基于三叉树的自动化网页正文信息提取

摘要：随着互联网技术的发展与应用，互联网已经成为获取有价值信息的主要平台,通过从互联网上获取文本信息，对文本信息进行大数据分析，可以预测某些活动，防止某些可疑事件的发生，但是各个网站之间的网页结构存在很大的差别，不能采用统一标准来获取主题文本信息。本文基于三叉树提取模型来提取网页的主题文本信息，通过分析对比同一个网站之下的相同结构的网页模板，使用程序自动生成网页信息的正则表达式，通过对国内各大新闻网站（新浪新闻，网易新闻，新华网等）进行文本信息提取实验，可以正确的提取出网页文本的主题信息，实现自动化的网页文本信息抽取。

关键词：三叉树、主题文本、自动化、信息抽取。

引言：文本抽取是指从半结构化或者非结构化的网页中提取主题相关的文本信息。网页文本抽取是网络信息采集系统中最终主要的部分，通过信息抽取来获取主题相关的文本信息，然后对这些数据进行分析处理。现有的网页信息抽取主要依赖于正则表达式的制定，正则表达式在很大程度上依赖于网页的结构，特别是主要标签的位置，当从多个网站获取主题文本信息时，需要对每个网站制定特定的正则表达式。虽然将半结构化文本结构使用Jtitydy可以生成结构化的文本信息，但是还是会依赖于网页中的标签的位置以及属性，特别是div,table标签等，对于属性的依赖主要依赖于标签的id,class属性。基于基于行块分布函数的通用网页正文抽取 虽然从一定程度上能够解决正则表达式不统一的结果，但是当出现网页正文稀疏时，该模型会需要会出现文本抽取为空的现象。自由式文本信息抽取，从自由文本中抽取出有限的主要信息，处理自由文本IE系统通常使用自言语言处理的技巧，其抽取规则主要建立在词或词类间语法关系上，需要经过的处理步骤包括:句法分析、语义标注、专有对象识别和抽取规则，规则需要人工制定，该方法也从一定程度上依赖于文本的标签以及标签位置。

本文针对上面上面提出的问题，提出了一种基于网页结构聚类的文本信息抽取算法。主要的研究成果包括：（1）通过网页结构聚类算法可以正确将结构相同的网页正确聚类。（2）通过三叉树信息抽取模型，可以将网页正文信息进行提取，通过对国内各大新闻网站实验，可以正确的提取新闻的正文信息，实验的准确率在 以上。

本文的结构如下:第二节，介绍网页文本信息抽取的相关工作，主要涉及到网页信息抽取的发展历程；第三节，介绍基于网页结构的聚类模型以及基于三叉树算法信息抽取模型；

第四节，主要展示实验的环境以及实现的结果；第五节，主要描述本模型存在的缺点以及下一步的展望。