四川大学计算机学院（软件学院）

硕士学位论文修改报告书

|  |  |  |  |
| --- | --- | --- | --- |
| 论文题目 | 基于链接和页面内容的主题爬虫算法的研究与应用 | | |
| 研究生姓名 | 2012223040119 | 导师姓名 |  |
| 论文修改对照表 | | | |
| 论文评阅意见 | | 修改后的内容（或解释说明） | |
| 图2-1搜索引擎架构图没有数据流向的顺序 | | 添加图2-1搜索引擎架构图的数据流向箭头 | |
| 主题爬虫最近的参考文献没有引用 | | 添加了最近的主题爬虫算法参考文献10、11、12、13、14，以及各个阶段典型的主题爬虫算法参考文献15、16、17、18 | |
| P32页伪代码17-22 23-29并列for循环应改写 | | 将本文代码中将子链接插入到待爬取链接队列和计算深度的两个并列for循环写到一个for循环中 | |
| P36页Page Rank算法的伪代码命名不对 | | 将代码名称改为基于Page Rank的主题爬虫算法的伪代码 | |
| 只考察1000个爬取页面，实验力度不够，建议多在一些数据集上做些实验，增加论文的力度 | | 增加了初始节点为10个、爬取页面总数为10000个的实验 | |
| 为了计算Page Rank爬虫要对新解析出来的URL进行基于链接的迭代计算，这使爬虫效率大大降低，应如何解决该问题？ | | 本文主要意在提高爬虫的查准率，爬虫效率可以后通过搭建GPU平台进行加速，软硬件结合的方式解决该问题 | |
| 研究生签名：  年 月 日 | | 导师签名：  年 月 日 | |

说明：

1）论文评阅意见请根据评阅书进行综合整理后分类表述，尽量做到条目化；

2）修改报告书在硕士论文正式答辩前填写完整，连同论文评阅意见表（各复印5份），提交答辩委员会，并作为论文工作的评定依据之一。