哈爾濱工業大學

毕业设计（论文）

题　　目 微博博主情感倾向分析系统的设计与实现

专　　业 计算机科学与技术

学　　号 110410413

学　　生 白辰甲

指 导 教 师 黄俊恒

答 辩 日 期 2015 年 6 月 23 日

哈尔滨工业大学

摘　要

微博的出现极大的丰富了人们的生活，海量微博数据中蕴含了丰富的信息。情感分析是判断和评估文本情感倾向的过程，其结果对于完善互联网的舆情监控，拓展企业的营销能力有很大帮助。

本文针对新浪微博博主进行情感倾向性分析，主要包括情感分析和兴趣识别两个方面。本文设计了相应的情感分析方法和兴趣识别方法，并最终设计实现了一个基于微博博主的情感倾向分析系统。

本文的主要工作概况为以下几个方面：

（1）新浪微博的数据获取。指定微博博主，系统将爬取该博主的微博正文。

（2）博主情感分析方法的设计。本文使用两种方法分别进行尝试：方法一基于支持向量机模型，通过特征抽取和模型训练得到一个情感极性分类器，可以对微博进行积极和消极的情感分类；方法二基于情感词典，通过对微博中存在的否定词、程度副词、感叹号、以及微博表情符号等进行相应分析处理，从而计算整条微博的情感分值。

（3）博主兴趣识别方法的设计。博主的兴趣识别我们将通过提取博主的兴趣标签来实现，本文使用三种方法分别进行尝试：方法一基于 TF-IDF 权重，通过构建语料库，计算候选词的权重，从而获得关键词并生成标签；方法二基于 TextRank 权重，通过分析微博中词语的共现关系，构建词语网络，抽取较为重要的词用于标签生成；方法三基于 K-means 聚类，词语相似度基于同义词词林进行计算，根据相似度进行词语聚类，随后从每个聚类簇中提取代表词用于标签生成。

（4）结合新浪微博爬虫、博主情感分析方法以及博主兴趣识别方法，本文设计并实现了一个基于微博博主的情感倾向分析系统，并对系统进行测试。

经测试，情感分析的两种方法准确率分别为 75.3%和 71.4%，标签提取的三种方法在人工评价中也取得了较好的效果。

关键词　情感分析；支持向量机；文本分类；聚类

Abstract

The emergence of microblog greatly enriched people's life, huge amounts of microblog contains a wealth of information in the data. Sentiment analysis is a process of judgment and evaluation of text tendency of emotion, the analysis results to improve the Internet public opinion monitoring and expand enterprise's marketing ability are of great help.

In this paper, we analyze the Sina weibo blogger’s emotional tendency, including sentiment analysis and interest identify. We propose a method for sentiment analysis and interest recognition. Finally, we achieved a sentiment analysis system.

The main work of this paper is as follows:

(1)Sina weibo data acquisition. Specify the weibo bloggers, system will crawl blog weibo text of the blogger.

(2)Blogger sentiment analysis method. This paper uses two methods respectively to try: The first method based on support vector machine (SVM) model, through feature extraction and model training, we get an emotional polarity classifier, micorblog can be divided into positive and negative; the second method based on emotion dictionary, word by denying the existence of micro-Bo, the degree adverbs, exclamation point, and microblog emoticons and other corresponding analysis process to calculate the whole microblogging emotional score.

(3)Bloggers interest recognition. Bloggers interest recognition our interest by extracting bloggers label to achieve, we were to try to use three methods: Method one based on TF-IDF weighting, by building the corpus, the right to re-calculate the candidate word to obtain keywords and generate label; method two weights based TextRank by analyzing microblogging words co-occurrence relations, building word network, the more important words extracted for label generation; method three based on K-means clustering, word similarity were based on synonyms Cilin calculations were based on the similarity of word clusters, followed by extraction Representation from each cluster to cluster label generation.

(4)We combine crawler Weibo blogger sentiment analysis methods and bloggers interest recognition, designs and realizes a weibo bloggers emotional tendency analysis system, and testing system.

The accuracy of the two methods of sentiment analysis are 75.3% and 71.4%, three label extract methods in evaluation of artificial and achieved a good result.

Keywords： sentiment analysis，support vector machine，text classification，clustering

目　录

第 1 章 绪论

1.1 课题来源

来源于导师的实际科研项目

1.2 课题背景及研究的目的和意义

微博是一个基于用户关系的信息分享传播以及获取平台，用户可以通过 WEB，WAP 以及各种客户端组件，以 140 字左右的文字更新信息，并实现即时分享。微博给予网络用户更自由更快捷的方式来沟通信息、表达观点、记录心情，已经成为国内最为热门的互联网应用之一。新浪微博是目前国内最大的微博平台，目前其注册用户已突破 3 亿，用户每日发博量超过 1 亿条。

我国目前的微博平台日益完善，用户数目不断增多，基于微博的各种新技术、新应用不断出现，使得微博成为了一个巨大的商业推广平台。微博文本非常口语化，用户多发表自己对社会热点事件的看法，其方便性和易用性正在吸引更多用户的加入。微博之所以能够成为当今国内外的主流社交媒体，主要是因为它具有强大的用户实时交互性。用户在使用微博的过程中，往往在微博网络空间中结成了种种关系。这不仅提高了用户上网体验的满意度，也为商业信息的准确推送提供了广阔的用户群体，微博的实际应用价值为其自身的持续健康发展形成了强劲的驱动力。

基于微博博主的情感分析是指以微博博主为研究对象，分析该博主微博中隐含的情绪状态，从而对该博主的整体情感进行判断或评估。通过对博主的情感倾向进行识别，可以了解和掌握该博主的思想和精神状态，为企业和政府及时提供相关信息。这对于企业的生存、发展和稳定举足轻重，对于政府的管理与和谐社会的构建至关重要。

