****

**本 科 毕 业 设 计（论文）**

题 目：基于语义的大规模中文网页模块识别与信息

提取系统设计与实现

——搜索引擎预处理模块

|  |  |
| --- | --- |
| 学生姓名： | 张 三 |
| 学 号： | 1707010101 |
| 专业班级： | 计算机科学与技术17-1班 |
| 指导教师： | 李 四 |

2023年6月11日

学位论文原创性声明

本人所提交的学位论文 ，是在导师的指导下，独立进行研究工作所取得的原创性成果。除文中已经注明引用的内容外，本论文不包含任何其他个人或集体已经发表或撰写过的研究成果。对本文的研究做出重要贡献的个人和集体，均已在文中标明。

本声明的法律后果由本人承担。

论文作者（签名）： 指导教师确认（签名）：

年 月 日 年 月 日

学位论文版权使用授权书

本学位论文作者完全了解中国石油大学（华东）有权保留并向国家有关部门或机构送交学位论文的复印件和磁盘，允许论文被查阅和借阅。本人授权中国石油大学（华东）可以将学位论文的全部或部分内容编入有关数据库进行检索，可以采用影印、缩印或其它复制手段保存、汇编学位论文。

保密的学位论文在\_\_\_\_\_\_\_年解密后适用本授权书。

论文作者（签名）： 指导教师（签名）：

年 月 日 年 月 日

摘 要

本文提出了一套基于语义的网页分块和主题内容信息提取算法，在搜索引擎预处理模块中将其实现，并且在国际会议中，以这套算法为框架，组织了主题型网页识别和网页主题内容信息块提取两个中文Web信息检索评测项目。在这套算法的基础上，基于文件系统与Map-Reduce计算平台，实现了分布式的网页块级别QuarkRank算法，改进了PageRank算法的效果。实际检验表明，该套算法具有很好的适应性与可扩展性，并达到了很高的精度和召回率。

**关键词**：网页分块；信息提取；评测；Map-Reduce；PageRank

**Abstract**

This paper presents a semantic web-page blocking and information extraction of thematic content algorithm, which is achieved in the pretreatment module of search engine, and in SEWM meeting, using this algorithm, we organized two Chinese Web Information Retrieval Evaluation Projects, which are theme-based Web page identification and block extraction of the information theme content. In this method, based on file system and the Map-Reduce computing platform, this paper reports the distributed block-level QuarkRank algorithm, which improves the result of PageRank algorithm. The actual test showed that this algorithm is good at adaptability and scalability, and reaches a very high precision and recall.

**Keywords**: Web-Page Blocking, Information Extraction, Evaluation, Map-Reduce, PageRank

目 录

[第1章 序言 1](#_Toc69482696)

[第2章 相关研究工作 2](#_Toc69482697)

[2.1 基于语义的网页信息提取算法 2](#_Toc69482698)

[2.2 Block Level PageRank算法 2](#_Toc69482699)

[2.2.1 Block Level Web Graph 2](#_Toc69482700)

[2.2.2 Page Graph 3](#_Toc69482701)

[2.3 基于视觉的网页分块算法 3](#_Toc69482704)

[第3章 搜索引擎Quark模块 4](#_Toc69482705)

[3.1 模块结构 4](#_Toc69482706)

[3.2 网页分块算法 5](#_Toc69482707)

[3.3 网页主题内容提取 6](#_Toc69482708)

[3.4 算法效果演示 6](#_Toc69482709)

[第4章 SEWM中文Web信息检索评测 7](#_Toc69482710)

[4.1 评测任务介绍 7](#_Toc69482711)

[4.2 评测格式 7](#_Toc69482712)

[4.3 评测结果 7](#_Toc69482713)

[4.3.1 主题型网页发现任务评测结果 7](#_Toc69482714)

[4.4 评测综述 7](#_Toc69482716)

[第5章 网页分块的分布式应用 8](#_Toc69482717)

[第6章 总结与展望 9](#_Toc69482718)

[6.1 总结 9](#_Toc69482719)

[6.2 展望 9](#_Toc69482720)

[致 谢 10](#_Toc69482721)

[参考文献 11](#_Toc69482722)

[附 录 12](#_Toc69482723)

[附录A名词术语及缩略词 12](#_Toc69482724)

# 第1章 序言

信息时代，非Web无以制胜。互联网的高速发展，改变了我们的生活方式，打破了我们的时空界限，重塑着我们的社会形态。经济、政治、学习、工作、生活、娱乐等等各个层面都在Web网络中激荡起伏，深刻地影响着人类的未来。而Web网络的灵魂，就是流动在其中的无穷无尽的信息。Web2.0的意义就在于网络内容的提供方从商人和专业人员转变为网络上的每一个普通用户，从而几何级数地增长了Web的信息量。然而信息量的增大，随着而来的就是存储成本的增大和信息提取难度的增大，如何有效的获取和整合Web信息成为大家面对的共同课题。

