Chapter 8
Applications of Machine Learning for Linguistic Analysis of Texts

Rosemary Torney
University of Ballarat, Australia

John Yearwood
University of Ballarat, Australia

Peter Vamplew
University of Ballarat, Australia

Andrei Kelarev
University of Ballarat, Australia

ABSTRACT

This chapter describes a novel multistage method for linguistic clustering of large collections of texts available on the Internet as a precursor to linguistic analysis of these texts. This method addresses the practicalities of applying clustering operations to a very large set of text documents by using a combination of unsupervised clustering and supervised classification. The method relies on creating a multitude of independent clusterings of a randomized sample selected from the International Corpus of Learner English. Several consensus functions and sophisticated algorithms are applied in two substages to combine these independent clusterings into one final consensus clustering, which is then used to train fast classifiers in order to enable them to perform the profiling of very large collections of text and web data. This approach makes it possible to apply advanced highly accurate and sophisticated clustering techniques by combining them with fast supervised classification algorithms. For the effectiveness of this multistage method it is crucial to determine how well the supervised classification algorithms are going to perform at the final stage, when they are used to process large data sets available on the Internet. This performance may also serve as an indication of the quality of the combined consensus clustering obtained in the preceding stages. The authors’ experimental results compare the performance of several classification algorithms incorporated in this multistage scheme and demonstrate that several of these classification algorithms achieve very high precision and recall and can be used in practical implementations of their method.

DOI: 10.4018/978-1-4666-1833-6.ch008
INTRODUCTION

The Internet and email have revolutionised both business and personal communication methods, negating the problems of distance and time-zones Alrawi & Sabry (2009). Although there have always been a small percentage of dubious enterprises that are prepared to prey on unsuspecting customers, in the real world it is usually possible to trace these unscrupulous establishments. The anonymity of the Internet makes this far more difficult. There is no physical location to return to and the victim has not seen or heard the perpetrator to give a description to law enforcement agencies. Criminal elements seem to be relying on the anonymity of cyberspace to protect them while they engage in illegal activities such as scams, phishing and predatory behavior Chaski (2008). However, they must make contact with their victims, and this is usually achieved with some form of text communication. This is where authorship analysis can be applied to extract some details about the identity or profile of the author on the basis of their use of language. It has been discovered, for example, in Abbasi & Chen (2008), Baayen et al. (2002), Chaski (2005), that authors leave a textual “fingerprint” behind in their choice of language. Stylometry or authorship analysis, has been used to determine the authenticity of evidence presented for both the prosecution and defense in USA courts, as reported in Chaski (2008).

The development of automated methods for various aspects of linguistic analysis based on machine learning techniques is one of the major research topics which has been very actively investigated. To illustrate let us refer to just a few recent articles Agarwal et al. (2009), Bao et al. (2009), Bian & Tao (2009), Ikeda et al. (2009), Long et al. (2009), Malik & Kender (2008), Momma et al. (2009), Nakajima et al. (2005), Negi et al. (2009), Ni et al. (2007), Park et al. (2009), Roth et al. (2009), Sindhwani et al. (2008). Clustering of documents based on similar linguistic features often forms an early stage in these automated analysis methods.

Several authors have demonstrated that ensemble clustering approaches can be highly useful for solving various problems, as in Aho & Dzeroski (2009), Domeniconi et al. (2009), Lu et al. (2009), Read (2008). Highly sophisticated and effective consensus functions and heuristics for clustering ensembles have been developed, for example, by Ailon, Charikar and Newman (2005), Fern & Brodley (2004, 2004A), Goder & Filkov (2008), Strehl & Ghosh (2002). However such methods are not practically applicable to the very large number of documents that are often encountered in linguistic analysis tasks.

This article proposes a novel multistage method for linguistic clustering of very large collections of documents available on the Internet. The method is based on creating a multitude of independent initial clusterings of a randomized sample of texts from the International Corpus of Learner English. The ICLE corpus represents a unique collection of essays with detailed authorship information. Two substages of the method apply advanced consensus functions and sophisticated ensemble clustering algorithms to obtain final consensus clustering of the sample, which is then used to train fast supervised classifiers.

The sets of documents available on the Internet are huge and have many possible initial clusterings worth considering. It is computationally infeasible to apply highly sophisticated consensus ensemble clustering algorithms to these enormous data sets directly. Instead, we propose a novel more advanced multistage approach, as shown in Figure 1. First, a multitude of independent clustering algorithms were applied to a randomized sample from the International Corpus of Learner English (ICLE). Second, several consensus functions and sophisticated algorithms were used in two substages to combine these independent clusterings into one final consensus clustering. Third, several fast supervised classification algorithms were trained on the consensus clustering of the ran-
Related Content

Investigations on the Brain Connectivity Parameters for Co-Morbidities of Autism Using EEG
[www.igi-global.com/article/investigations-on-the-brain-connectivity-parameters-for-co-morbidities-of-autism-using-eeg/202954?camid=4v1a](www.igi-global.com/article/investigations-on-the-brain-connectivity-parameters-for-co-morbidities-of-autism-using-eeg/202954?camid=4v1a)

A Value-Based Framework for Software Evolutionary Testing
[www.igi-global.com/article/value-based-framework-software-evolutionary/55129?camid=4v1a](www.igi-global.com/article/value-based-framework-software-evolutionary/55129?camid=4v1a)

Magnetic Resonance Image Analysis for Brain CAD Systems with Machine Learning
[www.igi-global.com/chapter/magnetic-resonance-image-analysis-brain/62234?camid=4v1a](www.igi-global.com/chapter/magnetic-resonance-image-analysis-brain/62234?camid=4v1a)

Fuzzy Logic Classifiers and Models in Quantitative Software Engineering
[www.igi-global.com/chapter/fuzzy-logic-classifiers-models-quantitative/4860?camid=4v1a](www.igi-global.com/chapter/fuzzy-logic-classifiers-models-quantitative/4860?camid=4v1a)