

学校编码: 10384

学号: 23020091152768

分类号 _____ 密级 _____

UDC _____

厦 门 大 学

硕 士 学 位 论 文

基于视觉分块及多特征的 web 信息抽取

Web Information-Extraction Based on Vision Block and Multi-Features

郑 艳 红

指导教师姓名: 张东 副教授

专业名称: 计算机应用技术

论文提交日期:

论文答辩时间:

学位授予日期:

答辩委员会主席: _____

评 阅 人: _____

2012 年 6 月

厦门大学博硕士学位论文摘要库

厦门大学学位论文原创性声明

本人呈交的学位论文是本人在导师指导下,独立完成的研究成果。本人在论文写作中参考其他个人或集体已经发表的研究成果,均在文中以适当方式明确标明,并符合法律规范和《厦门大学研究生学术活动规范(试行)》。

另外,该学位论文为()课题(组)的研究成果,获得()课题(组)经费或实验室的资助,在()实验室完成。(请在以上括号内填写课题或课题组负责人或实验室名称,未有此项声明内容的,可以不作特别声明。)

声明人(签名):

年 月 日

厦门大学博硕士学位论文摘要库

厦门大学学位论文著作权使用声明

本人同意厦门大学根据《中华人民共和国学位条例暂行实施办法》等规定保留和使用此学位论文，并向主管部门或其指定机构送交学位论文（包括纸质版和电子版），允许学位论文进入厦门大学图书馆及其数据库被查阅、借阅。本人同意厦门大学将学位论文加入全国博士、硕士学位论文共建单位数据库进行检索，将学位论文的标题和摘要汇编出版，采用影印、缩印或者其它方式合理复制学位论文。

本学位论文属于：

1.经厦门大学保密委员会审查核定的保密学位论文，于 年 月 日解密，解密后适用上述授权。

2.不保密，适用上述授权。

（请在以上相应括号内打“√”或填上相应内容。保密学位论文应是已经厦门大学保密委员会审定过的学位论文，未经厦门大学保密委员会审定的学位论文均为公开学位论文。此声明栏不填写的，默认为公开学位论文，均适用上述授权。）

声明人（签名）：

年 月

厦门大学博硕士学位论文摘要库

摘 要

随着信息社会的快速发展,web数据已经发展成为一种巨大的信息资源。Web信息抽取作为一种从web数据中抽取主题信息的研究内容,是数据分类、自然语言处理等研究领域的基础。因此,如何准确快速的从海量的web数据中抽取关注的信息变得越来越重要。本文对web信息抽取的方法进行了研究,并针对研究过程中遇到的问题,提出相应的解决方法。本文的主要研究内容如下:

(1) 对已存在的各种 web 信息抽取算法做出了详细的研究比较。

(2) 本文的主要目的是对具有主题信息的主题型网页进行正文抽取,而对于链接型网页不予处理。因此要先判断输入网址的网页类型。本文对两种网页进行了详细的比较,提炼出五个明显的特征,并提出一种基于多特征的网页类型划分方法。该方法利用遗传算法对数据集进行训练求得各个特征的权重,再通过计算网页各个特征的加权和来判断类型。

(3) 网页类型划分完成之后,对主题型网页进行正文抽取工作。本文对微软亚洲研究院所提出的基于视觉的分块算法 VIPS 算法进行了改进,提出了 nVIPS 算法,并在此基础上提出新的算法对正文标题、正文发表时间、正文内容进行抽取。

(4) 对网易、腾讯、人民网等八大网站共 800 篇文章进行抽取实验。并在相同数据集和运行环境下实现了基于多特征的正文抽取算法以及 VIPS 算法。通过实验结果对比表明本文提出的方法是快速有效的。

关键字: web 信息抽取; 网页类型判断; VIPS

厦门大学博硕士学位论文摘要库

Abstract

With the rapid development of information society, the web data has developed into a huge information resource. Web information-extraction is one Research based on extracting theme information from web data set, it is the basis of data classification, natural language processing and some other research areas. Therefore, how extract concerned information from the vast amounts of web data fast and accurate has become increasingly important. We study the methods of web information extraction and put forward the corresponding solution for the problem encountered in the course of the study. The main content is as follows:

- (1) Make a detailed study and comparison about the existed web information extraction.
- (2) The main purpose of this paper is extracting content from the Theme-oriented pages, and ignoring the Hub-oriented pages. Therefore, when we enter a URL we have to judge the page's type first. A detailed comparison of the two pages and five obvious features are showed in this paper. At the same time, this paper proposed a page divided method based on multi-features. The method uses GA to obtain the weights of the five features by training the data, and then calculates the values to judge the page's type.
- (3) After the completion of Page divide, extracting information of the theme-oriented pages. This paper proposed an algorithm nVIPS which improved the VIPS proposed by Microsoft Research Asia, and then proposed a new algorithm to extract article title, article time and article content from pages.
- (4) Extraction experiments are made on 800 articles from Netease, Tencent, People and other 8 sites. And then achieve the multi-features extraction algorithm and VIPS in the same data set and the same run environment. Contrast to the experimental results show that the method is fast and efficient.

