计算机工程与应用 2008 44 (5): 187-189 ISSN: 1002-8331 CN: 11-2127/TP

数据库与信息处理

# 文本分类中词语权重计算方法的改进与应用

熊忠阳, 黎 刚, 陈小莉, 陈 伟

重庆大学 计算机学院，重庆 400030

收稿日期 2007-5-28 修回日期 2007-7-25 网络版发布日期 2008-1-31 接受日期

摘要 文本的形式化表示一直是信息检索领域关注的基础性问题。向量空间模型（Vector Space Model）中的tf.idf文本表示是该领域里得到广泛应用，并且取得较好效果的一种文本表示方法。词语在文本集合中的分布比例量上的差异是决定词语表达文本内容的重要因素之一。但是其IDF的计算，并没有考虑到特征项在类间的分布情况，也没有考虑到在类内分布相对均匀的特征项的权重应该比分布不均匀的要高，应该赋予其较高的权重。用改进的TFIDF选择特征词条、用KNN分类算法和遗传算法训练分类器来验证其有效性，实验表明改进的策略是可行的。

关键词 文本表示 向量空间模型 特征选择 TFIDF

分类号

# Improvement and application to weighting terms based on text classification

XIONG Zhong-yang,LI Gang,CHEN Xiao-li,CHEN Wei

College of Computer，Chongqing University，Chongqing 400030，China

**Abstract**

Text representation has been the fundamental problem in Information Retrieval.tf.idf（term frequency，inverse document frequency） as one of term weighting schemes in Vector Space Model is a good text representation，Which is popular and make good results in the field of Information Retrieval.The difference of the proportion of distribution of terms in text collection is one of the most important factors of expressing the content of text.But the calculation of IDF，don’t consider the information of distribution about terms among classes，and don’t consider the more term weighting for the terms of the relative distributed balance inner classes.The improved TFIDF are used to select feature，KNN algorithm and genetic algorithm are used to train the classifier.and proves that the improved TFIDF method is feasible.

**Key words** text representation Vector Space Model feature selection TFIDF

DOI:

通讯作者 熊忠阳