。。。

在这篇论文中，第2章介绍了本文的相关研究工作，包括常见的网页分块和信息提取算法、基于视觉的网页分块算法，以及网页分块的一个应用Block LevelPageRank算法；第3章介绍了实现的网页分块和主题信息提取算法——Quark算法；第4章介绍了Quark算法在SEWM中文Web信息检索评测项目中的实际检验；第5章介绍了在Quark算法基础上实现的一个分布式QuarkRank程序。第6章是对本文的总结和工作展望。

【本段为提示，撰写时删除。

正文页眉写每一章对应的章名；摘要、目录、致谢、参考文献、附录等页眉为“中国石油大学（华东）本科毕业设计(论文)”，采用“楷体\_GB2312”五号字，居中。封面部分无页眉。

从正文到附录，页脚用阿拉伯数字连续编排页码；页码位于页脚中间，采用“Times New Roman 五号字体”。封面、摘要、目录部分不用编写页码。

】

# 第2章 相关研究工作

## 2.1 基于语义的网页信息提取算法

由于对Web页面有效分块之后可以极大地方便内容提取、数据挖掘、Web结构分析等各项Web信息检索领域的相关工作，所以早有很多研究人员前赴后继，就此展开了很多工作。其中，基于语义信息对网页分块是最简便，也最基础的一种方法。所谓语义信息，通常包括网页中包含的HTML标签信息，HTML DOM树的结构信息，文字内容信息，超链接信息，以及其他通过统计或学习而得到的全局信息等等，也可以理解成为除了网页中的视觉信息之外的所有可以得到的信息。

。。。

Site-Level的算法顾名思义，就是分析一个网站或者网页集内部的所有网页，从中提取反复出现的模式，而一般来说，在多个网页里重复出现的模式（可理解为Dom-Tree 子树）就是导航栏、广告等噪音信息。单个网页中减去这些信息，剩下的就是主题信息内容。关于Site-Level的研究一直在继续，WWW2008上就有一篇名为Web page sectioning using regex--based template[1]的论文使用正则表达式来提取重复模式，从而更适应网页间的细微变化，增加了Site-Level的召回率。

。。。

## 2.2 Block Level PageRank算法

在VIPS算法的分块基础上，微软2004年的论文Block-level Link Analysis[5]中提出了Block Level PageRank(BLPR)算法。之前的大多数链接分析算法都是以一个Web页面为Web图中的一个节点，而BLPR算法以网页中的语义块为原子节点，从链接结构和页面结构中提取出Page-to-Block，Block-to-Page关系矩阵，构建出新的Web语义图，并以此计算PageRank。实验表明，BLPR改进了PageRank的质量。

