Information Systems 38 (2013) 388–409

Contents lists available at SciVerse ScienceDirect

Information Systems

journal homepage: www.elsevier.com/locate/infosys

Temporal contexts: Effective text classification in evolving document collections

Leonardo Rocha [a,*], Fernando Mourão [b], Hilton Mota [c], Thiago Salles [b], Marcos André Gonçalves [b], Wagner Meira Jr. [b]

[a] Federal University of São João del-Rei, Computer Science Department—São João del-Rei, Brazil
[b] Federal University of Minas Gerais, Computer Science Department—Belo Horizonte, Brazil
[c] Federal University of Minas Gerais, Electrical Engineering Department—Belo Horizonte, Brazil

ARTICLE INFO

Article history:
Received 21 October 2011
Received in revised form
29 October 2012
Accepted 8 November 2012
Recommended by: T.G.K. Calders
Available online 21 November 2012

Keywords:
Classification
Text mining
Temporal evolution

ABSTRACT

The management of a huge and growing amount of information available nowadays makes Automatic Document Classification (ADC), besides crucial, a very challenging task. Furthermore, the dynamics inherent to classification problems, mainly on the Web, make this task even more challenging. Despite this fact, the actual impact of such temporal evolution on ADC is still poorly understood in the literature. In this context, this work concerns to evaluate, characterize and exploit the temporal evolution to improve ADC techniques. As first contribution we highlight the proposal of a pragmatical methodology for evaluating the temporal evolution in ADC domains. Through this methodology, we can identify measurable factors associated to ADC models degradation over time. Going a step further, based on such analyzes, we propose effective and efficient strategies to make current techniques more robust to natural shifts over time. We present a strategy, named temporal context selection, for selecting portions of the training set that minimize those factors. Our second contribution consists of proposing a general algorithm, called *Chronos*, for determining such contexts. By instantiating Chronos, we are able to reduce uncertainty and improve the overall classification accuracy. Empirical evaluations of heuristic instantiations of the algorithm, named *WindowsChronos* and *FilterChronos*, on two real document collections demonstrate the usefulness of our proposal. Comparing them against state-of-the-art ADC algorithms shows that selecting temporal contexts allows improvements on the classification accuracy up to 10%. Finally, we highlight the applicability and the generality of our proposal in practice, pointing out this study as a promising research direction.

Contents

* Corresponding author. Tel.: +55 32 3373 3985.
E-mail address: lcrocha@ufsj.edu.br (L. Rocha).

0306-4379/$ - see front matter © 2012 Elsevier Ltd. All rights reserved.
http://dx.doi.org/10.1016/j.is.2012.11.001

1. Introduction

The widespread use of the Internet has increased the amount of information being stored and accessed through the Web in a very fast pace. This information is frequently organized as textual documents [1] and is the main target of search engines and other retrieval tools, which perform tasks such as searching and filtering. A common strategy to deal with this information is to associate it with semantically meaningful categories, a technique known as Automatic Document Classification (ADC) [2]. This automatic document class assignment can support and enhance several tasks, such as automated topic tagging [3], digital library creation [4], and improvement of Web searching precision [5].

ADC usually employs a supervised learning strategy, in which a classification model is first built using pre-classified documents, i.e., a training set, and this model is then used to classify unseen documents. Building text classification models consists of finding and weighting a set of features (e.g., terms) that help to identify classes of documents. The concept is illustrated in Fig. 1.

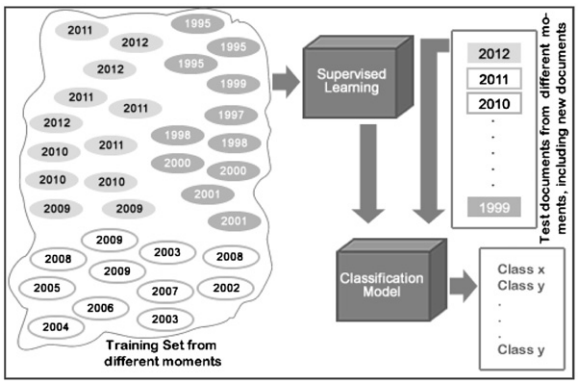

Fig. 1. Traditional supervised learning strategy. The whole training set, composed of documents from distinct moments, is given to a supervised learning technique. Then, it generates a classification model that infers classes for each test document, disregarding its creation moment.

In [6], it is advocated that time is an important dimension of any information space and may be very useful in information retrieval. Despite the potential impact of the temporal evolution on the quality of classification models, most of the current techniques for ADC do not consider this evolution while building and using the models. Furthermore, there are some relevant questions related to temporal evolution that are still not clear, three of which we address in this paper, namely:

(1) How does the temporal evolution of the information space affect the performance of the classifiers?
(2) What are the temporal-related characteristics that affect the classification's effectiveness?
(3) How to exploit such characteristics to improve the classification's effectiveness?

The first two questions are addressed in the first part of the paper. We distinguish three temporal effects that may affect the performance of automatic classifiers. The first one, called *class distribution*, is related to the impact of temporal evolution on the distribution of class frequency. It is a consequence of the dynamic evolution of knowledge and the ways to express it, which may result in classes appearing, disappearing, splitting or merging. The second effect, called *term distribution*, refers to how the relationship between terms and classes changes over time. These changes may occur as a consequence of terms appearing, disappearing, and presenting variable discriminative power within classes. The third effect, *class similarity*, concerns how the similarities among classes vary over time, as a function of the terms that occur in their documents. For instance, two classes may be very similar at a given moment, and less similar later in the future.

In order to understand and characterize these three factors more easily, we propose a novel and pragmatical methodology for assessing the impact of the temporal evolution in ADC applicable to distinct domains. Our methodology allows us to analyze each factor separately,