基于微博博主的标签抽取是指以微博博主为研究对象，提取最能反映博主兴趣的一系列标签，作为对博主兴趣取向的概括描述。其关键点在于识别、描述微博用户的兴趣。通过标签，微博官方和商业服务者可以更好地了解微博用户的兴趣和视角，以便使其提供的服务更加个性化，从而为用户提供更为准确的兴趣推荐服务，扩大自身的营销，同时，用户也能及时得到自身需要的资讯和服务，获得更佳的用户体验。

鉴于以上两点，本文将以微博博主为研究对象，首先通过对博主所发微博进行极性分类和情感打分，统计博主所发积极、消极、中性微博占比，进而分析该博主的整体情感倾向和情感波动情况。其次通过对博主的标签抽取，进而识别博主的兴

趣点和关注点。

1.3 国内外在该方向的研究现状及分析

1.3.1 国外研究现状

在微博情感分析方面，英文主要是针对 Twitter 上用户发布信息作为语料展开分析。Jiang 等[1]运用机器学习方法的方法对 1939 条微博进行训练，采用五折交叉验证的方法对结果进行测试。从测试结果可以看出，通过增加情感词典特征和与主题相关的特征，可以使情感分析结果有所提升。Dmitry Davidiv 等[2]把微博文本中的标签和表情符号作为特征，通过训练语料设计实现了一个 K 近邻的情感分类器，从而实现了微博情感分类。Barbosa 和 Feng[3]使用 Twendz、MicroblogSentiment、TweetFeel 提供的情感分析工具对 Twitter 进行情感分析，并抽取情感标签。他们自行制定规则来对微博进行预处理，去除含有多种情感的微博。经过预处理后带有情感标签的微博被用作训练数据，针对微博的情感分类问题，他们采用二部法：即采用抽象特征训练分类器进行主客观性分类和采用相同特征但修改词的情感极性的权重来进行情感极性分类两种方法。实验结果表明：在上述两大类特征中，按照作用大小排序依次为：负面情感极性最大，其次是正面情感极性、动词、表示正面情感的表情符号，最后为大写字母开头的词的个数。Go 和 Bhayani[4]通过分类文本，提出了一种基于距离的监督学习方法。即通过指定一个关键词，微博被自动被分为正面或负面情感。作者通过抽取含有表情图标的消息作为训练集，利用朴素贝叶斯，最大熵以及 SVM 等分类算法进行了实验，达到 80%以上的精度。

在微博用户的标签抽取方面，Matthew Michelson 和 Yegin Genc 等人[5]使用微博和维基百科资源之间的关系，提取博主微博中的句子实体和评价对象，将其映射到维基百科词条的某个类别节点上，通过统计所有实体所在节点的位置，识别用户感兴趣的维基百科节点，从而对用户的兴趣加以识别。Genc 和 Sakamoto[6]对单条微博所在的领域进行识别，首先抽取出微博中的评价实体，将其映射到相应的维基百科节点。维基百科各节点是一棵有层次结构的树，作者通过开发一个基于路径的算法自动识别每条微博的最终的类别。Wei Wu 等[7]根据微博句子中各个词之间的语义相关性，提出了基于 PageRank 关键词提取的方法抽取用户微博的关键词作为用户标签。

1.3.2 国内研究现状

在国内，近年来关于文本情感分析的研究发展迅速，目前在学术界提出的中文

情感分析的策略主要分为两种：

（1）基于情感词典及语义分析

徐琳宏、林鸿飞[8]从句子的词汇和结构等方面考虑，提取影响语句情感的 9 个语义特征，人工构建了情感词汇本体库，对情感分析的研究做了初步尝试。闻彬，何婷婷等[9]提出一种基于语义理解的文本情感分类方法，在情感词识别中引入了情感义原，通过赋予概念情感语义，重新定义概念的情感相似度，得到词语情感语义值。通过分析语义层副词的出现规律及其对文本倾向性判定的影响，实现了基于语义理解的文本情感分类。实验表明，该方法对于有效的判定文本情感倾向性有很大提高。赵妍妍等[10]提出了一种基于句法路径的方法自动识别情感句中的情感评价单元。句法路径是指链接评价词语和评价对象的一种有向句法结构。事实证明，它可以很好地描述评价词语和评价对象之间的句法关系。作者首先提出了自动采集大量句法路径的方法，继而基于句法路径精确匹配算法来自动获取情感句中的情感评价单元。进一步地，作者还提出了一种基于编辑距离的句法路径匹配改进策略来提高系统的性能。实验结果表明，作者提出的基于句法路径的方法对于情感评价单元的识别是有效的，即句法路径能够较好地挖掘评价词语和评价对象之间的关系。陈晓东[11]对当前已有情感词汇资源加以总结和整理，并运用了扩展的情感倾向点互信息算法对新浪微博语料进行实验，自动获得领域情感词，构建了一个面向中文微博的情感词典。其次，基于中文微博表达多元化的特点，对微博文本进行了相应预处理，并采用微博消息文本中的情感词作为特征选择方法，对微博消息文本中存在的否定词、程度副词、感叹句、反问句、以及微博表情符号等进行相应分析处理。最后对整条微博消息作加权计算获得其情感倾向性，实现了一个面向中文微博的情感倾向分类系统。该方法获得的最高准确率为 74.2%，平均准确率为 70.5%，取得了较好的效果。

总体来看，使用情感词典及与语法语义的方法来分析文本情感，其优点是应用在词语特征级，句子级，粒度细，分析精准。但受到自然语言处理技术及相关抽取技术的限制，该方法容易丢失数据集中隐藏着的重要模式，使得未来研究工作中还有很大的提高空间。

