计算机工程与应用 2007 43 (30): 148-151 ISSN: 1002-8331 CN: 11-2127/TP

网络、通信与安全

基于FFT的网页正文提取算法研究与实现

李 蕾[1,2], 王劲林[1], 白 鹤[1,2], 胡晶晶[1,2]
1.中国科学院 声学研究所 DSP中心,北京 100080
2.中国科学院 研究生院,北京 100039

收稿日期 修回日期 网络版发布日期 2007-10-11 接受日期

摘要 主要研究“正文式”网页的有效信息提取算法。该种底层网页真正含有Web页面所表达的主题信息,通常包含一大段的正文信息,正文信息的前后是一些格式信息(例如导航信息、交互信息、JavaScript脚本等)。分析了此种网页的页面结构特征,将问题转化为——给定一个底层网页的HTML源文件,求解最佳的正文区间;从而提出了一种基于快速傅立叶变换的网页正文内容提取算法。采用窗口分段的方法,利用统计学原理和FFT,得出每个可能区间的权值,从而求解出最佳正文区间。实验结果表明,此种方法能比较准确的对“正文式”网页的有效信息进行提取。

关键词 中文信息处理 Web页面 信息提取 页面结构 FFT 区域分割
分类号

Research and implementation of FFT-based extraction algorithm of webpage content main body

LI Lei[1,2],WANG Jin-lin[1],BAI He[1,2],HU Jing-jing[1,2]

1.DSP and Network Research Center,Institute of Acoustics,Chinese Academy of Sciences,Beijing 100080,China
2.Graduate University of Chinese Academy of Sciences,Beijing 100039,China

Abstract
This paper studies the extraction algorithm of the effective information of “Content-Dominated” Web pages.This kind of Web pages contains the major content information of the Web sites.It includes a long paragraph of content main body,and format information in the beginning and the ending（e.g.navigation information,interaction information,JavaScript and so on）.This paper analyzes the structural characteristics of this kind of Web page,and transformed the problem as：given an HTML source file of a “Content-Dominated” Webpage,to find the best range of the content main body.Presents an FFT-based extraction algorithm of webpage content main body.By applying window-segmentation,statistics theory and FFT,this method calculates the weight of every possible range;and thereby selects the best one as solution.The experimental result proves that this algorithm can efficiently extract the effective information of “Content-Dominated” Web pages.

Key words Chinese information processing Web page information extraction Web page structure Fast Fourier Transformation（FTT） page segmentation

DOI:

通讯作者 李 蕾 E-mail：lil@dsp.ac.cn