### 2.2.1 Block Level Web Graph

首先定义两个集合P和B。P为所有网页的集合，P = {p1, p2, …, pk}，k为网页总数。。。

。。。

【正式撰写时删除

三级以下标题用“1、”“1）”、“（1）”、“①”等，内容为小四号宋体，1.5倍行距，首行缩进2字符。】

。。。

### 2.2.2 Page Graph

。。。

## 2.3 基于视觉的网页分块算法

。。。

# 第3章 搜索引擎Quark模块

## 3.1 模块结构

基于可扩展性和独立性特点，初步实现了Quark模块，类结构如图3-1所示。

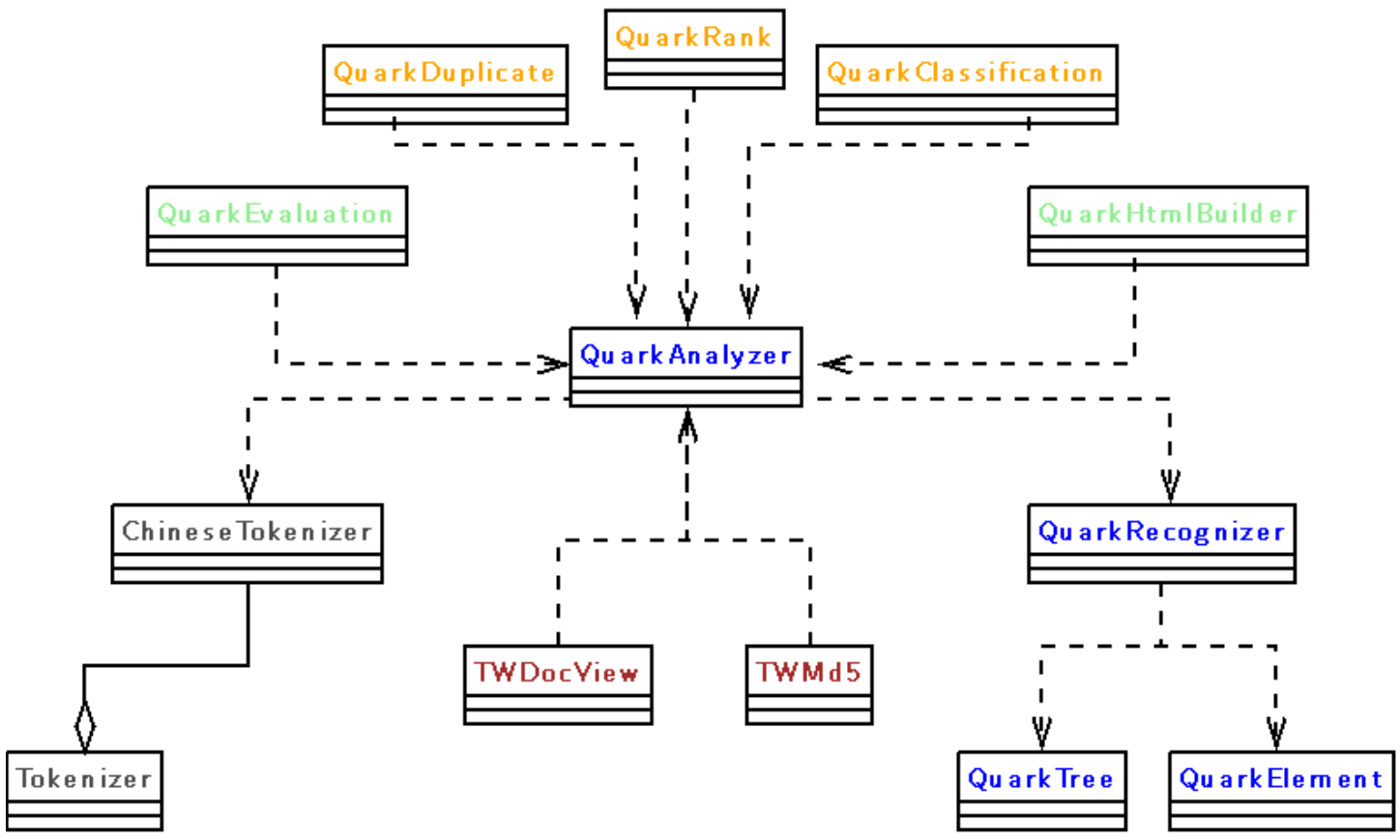


图3-1 Quark模块类结构

1、图中右下及中间蓝色的部分为Quark模块的核心功能类，包括QuarkTree、QuarkElement、QuarkRecognizer、QuarkAnalyzer等四个类。

1）QuarkTree 类的作用有两个，一个是以原始网页为输入，建立Html的Dom Tree；另一个是存储分好的网页块（在系统中，每一个网页块就叫做一个Quark）并记录Quark与Quark之间的组织架构。

QuarkElement类指代一个Quark，即每个Quark自身就是一个QuarkElement类的对象。

QuarkRecognizer类肩负网页分块的重任，从网页中识别出所有语义块。它依赖于前面的两个类。

QuarkAnalyzer类依赖于QuarkRecognizer类，它在分好的块的基础上，判断各个块的类型，提取正文信息。这个类是整个Quark模块最核心的类，目前功能只是初步实现，还有很大的改进空间，将来也可以根据功能将其分割成多个类。