（2）基于统计和机器学习方法

这类方法常用的机器学习模型有：中心向量分类法，朴素贝叶斯(NaiveBayes)，最大熵(MaximumEntropy)，K 最近邻分类和支持向量机(SVM)。唐慧丰等[12]通过用名词、副词、形容词、动词做不同的文本表示特征，以信息增益、文档频率、CHI 统计量和互信息做不同的特征选择方法，分别以中心向量法、贝叶斯分类、K 最近邻和支持向量机做不同的文本分类方法做对比实验，其结果显示：在足够大的训练

集与选择合适特征的情况下，采用 n-Gram 特征表示、信息增益特征选择和支持向量机分类方法，能取得较好的情感分类效果。谢丽星[13]选择了 4 种特征共用，采用支持向量机的方法对新浪微博数据展开情感分析研究。实验结果显示，使用主题无关与主题相关的特征时所获得的最高准确率分别为 66.467%与 67.283%。

基于机器学习方法的情感分析关键在于特征信息的有效提取。优点是分析客观，准确性较高。缺点是对训练语料依赖性比较高，训练周期相对较长。总体来看，使用机器学习方法并不比使用情感词典及与其关联信息方法具有明显的优势，机器学习方法应该有较好的发展空间。

在微博用户的标签抽取方面，方维[14]在分析主流中文微博的信息特点和用户行为特点的基础上，研究了适合针对中文微博系统的信息采集、中文分词、兴趣识别和自动推送技术。提出了文本分类和主题库词匹配相结合的策略，对用户兴趣进行识别；针对微博的时效性特点，采用监听器技术实现对微博信息的实时推送。谢毓彬[15]从生成标签的不同粒度出发，分别从基于关键词和基于类别的角度自动生成微博用户标签，取得了较好的效果。

1.4 主要研究内容

本文选取特定的新浪微博博主进行数据获取，随后基于微博博主进行情感分析和兴趣识别，最后结合以上几点，利用 web 技术设计并实现一个微博博主情感倾向分析系统。本文的具体工作如下：

（1）新浪微博的数据获取。采用基于模拟浏览器的方法，实现对新浪微博的模拟登陆、数据抓取、数据解析、数据存储等功能。指定博主的新浪微博 ID 和需要爬取的页面数量，爬虫将爬取该博主固定页数的微博，作为下一步分析的数据源。

（2）博主情感分析方法的设计。分别使用基于机器学习的方法和基于情感词典方法对爬取的微博进行情感分析。基于机器学习的方法是一个二分分类，可以判断单条微博语句是积极倾向还是消极倾向。基于情感词典的方法可以对微博语句的情感值进行打分，分值为正数表示积极情感，分值为负数表示消极情感，绝对值越大表示情感倾向越强。

（3）博主兴趣识别方法的设计。博主兴趣识别我们通过提取博主的兴趣标签来实现，分别使用基于 TF-IDF 的权重算法、基于 Textrank 的权重算法和基于聚类的方法对博主进行兴趣标签的抽取。前两种方法基于词语权重，第三种方法基于词语相似度。这三种方法都需要首先提取关键词，然后将关键词扩展得到兴趣标签。

（4）基于以上设计，结合新浪微博爬虫、博主情感分析方法以及博主兴趣识别方法，本文设计并实现了一个基于微博博主的情感倾向分析系统，并对系统进行

测试。

1.5 本文结构

本文主要内容如下：

第 2 章中，主要介绍了基于模拟浏览器技术的新浪微博数据爬虫的设计。基本步骤为：新浪微博的模拟登录、爬取指定用户页面的网页源代码、原始页面解析并提取微博正文。

第 3 章中，主要介绍了基于微博博主进行情感分析的方法。分别使用基于机器学习的方法和基于情感词典的方法对特定博主的微博进行情感分析，并对这两种方法进行了相应的比较。

第 4 章中，主要介绍了基于微博博主的兴趣标签提取方法。分别使用三种方法进行尝试，第一种方法基于 TF-IDF 权重算法，首先构建语料库，计算候选词权重，将关键词扩展为标签。第二种方法基于 TextRank 权重，通过分析词语的共现关系，构建词语网络，抽取较为重要的词用于标签生成。第三种方法基于 K-means 聚类，首先提出了基于同义词词林的词语相似度的计算方法，随后进行聚类。从较重要的聚类簇中提取代表词用于标签生成。

第 5 章中，主要介绍基于微博爬虫，情感分析，兴趣标签提取三个模块的微博博主情感倾向分析系统的实现。

第 2 章 微博爬虫的设计与实现

本章主要介绍新浪微博爬虫的设计。采用模拟浏览器技术，基本步骤为：新浪微博的模拟登录、爬取指定用户页面的网页源代码、原始页面解析和提取微博正文。其中新浪微博的模拟登录是前提，解析网页源代码提取正文是关键。

2.1 两种数据挖掘方案的比较

互联网数据的获取通常是通过网络爬虫实现的。在一段网络爬虫程序中，通过设定入口 URL 地址，程序按照一定的爬行策略将网页内容以文本文件的形式保存在本地存储系统中。在新浪微博中可供选择的数据挖掘方案有以下两种。

2.1.1 基于新浪微博 API

基于新浪微博的开放 API 接口可以简洁高效的获取相应的数据。首先，开发者必须通过 OAUTH 认证，该认证使用户在不向第三方透露自己的用户名密码的条件下，使第三方软件提供方申请获得该用户资源的授权。OAUTH 认证为用户资源的授权提供了一个安全、开放而又简易的标准，因此被用于新浪微博 API 的用户验证协议。认证后开发者可以通过调用 API 接口实现新浪微博数据的便捷抓取与解析。新浪 API 可根据请求内容的不同，返回特定的 XML 或 JSON 文件。

