Authorship Analysis based on Data Compression

Daniele Cerra, Mihai Datcu, and Peter Reinartz

German Aerospace Center (DLR), Muenchner str. 20, 82234 Wessling, Germany
Corresponding author's email: daniele.cerra@dlr.de,
phone: +49 8153 28-1496, fax: +49 8153 28-1444.

Abstract

This paper proposes to perform authorship analysis using the Fast Compression Distance (FCD), a similarity measure based on compression with dictionaries directly extracted from the written texts. The FCD computes a similarity between two documents through an effective binary search on the intersection set between the two related dictionaries. In the reported experiments the proposed method is applied to documents which are heterogeneous in style, written in five different languages and coming from different historical periods. Results are comparable to the state of the art and outperform traditional compression-based methods.

Keywords: Authorship Analysis, Data Compression, Similarity Measure

1. Introduction

The task of automatically recognizing the author of a given text finds several uses in practical applications, ranging from authorship attribution to plagiarism detection, and it is a challenging one (Stamatatos, 2009). While the structure of a document can be easily interpreted by a machine, the