山东大学学报(工学版) 2009, 39(5) 27-31 DOI: ISSN: 0412-1961 CN: 21-1139

机器学习与数据挖掘

作者写作特征提取引擎

董乃鹏 赵合计 SCHOMMER Christoph

董乃鹏，赵合计：山东大学计算机科学与技术学院，山东 济南 250101; SCHOMMER Christoph： 卢森堡大学信息与计算机学院，卢森堡， 2311

摘要：

随着计算机网络的发展,电子文章逐渐繁荣.电子文章版权保护近年来也越来越受关注.电子文章版权保护的一个解决方案是,首先提取一个作者的写作特征,通过写作特征的比较来判断版权所属.目前作者特征提取方向的研究多集中在寻找新的更有效的特征上.如何更加有效的提取一个作者的写作特征仍是一件富有挑战性的工作.本文建立了一个作者特征提取引擎模型,该引擎以某个作者某一类型的文章作为输入,以该作者在这一类型文章上的写作特征为输出.应用这个引擎模型,在可能的作者列表中,可以确定一篇文章倾向属于某个作者的可能性.本文主要对英文文章进行特征提取.作者的特征通过各种语言学上特征和语言学度量来表示,并采用标准差和主成分分析法分析这些特征的有效性.

关键词： 作者特征提取;文本处理;自然语言处理;数据挖掘;人工智能

A fingerprint engine for author profiling

1. Department of Computer Science and Technology, Shandong University, Jinan 250101, China;
2. Department of Information and Computer Sciences, Luxembourg 2311, Luxembourg

Abstract:

With the development of the internet, digital texts are proliferating. Protection of a copyright has become increasingly important in recent years. To solve the copyright problem, one way is to profile an author's writing style. By comparing writing styles, we could tell whether a text has been written by a certain author. Most of the current researches in author profiling focused on examining linguistic attributes or finding new attributes. However, the appropriate profiling of an author is still a challenging task. This paper aims to build a model to fingerprint an author, and took texts of an author of a certain domain as input and produced a profile of the author as output. Using this fingerprint engine we can tell with a certain probability whether an input text has been written by an author among a list of possible authors. This paper focused on author profiling of English texts. Writing styles were measured using linguistic attributes and linguistic measurements. Statistical methods, such as standard deviation analysis and principal components analysis, were used to evaluate the linguistic measurement's efficiency.

Keywords: author profiling; text analysis; natural language processing; data mining; artificial intelligence

收稿日期 2008-12-10 修回日期 网络版发布日期 2009-10-16

DOI:

基金项目:

通讯作者:

作者简介:

作者Email:

PDF Preview

参考文献：

本刊中的类似文章