部分新浪微博提供的数据读取接口如图 2-1 所示：

微博		
读取接口	statuses/public_timeline	获取最新的公共微博
	statuses/friends_timeline	获取当前登录用户及其所关注用户的最新微博
	statuses/home_timeline	获取当前登录用户及其所关注用户的最新微博
	statuses/friends_timeline/ids	获取当前登录用户及其所关注用户的最新微博的ID
	statuses/user_timeline	获取用户发布的微博
	statuses/user_timeline/ids	获取用户发布的微博的ID
	statuses/timeline_batch	批量获取指定的一批用户的微博列表 高
	statuses/repost_timeline	返回一条原创微博的最新转发微博
	statuses/repost_timeline/ids	获取一条原创微博的最新转发微博的ID
	statuses/mentions	获取@当前用户的最新微博
	statuses/mentions/ids	获取@当前用户的最新微博的ID

图 2-1 新浪微博部分 API 接口示意图

但是基于新浪微博 API 的数据获取方式有以下两点不足：

（1）API 数据获取方式决定了所获得的数据仅限于接口限定的数据内容，因此无法满足不同用户对数据的多样化需求，缺少数据内容的灵活性。

（2）鉴于数据的有价性和网络资源的有限性，官方平台对数据获取频率是有限制的，数据获取频率越高，其收费亦越高，故其数据获取成本比较高。

综合 API 获取数据方式如上所述的优缺点，本文不采用这种方式。

2.1.2 基于模拟浏览器技术

模拟浏览器行为，即指通过程序设计的方式，将正常的人为操作浏览器访问 web 站点的行业进行程序化，从而获得和人为去浏览 web 站点相同的数据。

它的优点主要包括两点：

（1）由于是模拟人为登录的操作流程，故通过浏览器看到的内容，都可以通过这种式获取，而我们分析挖掘的内容恰恰是人们所能看到的数据，看不到的也没有分析的价值，故它可以满足不同用户对数据的多样性需求，增加了数据获取的灵活性。

（2）它不受限于 API 方式的请求频率，只要设计合理、带宽充分就可以大量抓取所需数据。

其难点在于：数据返回的格式是网页源数据格式，一般为 html 代码、json 数据、html+json 混合等多种格式，需要针对不同的数据格式做正则匹配等解析。如果官方平台在某些页面有所改动，往往会影响此时的数据解析，从而影响网络爬虫的准确性。本文采用该方式获取新浪微博数据。

2.2 模拟登录

通过分析浏览器登录新浪微博时的数据包，可以看出在获取 cookie 之前需要进行的一系列预请求。我们将这些预请求进行分析，就可以获取 cookie 请求 URL 的参数。

2.2.1 数据包分析

一般静态的 HTML 页面可以直接通过程序下载网页源代码，经过解析获取到微博正文数据。新浪微博的页面是动态网页，需要登陆后才能获取页面。因此，网页爬虫的第一步就是要实现模拟登录。一般而言，网站的登录过程按实际登陆步骤可以分为 4 步：(1)访问网站初始登录页面；(2)访问中间预处理页面；(3)访问 POST 数据提交页面；(4)访问生成的登陆页。

登录所需要的数据包括：用户名，密码，是否使用代理服务，新浪微博服务器地址，请求新浪微博通行证地址，新浪微博浏览器代理等。这些数据都可以通过抓包分析和查阅新浪微博相关文档获取。

本文以 Firefox 的 Firebug 进行抓包分析，说明新浪微博的登陆过程。首先打开抓包工具，清除原有记录，然后按正常的登陆顺序人工访问网站，输入用户名和密码，直至登陆成功。此时开发者工具可以记录下浏览器与服务器交互的整个过程。图 2-2 显示了新浪微博登录时抓取的数据包：

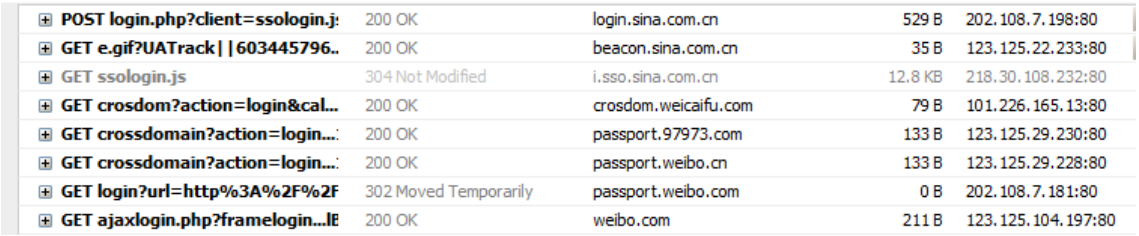

图 2-2 新浪微博登录抓包分析

初始化参数后请求新浪微博服务器，服务器将返回 json 格式的数据。通过正则解析该数据，可以获取登录所需要的 servertime（新浪时间戳），nonce（随机数），pubkey（公钥）等数据，这些数据将用于用户名和密码的加密。表 2-1 总结了新浪微博模拟登录时的基本流程：

表 2-1 新浪微博模拟登录的基本流程

顺序	访问地址	访问类型	发送值	返回值
1	http：//login.sina.com.cn/sso/prelogin.php?entry=microblog&callback=sinaSSOController.preloginCallBack&su=&rsakt=mod&client=ssologin.js(v1.4.18)&_=1407721000736	GET	无	随机数 nonce 和时间戳 servertime
2	http：//login.sina.com.cn/sso/login.php?client=ssologin.js(v1.4.18)	POST	请求头	进一步登录的 URL
3	http：//microblog.com/ajaxlogin.php?framelogin=1&callback=parent.sinaSSOController.feedBackUrlCallBack&ssosavestate=1466045982&ticket=ST-NTI0NjIxMzA3MQ==-1434509982-gz-4FE213069D565EFC335E1E9E5C3A98F7&retcode=0	GET	无	重定向到用户首页的链接

