计算机工程与应用 2009 45 (29): 25-27 ISSN: 1002-8331 CN: 11-2127/TP

博士论坛

汉语语体的计量特征在文本聚类中的应用

黄 伟[1，2]，刘海涛[2]

1.北京语言大学 汉语水平考试中心，北京 100083
2.中国传媒大学 应用语言学研究所，北京 100024

收稿日期 2009-7-31 修回日期 2009-8-31 网络版发布日期 2009-10-10 接受日期

摘要 提出了将语言计量研究成果应用于文本聚类研究的方法。通过两个50万词的语料样本发现了在现代汉语口语体和书面语体中具有显著分布差异的16个语言结构特征；以其中7个作为文本表示特征准确地将实验文本聚类为口语体（相似度89. 84%）和书面语体（相似度86. 93%）两类。以语言结构的计量特征表示文本的方法加强了聚类/分类研究的可解释性，具有较高的理论和应用价值。以语料库和统计方法进行语体特征计量研究是汉语语体描写研究的重要方法，阐述了其理论基础。

关键词 文本聚类 语体特征 语言结构 汉语口语 汉语书面语
分类号 TP391.1

Application of quantitative characteristics of Chinese genres in text clustering

HUANG Wei[1，2]，LIU Hai-tao[2]

1.Chinese Proficiency Test Center（HSK），Beijing Language and Culture University，Beijing 100083，China
2.Institute of Applied Linguistics，Communication University of China，Beijing 100024，China

Abstract
The method of applying the findings in quantitative study on linguistics to research on text clustering is presented.16 linguistic structures，which distribute distinctively between oral and written Chinese，are investigated based on two sample corpora with size of half million words for each.Test texts represented by using 7 of those linguistic structures are correctly clustered into spoken（similarity=89.84%） and written（similarity=86.93%） classes in a text clustering experiment.The method of representing texts with quantitative characteristics of linguistic structures enhances the interpretability of the results，and is feasible and theoretically and practicably significative in text clustering and text classification.Corpus and statistics are methodologically significant in describing study on Chinese genres，the theoretical foundations of which are also included.

Key words text clustering genre characteristics linguistic structure spoken Chinese written Chinese

DOI: 10.3778/j.issn.1002-8331.2009.29.007

通讯作者 黄 伟 huangwei@blcu.edu.cn