**3.1.1需求分析**

随着微博应用在众多网络用户心中的地位越来越重，越来越多的研究者重

视对微博数据的研究。要充分对庞大的微博数据进行分析研究，首先必须要构

建一个能够抓取微博信息的微博爬虫。目前存在的微博爬虫一方面是基于微博

平台API的，在数据获取数量上有限制;另一方面获取的微博数据也是杂乱无

章的，没有规律性可言。基于上面的两个当前微博爬虫存在的问题，本文提出

了基于主题的微博Web爬虫。那么根据这些描述，本文研究的爬虫需要实现以

下的功能:

(1)模拟微博平台的登陆过程,对抗微博的反爬技术，获得微博平台的合法URL地址，以便后面URL地址的扩大采集。

（2)本毕设不使用微博平台的API，利用网页分析技术分析微博网页页面，回避现有的API的种种限制，尽可能的抓取更多的微博数据。

(3)爬取的数据包括对用户发表的微博点赞、转载、评论数量与发表时间、定位、发表平台等等进行爬取；爬取微博用户ID、粉丝、省市、关注、首页链接等等基本信息；爬取用户间关系方便以后建立关系分析等。

（4）

(4)设计数据库Mongodb，存储微博数据以及微博用户信息。

根据上述功能需求的分析，可以将整个爬虫大致分为三个模块。这三个模

块分别为:模拟微博平台登陆模块、微博信息抓取模块、微博数据存储模块。

其中，第一个模块为起始模块，是实现后续模块的基础，而第二个模块为核心

模块，是实现抓取主题相关微博的主要模块，其中最重要的是微博主题相关性

分析子模块，而第三个模块便是爬虫的收尾模块，是对于前几个模块成果的体

现。每个模块实现的具体功能为:

（1）模拟微博平台登陆模块:根据微博平台登陆的流程，模拟其登陆过程，

获取合法的初始URL地址，将这个URL地址作为seed，后续模块将根据这

个“引子’，获取更多URL地址;

（2）微博信息抓取模块:利用合法的URL地址，获取对应的HTML代码，

该模块主要是分析这一HTML代码，从中抽取出微博数据以及相关的微博用户

信息。同时这一模块也需实现获取新的URL地址的功能，并且需要判断微博是