计算机工程 2010 36 (7): 173-175,178 ISSN: 1000-3428 CN: 31-1289/TP 首页 当期目录 上一期

人工智能及识别技术

含有语义特征的网页新闻自动抽取

施 洋，张 奇，黄萱菁

(复旦大学计算机科学技术学院，上海 200433)

收稿日期 修回日期 网络版发布日期 接受日期

摘要 通过分析新闻网页的语义特征以及网页之间存在的通用性质，提出一种含有语义特征的网页新闻自动抽取方法，包括利用语义分类器识别新闻网页中的种子信息以及页面中的局部信息来完成抽取。在分类器中加入语义特征可以使F1值达到94.2%。在语义分类器与局部特征结合的情况下，F1值可以达到96.9%。实验结果证明，该方法能有效提高网页信息抽取算法的精度，降低机器学习所需要的标注成本。

关键词 网络信息抽取；语义特征；局部特征

分类号 TP393

DOI:

通讯作者:

作者个人主页:

施 洋;张 奇;黄萱菁