Key words: Web information-extraction; Sort of web page; VIPS

厦门大学博硕士学位论文摘要库

目 录

目 录	1
Contents	1
第一章 绪论	1
1.1 研究背景及意义	1
1.2 国内外研究现状	2
1.3 本文的组织结构	4
第二章 关键技术与算法分析	7
2.1 引言	7
2.2 HTML	7
2.2.1 HTML 简介	7
2.2.2 HTML 网页标签分析	8
2.3 DOM.....	9
2.3.1 DOM 技术研究	9
2.3.2 HTML 文档的树模型	10
2.3.3 基于 DOM 的 web 信息抽取算法.....	11
2.4 VIPS	13
2.4.1 VIPS 算法简介	13
2.4.2 基于 VIPS 的 web 信息抽取算法	17
2.4.3 VIPS 算法和 DOM 技术相结合的 web 信息抽取算法.....	18
2.5 同义词词林及其扩展版	18
2.5.1 同义词词林的编码方法.....	19
2.5.2 基于同义词词林的词语相似度算法.....	20
2.5.3 其他词语相似度计算方法.....	21
2.6 本章小结	21
第三章 基于多特征的网页类型划分算法.....	23

3.1 引言	23
3.2 现有的网页类型划分方法	23
3.2.1 整体判断方法.....	23
3.2.2 基于局部的方法.....	24
3.2.3 基于分块的方法.....	24
3.3 基于多特征的网页类型划分方法	25
3.3.1 网页类型特征对比.....	25
3.3.2 特征表示及问题描述.....	26
3.3.3 特征值统计.....	26
3.3.4 遗传算法求特征权值 W	29
3.3.5 网页类型划分算法.....	31
3.4 实验结果及分析	32
3.4.1 实验数据.....	32
3.4.2 实验结果评价指标.....	32
3.4.3 实验结果分析及比较.....	33
3.5 本章小结	33
第四章 基于改进 VIPS 的正文抽取算法	35
4.1 引言	35
4.2 基于 nVIPS 的网页正文抽取方法.....	35
4.3 改进的 VIPS 算法—nVIPS.....	35
4.3.1 算法改进.....	36
4.3.2 算法比较.....	37
4.4 正文信息抽取算法	39
4.4.1 分块标注算法.....	39
4.4.2 正文标题抽取算法.....	41
4.4.3 正文发表时间抽取算法.....	42
4.4.4 正文内容抽取算法.....	43
4.5 实验结果分析	44
4.5.1 实验数据集.....	44

4.5.2 结果评价指标.....	45
4.5.3 实验结果.....	45
4.5.4 结果分析.....	47
4.6 本章小结	48
第五章 结论	49
5.1 总结	49
5.2 后续工作	49
参考文献	51
攻读硕士学位期间发表的论文	55
致 谢	57

厦门大学博硕士学位论文摘要

厦门大学博硕士学位论文摘要库

Contents

Chapter1 Introduction	1
1.1 Backgroud and Signification.....	1
1.2 Research Status	2
1.3 Organizational Structure.....	4
Chapter2 Key technologies and algorithm analysis	7
2.1 Introduction.....	7
2.2 HTML	7
2.2.1 HTML brief introduction	7
2.2.2 HTML lable analysis.....	8
2.3 DOM.....	9
2.3.1 DOM technical study	9
2.3.2 Tree model of HTML file.....	10
2.3.3 Web infromation extraction Based on DOM Tree	11
2.4 VIPS	13
2.4.1 VIPS Algorithm brief introduction	13
2.4.2 Web infromation extraction Based on VIPS	17
2.4.3 Web infromation extraction combine DOM with VIPS.....	18
2.5 TongYiCi CiLin	18
2.5.1 TongYiCi CiLin’s coding way	19
2.5.2 Word similarity algorithm based on TongYiCi CiLin.....	20
2.5.3 Other Word similarity algorithm	21
2.6 Chapter Summary	21
Chapter3 Page sorting algorithm based on multi-features.....	23
3.1 Introduction.....	23
3.2 Existing web page sorting methods	23
3.2.1 Overall judgment method	23

3.2.2 Method based on local	24
3.2.3 Method based on block	24
3.3 Page type classification method based on multi-features	25
3.3.1 Characteristic contrast of the page type	25
3.3.2 Feature representation and description of the problem	26
3.3.3 Statistics of every eigenvalue	26
3.3.4 GA solving each eigenvalue's weight W	29
3.3.5 page sorting algorithm	31
3.4 Experimental results and analysis	32
3.4.1 Experiment Data	32
3.4.2 Evaluation	32
3.4.3 Analysis of experimental results	33
3.5 Chapter Summary	33
Chapter4 Content extraction algorithm Based on Improved	
VIPS algorithm--nVIPS	35
4.1 Introduction	35
4.2 Problem description	35
4.3 Improved VIPS algorithm--nVIPS	35
4.3.1 Algorithm improve	36
4.3.2 Algorithm compare	37
4.4 content extraction algorithm	39
4.4.1 Block-Mark	39
4.4.2 Page title extraction method	41
4.4.3 Page time extraction method	42
4.4.4 Page content extraction method	43
4.5 Experimental results and analysis	44
4.5.1 Experiment Data	44
4.5.2 Evaluation indicators	45

Degree papers are in the "[Xiamen University Electronic Theses and Dissertations Database](#)". Full texts are available in the following ways:

1. If your library is a CALIS member libraries, please log on <http://etd.calis.edu.cn/> and submit requests online, or consult the interlibrary loan department in your library.
2. For users of non-CALIS member libraries, please mail to etd@xmu.edu.cn for delivery details.

厦门大学博硕士论文摘要库