2.2.2 模拟登录

在分析数据包的基础上，我们模拟浏览器请求的方式，封装和发送数据包，从

而实现模拟登录。本节将对关键技术作详细阐述。

2.2.2.1 用户名加密

新浪微博的用户名加密目前采用 Base64 加密算法。 Base64 是一种基于 64 个可打印字符来表示二进制数据的表示方法。由于 2 的 6 次方等于 64，所以每 6 个比特为一个单元，对应某个可打印字符。三个字节有 24 个比特，对应于 4 个 Base64 单元，即 3 个字节需要用 4 个可打印字符来表示。在 Base64 中的可打印字符包括字母 A-Z、a-z、数字 0-9 ，这样共有 62 个字符，此外两个可打印符号根据系统的不同而有所不同。编码后的数据比原始数据略长，为原来的 4/3。

Python 的 hashlib 模块里面包含了多种哈希算法，包括了验证文件完整性的 md5，sha 家族数字签名算法，还有常用的 base64 加密算法包含于 base64 模块中。

实现该算法的伪代码如下：

Input： 用户名 UserName

output： 加密后的用户名 resultNameEncoded

```
1. import urllib，base64
2. urlNameTemp = urllib.quote(UserName)
3. userNameEncoded = base64.encodestring()
4. return userNameEncoded
```

2.2.2.2 密码加密

新浪微博登录密码的加密算法使用 RSA2。需要先创建一个 rsa 公钥，公钥的两个参数新浪微博都给了固定值，第一个参数是登录第一步中的 pubkey，第二个参数是 js 加密文件中的‘10001’。这两个值需要先从 16 进制转换成 10 进制，把 10001 转成十进制为 65537。随后加入 servertime 和 nonce 再次加密。

实现该算法的伪代码如下：

Input： 密码明文 password，时间戳 servertime， 随机数 nonce， 公钥 pubkey

output： 加密后的密码 passwd

```
1. import rsa， binascii
2. rsaPublickey = int(pubkey，16)
3. key = rsa.PublicKey(rsaPublickey， 65537)
4. message = str(servertime) + '\t' + str(nonce) + '\n' + str(password)
5. passwd = rsa.encrypt(message， key)
6. passwd = binascii.b2a_hex(passwd)
```

7. return passwd

2.2.2.3 请求新浪通行证登录

新浪微博请求登录的 URL 地址为：

http：//login.sina.com.cn/sso/login.php?client=ssologin.js(v1.4.18)

请求该 URL 需要发送的报头信息为包括：

{ 'entry'： 'microblog'， 'gateway'： '1'， 'from'： ''， 'savestate'： '7'， 'usertickeˊt'： '1'， 'ssosimplelogin'： '1'， 'vsnf'： '1'， 'vsnval'： ''， 'su'： encodedUserName， 'service'： 'miniblog'， 'servertime'： serverTime， 'nonce'： nonce， 'pwencode'： 'rsa2'， 'sp'： encodedPassWord， 'encoding'： 'UTF-8'， 'prelt'： '115'， 'rsakv'： rsakv， 'url'：'http：//microblog.com/ajaxlogin.php?framelogin=1&callback=parent.sinaSSOController.feedBackUrlCallBack'， 'returntype'： 'META'}

将参数组织好，POST 请求，然后查看查看 POST 后得到的内容。该内容是一个 URL 重定向地址，可以检验登录是否成功。如果地址末尾为“retcode=101”则表示登录失败，为“retcode=0”则表示登录成功。

登录成功后，在 body 中的 replace 信息中的 url 就是我们下一步要使用的 url。然后对上面的 url 使用 GET 方法来向服务器发请求，保存这次请求的 Cookie 信息，就是我们需要的登录 Cookie。

2.3 获取网页源代码

在获取 Cookie 之后就可以开始爬取网页了，本文使用 python 的第三方包 urllib2。urllib2 包是 Python 的一个获取 URLs 的组件。它以 urlopen 函数的形式提供了一个非常简单的接口，同样提供了一个比较复杂的接口来处理一般情况，例如：基础验证，cookies，代理和其他。它们通过 handlers 和 openers 的对象提供。

HTTP 是基于请求和应答机制的，客户端提出请求，服务端提供应答。urllib2 用一个 Request 对象来映射你提出的 HTTP 请求，通过调用 urlopen 并传入 Request 对象，返回 response 对象，然后调用 read 获取网页源代码存入文件中。

2.4 数据解析

从微博平台将数据下载到本地后，网页源代码中是 html 代码、json 值、或是 html 和 json 混合等，本文要进行抽取的是微博博主所发的中文微博。

首先用正则表达式对网页源代码进行解析。具体步骤为：去除 Javascrip 控制语句，去除 CSS 样式表，将
标签转为换行符，将连续的换行转化为一个，取

出 HTML 注释，去除连续的空格，去除多余的标签，取出@后的字符。进行这些初步的处理之后按行写入文件。

第二步抽取微博信息。通过观察可以发现博主所发的微博正文都在<script>FM.view({"ns"："pl.content.homeFeed.index"，"domid"："Pl_Official 之后，我们根据正文所在位置的 class 和 id 对微博正文进行解析。我们组合使用正则表达式和 BeautifulSoup 类库。BeautifulSoup 库类是一个高质量 HTML 解析器，它将 HTML 文件导入后，一次性以树型结构处理完毕，然后用户可以在树型结构内方便的抵达任何一个标签或数据，具有精准定位和查找标签的优势。BeautifulSoup 的劣势在于处理速度要比正则表达式要慢。因此，组合使用正则表达式和 BeautifulSoup 类库，可以满足本文网页解析的需要。

