****

**学士学位本科生**

**开题报告与文献总结**

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 |

2021年6月1日

**开题报告**

# 1 选题目的及研究意义

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。

。。。

【本段为提示，撰写时删除。

正文页眉为“中国石油大学（华东）本科生毕业设计开题报告”，采用“楷体\_GB2312”五号字，居中。封面部分无页眉。

正文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”。封面部分不用编写页码。

】

# 2 国内外研究现状

## 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。

。。。

## 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

。。。

# 3 相关技术原理及技术路线

## 3.1 技术原理

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。

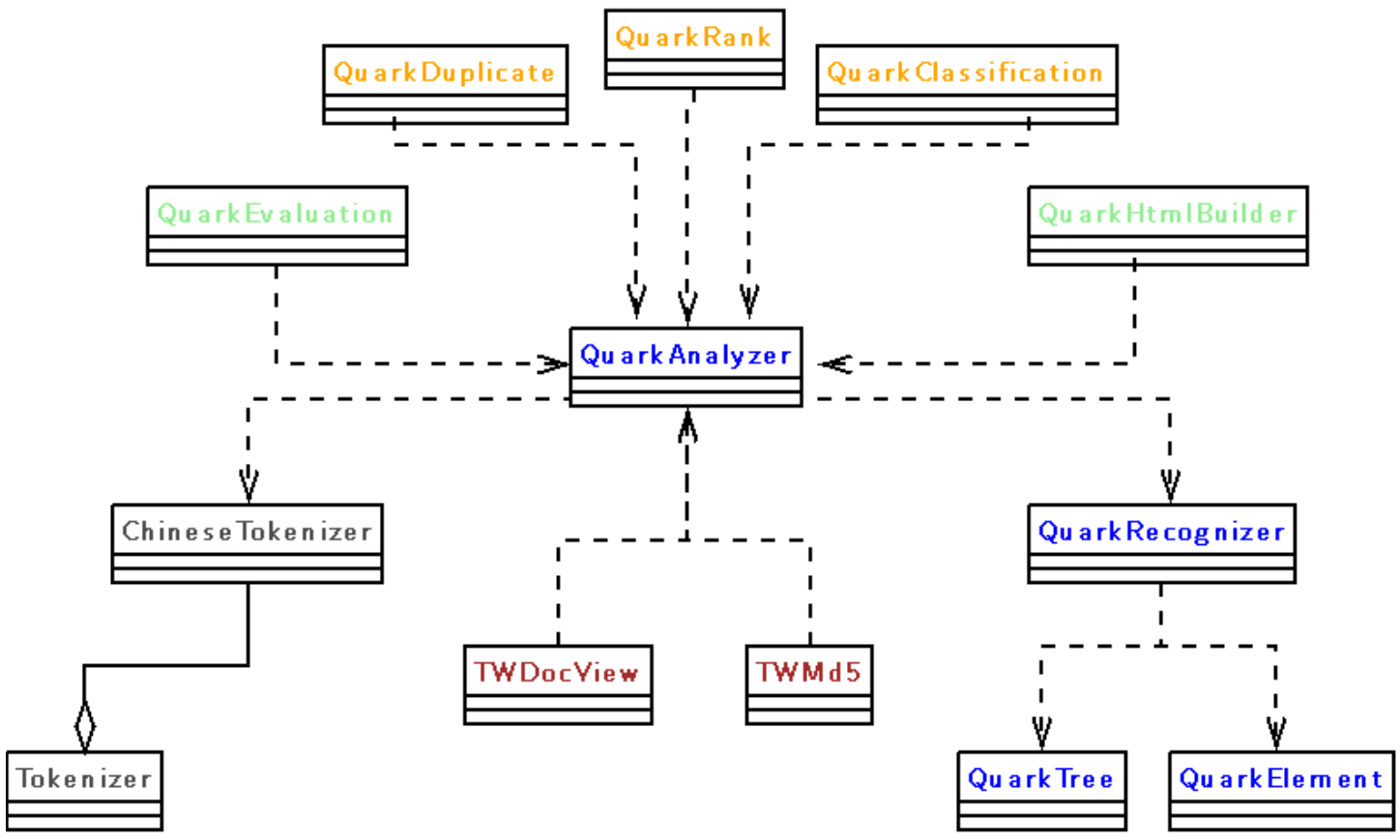


图3-1 Quark模块类结构

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。

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 技术路线

。。。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* | | | |
|  | *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** | | | |

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）。

# 4 研究难点及目前存在的问题

## 4.1 研究难点

## 4.2 目前存在的问题

## 4.3 拟采取的措施

### 4.3.1 xxxx

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 |

# 5 研究预期成果

。。。

# 6 进度安排

。。。

**文献总结**

【本段为提示，撰写时删除。

正文页眉为“中国石油大学（华东）本科生毕业设计文献总结”，采用“楷体\_GB2312”五号字，居中。

从文献总结到参考文献，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”。

】

# 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] …