature selection results are generally modest when the training data sets are small. The