第三步进一步去除噪音。微博正文中会嵌入一些广告，如“360 浏览器”，“百度音乐尊享版”，“百度视频”，“腾讯视频”，“新华网”等，如果该微博已被删除则会出现“抱歉，该条微博已经被删除”，另外还有一些网页提示信息如“正在加载中请稍后”，上述噪音都要进行剔除。

2.5 实验与结果分析

我们选取微博用户“高晓松”的微博进行分析，其新浪微博 ID 为“gaoxiaosong”，我们设置获取该博主的量两页微博。

（1）模拟登录。我们使用自行注册的用户名“weibotest1”和密码“passwordtest1”。经过加密后，和其他数据一起封装成数据包向新浪服务器发送请求。请求新浪通行证后，新浪服务器返回的重定向地址，如表 2-2 所示：

表 2-2 请求新浪通行证重定向地址

地址	http：//microblog.com/ajaxlogin.php?framelogin=1&callback=parent.sinaSSOController.feedBackUrlCallBack&ssosavestate=1462159100&ticket=ST-NTI0NjIxMzA3MQ==-1430623100-gz-3A295BB60B21AB408A699C43D47D8376&retcode=0

可以看出，末尾为“retcode=0”，表示登录成功。

（2）获取网页源代码。使用程序下载该用户主页的源代码，并保存至文本文件中。获取用户“高晓松”的微博主页（gaoxiaosong）的网页源代码如图 2-3 所示：

raw_html.txt

```
<!doctype html>
<html>
<head>
<meta charset="utf-8">
<meta content="高晓松，高晓松的微博，微博，新浪微博，weibo" name="keywords" />
<meta content="高晓松，知名音乐人。高晓松的微博主页、个人资料、相册,清华大学。新浪微博，随时随地分享身边的新鲜事儿。" name="descri
<meta name="viewport" content="initial-scale=1,minimum-scale=1" />
<link rel="dns-prefetch" href="//img.t.sinajs.cn/">
<link rel="dns-prefetch" href="//img1.t.sinajs.cn/">
<link rel="dns-prefetch" href="//js.t.sinajs.cn/">
<link rel="dns-prefetch" href="//js1.t.sinajs.cn/">
<link rel="dns-prefetch" href="//js2.t.sinajs.cn/">
<link rel="dns-prefetch" href="//biz.weibo.com/">
<link rel="dns-prefetch" href="//beacon.sina.com.cn/">
<link rel="dns-prefetch" href="//rs.sinajs.cn/">
<link rel="dns-prefetch" href="//tp1.sinaimg.cn/">
<link rel="dns-prefetch" href="//tp2.sinaimg.cn/">
<link rel="dns-prefetch" href="//tp3.sinaimg.cn/">
<link rel="dns-prefetch" href="//tp4.sinaimg.cn/">
<link rel="dns-prefetch" href="//ww1.sinaimg.cn/">
<link rel="dns-prefetch" href="//ww2.sinaimg.cn/">
<link rel="dns-prefetch" href="//ww3.sinaimg.cn/">
<link rel="dns-prefetch" href="//ww4.sinaimg.cn/">
```

图 2-3 微博用户“高晓松”主页的网页源代码

（3）对网页源代码进行解析，解析后得到的数据存储在文本文件中，每一行代表一条微博。解析后得到微博用户“高晓松”（gaoxiaosong）的微博正文如图 2-4 所示：

chinese_weibo.txt

```
开扒大得州！拥有小牛马刺火箭牛仔和的英雄得州！
为独立，得克萨斯人活捉墨西哥“魂斗罗”总统不熄火，狂放牛仔彪悍如昔上演摩托枪战。得州，英雄之乡，五星上将艾森豪威尔、尼米兹骁勇征战
晓松奇谈之扒一扒美利坚4枪与堕胎药4
为独立，得克萨斯游击队完败墨西哥“魂斗罗”桑塔安纳不熄火，狂放牛仔彪悍如昔化身摩托党上演洒吧枪战。得州，英雄之乡，五星上将艾森豪威
嘿嘿。去年复星郭老大来北美采买企业，我做为门客兼翻译陪他东西海岸转。他随口问我觉得哪家文化企业值得收，我随口说。我不懂商业，只是喜
复星收购太阳马戏团25股权中国企业又在买买买了！今年4月，被称为“加拿大国宝”的太阳马戏团宣布控股权易主，买家是来自中国的复星集团与财
复星购加拿大太阳马戏团股份4
中国企业又在买买买了！今年4月，被称为“加拿大国宝”的太阳马戏团宣布控股权易主，买家是来自中国的复星集团与财团。今天复星公开持股比例
据上海友谊商店记录，购物贵宾中出手最豪的是菲律宾总统马科斯夫人伊梅尔达，最寒酸的是里根总统夫人南希。让一贯崇美轻亚的上海同志们大跌
友谊商店1958年成立，也接待中国人，但只限党政军局级以上干部、统战对象、外宾接待陪同人员。1972年5月商业部、外贸部、外交部联合发布通知
印象中府右街养蜂夹道有个特供站，然而里面并没有珍贵的猪鸡牛肉，只有好多酒。回复特供商店是另外一套系统。在东华门大街，对行政8级以上干
出国人员服务部原在亮马河边，后迁至此。长期海归可买八大件几小件，短期两大两小。也可在门口黑市卖掉指标，亮马时期大件指标两千，到惠新
论斤白送，用途自便有史、有料、有趣、有情、有美好！回复比爱奇艺多一点。
最新力作晓松奇谈同名书来袭！史上最火爆中文脱口秀完整未删节版！将你陈旧伪饰的历史观彻底清盘！预售当当网页链接京东网页链接亚马逊网页
宝贝：预售包邮附书签海报晓松奇谈脱口秀完整未删节版高晓松《晓说》《鱼羊野史》《如丧》后新作出版社直发！4
卖家：凤凰联动图书专营店
十字军帮东罗马守卫君士坦丁堡？昏古七！1204年第四次十字军攻占并洗劫君士坦丁堡，一度灭掉东罗马帝国成立拉丁帝国咱讲过呀，这仇直到800年
看了鸡观和晓说一下搞懵了。突厥人搞定中亚后，成为了穆斯林，然后开始搞土耳其。十字军东征是帮东罗马帝国守卫君士坦丁堡，也就是对抗异教
这个！本来是北京地下水枯竭开始下陷，我为救故土乡亲只身远走加州。加州自此也加速了下陷。我虽只有80千克重，但质量高啊！下一步，去喜马
，矮大紧快说，跟你有没有关系？分享网易新闻美国加利福尼亚州在快速下沉美国加利福尼亚州在快速下沉
```

