计算机应用 2008 28 (03): 764-767 ISSN: 1001-9081 CN: 51-1307/TP

数据挖掘

一种基于新词发现的Web文本表示方法

吴春颖[1];王士同[1];蔡崇超[1]

江南大学[1]

收稿日期 2007-9-27 修回日期 2007-11-27 网络版发布日期 2008-3-1 接受日期

摘要 Web文本表示是Web文本特征提取和分类的前提，最常用的文本表示是向量空间模型（VSM），其中向量一般是基于词的特征项。由于向量空间模型本身没有考虑文本上下文间的潜在概念结构（如词汇间的共现关系），而Web文本是一种半结构化文本，同时经常有新词出现，因此在VSM基础上提出了一种基于新词发现的Web文本表示方法：首先进行预处理将网页转化为文本；然后进行文本分词；接着通过二元互信息进行新词发现，同时把新词加入字典重新分词；最后用词和新词共同来表示Web文本。实验结果表明，该方法可以帮助识别未登录词并扩充现有字典，能够增强Web文本表示能力，改善Web文本的特征项质量，提高Web文本分类效果。

关键词 中文分词 二元语法 互信息 新词发现 Web文本表示

分类号

DOI:

对应的英文版文章：A7095318

通讯作者:
吴春颖 chuny.wu@gmail.com;chuny.wu@hotmail.com;cxmxiaojie@yahoo.com.cn
作者个人主页: 吴春颖 王士同 蔡崇超