现代图书情报技术 2007 (4): 48-51 ISSN: 1003-3513 CN: 11-2856/G2

知识组织与知识管理

基于熵的新闻网页抽取方法的研究

朱红灿[1];龙朝阳[1;2]

湘潭大学管理学院[1]

收稿日期 2007-2-5 修回日期 2007-2-28 网络版发布日期 2007-4-30 接受日期

摘要 为了减少或根除新闻网站中大量非主题信息的干扰，提出一种新闻网页抽取方法，采用基于熵的计算和DOM树的知识，从新闻网页中抽取主题文档和相关链接。

Abstract In this paper,an approach for news article extraction from Web page is proposed and this approach applies information theory to DOM tree. Experiment on several news Web sites shows that it is practical.

关键词 熵 信息抽取 信息块 DOM

Key words Entropy;Information extraction;Informative block;DOM

分类号 TP181

DOI:

通讯作者:
朱红灿 zhuhongcan@xtu.edu.cn
作者个人主页: 朱红灿 龙朝阳