图 2-4 微博用户“高晓松”的微博正文

2.6 本章小结

本章首先分析了两种新浪微博数据挖掘方案的优缺点，随后介绍了基于模拟浏览器技术的新浪微博数据获取方式。指定微博用户后，通过模拟登录、页面爬取、数据解析等步骤成功获取了特定博主的微博。经过测试，该程序抓取 100 个网页所需的时间在 1 分钟左右。

虽然已经可以满足本文的需要，但程序仍有一些不足之处：

（1）新浪微博的登录加密技术不断更新，发送数据包的内容也会随之改变，使得爬虫程序需要因此改写，并不能做到一劳永逸。

（2）新浪微博具有反爬虫功能，一个用户在一段时间内爬取过多的网页可能会被封号。因此如果想获取更多的数据，则必须准备多个账号一同使用，轮流爬取

数据。这样每个账号平均爬取的数据量变少，从而避免被封号。

（3）暂时还不能实现输入微博用户的中文名就可以抓取其所发微博。原因是用户的中文名和新浪微博给该用户分配的域名并无直接关联。如“姚晨”的新浪微博域名是“yaochen”，“李开复”的新浪微博域名是“kaifulee”，“范冰冰”的新浪微博域名是“fbb0916”，高晓松的新浪微博域名是“gaoxiaosong”。因此本文的解决方案是尽可能搜集更多的博主和域名的对应关系，方便使用者查询。另外使用者也可以直接键入要爬取的用户的微博域名。

总之，程序成功实现了实时抓取数据，为基于微博博主的情感分析和标签提取提供了数据源。

第 3 章 情感分析方法的设计与实现

在获得指定博主的微博后，我们对单条微博进行情感分析。本文采用两种方法分别进行尝试，这两种方法分别基于机器学习技术和情感词典。并对这两种方法的分析结果进行比较。

3.1 情感词库构建

3.1.1 基础情感词典

目前，文本情感分析研究领域有多个高校和科研机构构建的情感词典可供使用。构建一个面向中文微博的情感词典，就需要对现有的词典资源进行总结。目前可用的资源有董振东先生开发的知网[16](HowNet)；张伟、刘缙等人[17]编著的《学生褒贬义词典》；史继林、朱英贵[18]编著《褒义词词典》；杨玲、朱英贵[19]编著的《贬义词词典》；哈尔滨工业大学信息检索实验室整理的《同义词词林扩展版》[20]以及台湾大学整理的中文情感词典[21]（NTUSD）等。

本文采用的情感词典包括知网情感词典，台湾大学情感词典，以及其他情感词典。其中知网情感词典中的正负面情感词语与正负面评价词语是基础，每种类别的词语个数如表 3-1 所示：

表 3-1 知网情感词典情感词数量

词语集名称	词语（个数）
“正面情感”词语	755
“负面情感”词语	1218
“正面评价”词语	3360
“负面评价”词语	3028

选用中文简体版 NTUSD 作为基础情感词典的扩充，并将正面情感词语加入褒义词典，负面情感词语加入贬义词典。然后进行合并去重，去除不常见的情感词。

构建情感词典后将正面情感词存入 pos_all_dict.txt，将负面情感词存入 neg_all_dict.txt。整理后的词典情感词数目如表 3-2 所示：

表 3-2 归纳整理后的情感词数量

词语集	数目
正面情感词	11523
负面情感词	18720

整理后的词典按行存储在文本文件中，如图 3-1 所示：

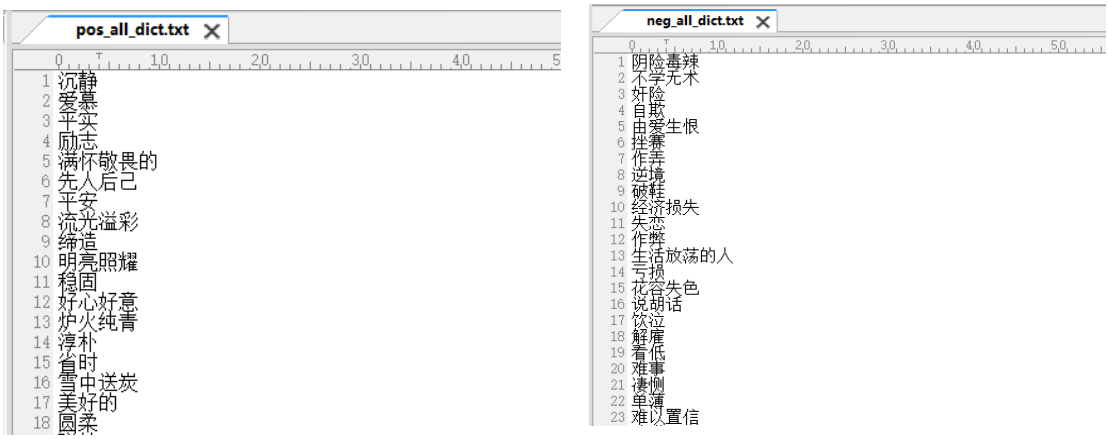

图 3-1 整理后的情感词典

3.1.2 词性标注

中文分词和词性标注技术是对汉语文本进行处理的基础要求，已经有众多机构开发出的多种中文分词系统可供使用。其代表有：结巴中文分词系统、庖丁中文分词系统，纯 C 语言开发的简易中文分词系统 SCWS，中国科学院计算技术研究所推出的汉语词法分析系统 ICTCLAS，哈尔滨工业大学信息检索研究室研制的 IRLAS 等。

