**7. 1毕业设计工作总结**

随着互联网络的迅速发展，Internet上网络信息的爆炸式增长，对于网络信息的搜索和需求显得至关重要。起初通用网络爬虫被运用到Internet的各个领域，因为它具备整合信息能力，能够实现用户的全面搜索。然而，尽管如此，通用爬虫的弊端也越来越明显，例如内容陈旧、查全率查准率偏低、信息冗余、分布不等。无法满足对于特定用户和人群的搜索需求，因此在这种背景下，主题网络爬虫为基础的主题搜索引擎诞生。其运行效率的高低直接关系到下一代搜索引擎的性能。

1. 设计并实现了微博的爬虫系统。通过模拟登录的方法解决身份认证，解决验证码问题，伪装成为移动端环境爬取微博移动端web等问题，爬虫采用广度优先搜索的思想。并结合网络爬虫、BeautifulSoup、Scrapy框架、正则表达式和多线程并发、Cookie池等技术实现多种用户信息和微博信息的采集。采用引入Cookie池输入多个微博用户账号代理的方法解决了微博服务器限制请求过频的问题，使用了内存非关系型Redis数据库作为多线程的队列，同时用Redis实现爬虫程序能中断的功能，并利用非关系型数据库MongoDB存储数据，实现了微博数据的持续采集与存储，为后续研究提供了数据支持。
2. 在数据库MongoDB中事先在爬虫的设计过程中对微博用户世界机型抽象建立数据库模型，并画出相应的E-R图与三张数据库表。它们分别是用户关系表，用户个人信息表，用户发表的微博表；然后根据微博的网页特性在爬虫系统各相应模块编写正则表达式。
3. 根据饰演的结果，对爬取道德400W条数据进行简单的统计分析，得出一些小结论达到实验的目的。

根据实验结果和相应的统计分析，显示了达到了预期的实验目标，成功地改

进了爬虫的工作方式和工作效率，取得了一定的创新成果。使得系统在采用该算

法时可以更好地过滤掉过程中与主题无关的网页，提高了效率。即便不完全遍历

整个互联网也可以较为准确的抓取主题页面。

**7.2不足**

在本文的研究上存在一些不足，以下分别列举出来：

1. 虽然本毕设采用了Scrapy-Redis架构这个实际上是分布式架构但是由于现实软硬件网络环境等的限制没有采用分布式，只是采取了单机上的多线程高并发式的爬取，在实际中是无法达到工业级别要求的只能算是模拟仿真的小demo。
2. 硬件上是采用我个人笔记本电脑，网络环境是采用的普通家庭的网络环境。
3. 由于软硬件网络环境等的限制只能爬取微博用户的400W条数据，数据量不够太少了。
4. 操作系统平台采取的Windows10专业版没有采用Linux网络操作系统，稳定性，鲁棒性等都不足。
5. 由于资金限制，不能部署到云端跑程序。

**7.3展望**

在本文的研究基础上，可以在以下五个方面进一步研究:

第一，本文的数据存储采用的是非关系型数据库MongoDB要把MongoDB分片在不同的数据库集群中,期间做好调度，同时为了更好地适应大数据分析，下一步考虑采用Scrapy-Redis架构搭建分布式爬虫系统，分布式文件系统更适用于大规模的数据处理,更能适应企业对公司业务的需求。实际工业水平上采用P2P分布式爬虫才能真正的达到企业对大型网站数据获取的大数据业务，比如百度搜索引擎就是在就在全国的50多个分布式集群。

第二，本文的研究由于硬件环境的限制是在自己笔记本电脑上进行的,只能做模拟仿真爬取一下,模拟形式下在实际工业生产过程中是不合用的,在工业级别中要求爬虫系统应该布置在集群上或者说至少是服务器哪怕是单台服务器上,相应的网速要更快,硬盘容量要更大达到TB、PB级别上，能是SSD（固态硬盘）更好。

第三，本文的研究是为了毕设的方便所以采用的操作系统平台是Windows10专业版，在实际工业中应该采取Linux操作系统，把程序放到Linux服务器上去跑，比如Redhat,Centos7等等，因为linux 稳定性、安全性、能全年365天跑程序都没有问题很适合爬虫。

第四，本文的研究能进一步改进，比如为了节省开支，去阿里云、腾讯云去购买VPS云服务器，数据库也采用云数据库等等。

第五，本文采用网络爬虫采集数据，由于微博网站的更新改版，带来了较高的维

护成本，今后可继续优化改进系统，降低爬虫系统的维护成本。