。。。

## 3.2 网页分块算法

。。。QuarkRecognizer算法的核心伪代码如算法3-1所示。

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **算法3-1: QuarkRecognizer算法** | | | | |
| */\* periodically verify node n’s broker list and topic lists \*/* | | | | |
| **Define:** | | | | |
|  | *bl*: the current broker list of the node *n* | | | |
|  | *bl[].QuorumKey*: the *QuorumKey* of current node in *bl* | | | |
|  | *bl[].Ip*: the *Ip* of the correct node in *bl* | | | |
|  | *bl[].InRow*: the correct node is in the same row or same column | | | |
|  | *bl[].NoRes*: the number of correct node doesn’t receive the response message | | | |
|  | *bl[i].tl*: the topic list of the correct node in *bl* | | | |
|  | *bl[i].tl.status*: the topic list of the correct node in *bl* | | | |
|  | *HbRes*: the return of the heart-beat message | | | |
|  | *SubsNode*: the return of the *RandomNode()* | | | |
| **Algorithm:** | | | | |
| 1 | *r* 🡨 *Sizeof(bl);* | | | |
| 2 | **For** *i*=1; *i*≤*r*; **do** | | | |
| 3 |  | *HbRes* 🡨 *SendHeartbeatMsg(bl[i].ip)*; | | |
| 4 |  | **If** *HbRes* = *Null* **then** | | |
| 5 |  |  | *bl[i].NoRes*++; | |
| 6 |  |  | **If** *bl[i].NoRes* ≥ 3 **then** | |
| 7 |  |  |  | *SubsNode* = *RandomNode(bl[i].ip)*; |
| 8 |  |  |  | *bl[i].QuorumKey* 🡨 *SubsNode.QuorumKey*; |
| 9 |  |  |  | *bl[i].Ip* 🡨 *SubsNode.Ip*; |
| 10 |  |  |  | *bl[i].InRow* = *SubsNode.InRow*; |
| 11 |  |  |  | *bl[i].NoRes* = 0; |
| 12 |  |  | **End** | |
| 13 |  |  | *bl[i].tl.status* = *disabled*; | |
| 14 |  | **Else** | | |
| 15 |  |  | *bl[i].tl* = *HbRes.tl*; | |
| 16 |  |  | *bl[i].tl.status* = *enable*; | |
| 17 |  | **End** | | |
| 18 | **End** | | | |

## 3.3 网页主题内容提取

为了简化以下内容的描述，可以使用一个三元组{*n, m, k*}表示标准布鲁姆过滤器。一个元素不属于集合的假阳性概率，也被称为假阳性率，可表示为公式（3-1）。

。。。

## 3.4 算法效果演示

。。。

# 第4章 SEWM中文Web信息检索评测

## 4.1 评测任务介绍

## 4.2 评测格式

## 4.3 评测结果

### 4.3.1 主题型网页发现任务评测结果

评测结果如表4-1所示。

表4-1 主题型网页发现任务评测结果

|  |  |  |  |
| --- | --- | --- | --- |
| TEAM | Macro-Precision | Macro-Recall | Macro-F1 |
| DLUT | 0.888888888889 | 0.869565217391 | 0.879120879121 |
| SCU | 0.846153846154 | 0.876811594203 | 0.861209964413 |
| SCUT | 0.883211678832 | 0.876811594203 | 0.883211678832 |
| SDU | 0.794871794872 | 0.905797101449 | 0.838926174497 |
| RUC | 0.670103092784 | 0.942028985507 | 0.783132530102 |

## 4.4 评测综述

# 第5章 网页分块的分布式应用

。。。

# 第6章 总结与展望

## 6.1 总结

本文提出了一套基于语义的网页分块和主题内容信息提取算法，在天网搜索引擎预处理模块中将其实现，并通过了SEWM中文Web信息检索评测项目的检验。在该套算法基础上，还实现了基于Map-Reduce的分布式QuarkRank算法，该算法改进了PageRank算法的效果。

。。。

【建议增加以下三个方面的分析或评价，正式撰写时删除

* 从社会、健康、安全、法律或文化等方面，分析与评价本课题对这些因素的影响
* 从绿色节能计算、低功耗设计或软硬件版本兼容性等方面，评价系统对环境和可持续发展的影响
* 将项目管理的原理、技术和方法应用于系统分析、设计、实现或运维过程的效果和心得

】

## 6.2 展望

。。。而且还存在着许多不足之处。如：。。。

# 致 谢

在即将要离开石大的时候，心中充满了留恋与感慨。。。

。。。

# 参考文献

[1] 严蔚敏,吴伟民.数据结构（C语言版）[M].北京:清华大学出版社,1997.4.

[2] 沈晴霓,聂青,苏京霞.现代程序设计—C++与数据结构面向对象的方法与实现[M].北京:北京理工大学出版社,2002.8.

[3] Thomas Connolly Carolyn Begg. Database systems[M].北京:电子工业出版社,2004.7.

[4] Roger Bate,Sandy Shrum. CMM Integration framework[J].CMU/SEI Spotlight 1998.9.

[5] J P Kuilboer,N Ashrafi. Software process and produt improvement[J].An empirical assessment,2000.4.

[6] 张美金 著.基于ASP技术的远程教育系统体系结构的研究.http://172.50.0.88:86 /~cddbn/Y517807/pdf/index.htm,2003-05-01.

[7] 王伟国,刘永萍,王生年,等.B/S模式网上考试系统分析与设计[J].石河子大学学报（自然科学版）,2003,6(2):145-147.

[8] …

[9] …

[10] …

# 附 录

## 附录A名词术语及缩略词

论文的附录用大写字母A，B，C等连续编号，如附录A。附录中的图、表、公式算法、关键代码等用阿拉伯数字另行编序号，但在数码前冠以附录序码，如图A1；表B2；式(B3)等。