计算机工程与应用 2007 43 (19): 192-194 ISSN: 1002-8331 CN: 11-2127/TP

数据库与信息处理

Web文档中词语权重计算方法的改进

初建崇[1]，刘培玉[2]，王卫玲[2]
1.海军航空工程学院 训练部，山东 烟台 264001
2.山东师范大学 信息科学与工程学院，济南 250014

收稿日期 修回日期 网络版发布日期 2007-6-20 接受日期

摘要 以向量空间模型作为Web文本的表示方法，对传统的TF*IDF公式进行了改进。首先，结合Web文本中HTML标签的修饰功能，体现了特征词在Web文本结构中的位置信息;其次，以广义信息论为理论基础，引入了基于二次熵的互信息作为权重计算公式的一项，体现了单词的类区分能力。实验验证了该方法的可行性和有效性。

关键词 向量空间模型 Web文本分类 权重调整 互信息
分类号

Improved approach to weighting terms in Web Text

CHU Jian-chong[1]，LIU Pei-yu[2]，WANG Wei-ling[2]

1.Naval Aeronautical Engineering Institute，Yantai，Shandong 264001，China
2.College of Information Science and Engineering，Shandong Normal University，Ji’nan 250014，China

Abstract
This paper uses vector space model as the description of the Web text，analyses and improves the traditional formula TF*IDF.First，we explore the feature of the Web pages which are written in HTML and describe the situation information of the terms in Web text.Second，we use generalized information theory as the theory base to introduce the quadratic entropy mutual information into the formula.The experiment shows the feasibility and the validity of this method.

Key words vector space model Web text classification weight adjustment mutual information

DOI:

通讯作者 初建崇 E-mail：wangweiling0714@163.com