本文拟采用结巴分词系统，因为该分词工具已经封装成 Python 语言的一个第三方包，免费使用，对 python 语言有良好的支持，并且可以加入用户自定义词典以提高分词的准确率。

将正面情感词词典 pos_all_dict 和负面情感词词典 neg_all_dict 用结巴分词器进行词性标注，标注后的文本分别存入 pos_all_cut.txt 和 neg_all_cut.txt 中。如图 3-2 所示：

图 3-2 词典标注后的结果

3.1.3 构建分类词典

用结巴分词器进行词性标注，标注后的词性标志及代表意义如表 3-3 所示：

表 3-3 词性标志及意义

标志	a	b	c	d	e	f	i	k
意义	形容词	区别词	连词	副词	叹词	方位词	成语	后接词
标志	l	m	n	o	P	q	R	S
意义	习用语	数词	名词	拟声词	介词	量词	代词	处所词
标志	t	u	v	w	x	y	z	un
意义	时间词	助词	动词	标点	非语素	语气词	状态词	未知词

经过词性标注后，我们选取 18 种词构建分类情感词典，分别是：

（1）8 种积极情感词：a（形容词），b（区别词），d（副词），i（成语），l（习用语），n（名词），v（动词），z（状态词）

（2）8 种消极情感词：a（形容词），b（区别词），d（副词），i（成语），l（习用语），n（名词），v（动词），z（状态词）

（3）2 种特殊词：否定词和转折词。

其中否定词和转折词需要自行收集，收集如表 3-4 所示：

表 3-4 否定词和转折词

否定词	不、没、无、非、莫、毋、勿、未、否、别、無、休、不曾、未必、没有、不要、难以、未曾
转折词	虽然、但是、然而、可以、但、不过

根据词性分类的结果，将不同词性的词语进行归类，存入不同的文本文件中，最终的情感词典如图 3-3 所示：

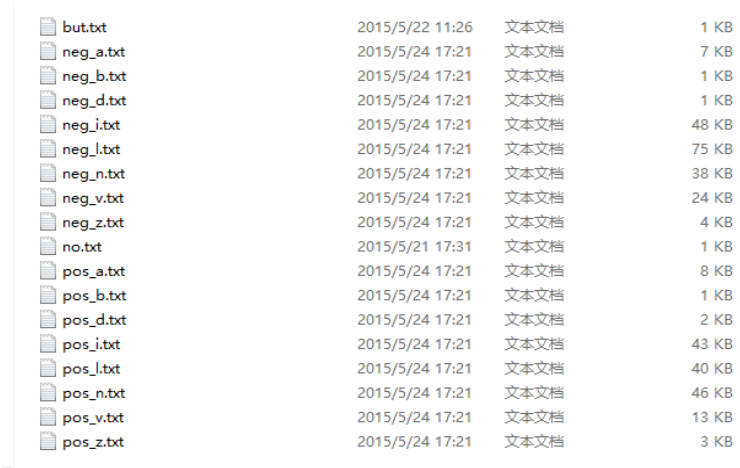

图 3-3 分类词典文本文件

3.1.4 构建程度副词词典

程度副词也是副词的一种，副词一般用于修饰或限制动词与形容词，表示范围、程度等。“程度”是指某个量处于相应层次序列中的某个层级上，是量的层级表现。程度副词的加入使用户在的情感倾向强弱程度上发生了变化，因此需要构建一个程度副词表，并根据程度的高低区分等级。

本文中的程度副词来源于知网情感词典中的“中文程度级别词语”共 182 个。本文将其分为五个等级，如表 3-5 所示：

表 3-5 程度副词词典

级别	代表词
Most	百分之百、倍加、备至、不得了、不堪、不可开交、不亦乐乎、不折不扣、彻头彻尾、充分…（共 64 个）
Very	不少、不胜、出奇、大为、多、多多、多加、多么、分外、非常、格外、够瞧的、够饿、好、好不、何等…（共 40 个）
More	大不了、多、更、更加、更进一步、更为、还、还要、较、比较、较比、较为、进一步、那般…（共 37 个）
Ish	点点滴滴、多多少少、怪、好生、还、或多或少、略、略加、略略、略微、略为、蛮、稍、稍稍、稍微…（共 30 个）
insufficiently	半点、不大、不丁点儿、不甚、不怎么、聊、没怎么、轻度、丝毫、微、相对…（共 11 个）

3.2 基于支持向量机方法的极性分类

3.2.1 支持向量机方法简介

支持向量机(Support Vector Machine)是 Cortes 和 Vapnik 首先提出的，它建立在统计学习理论的 VC 维理论和结构风险最小原理基础上，在解决小样本、非线性及高维模式识别中表现出许多特有的优势，相对于传统的学习方法具有极强的推广能力。其基本思想是构造一个超平面作为决策平面，使正负模式之间的空白最大。

这里我们考虑的是一个二元分类问题，数据点用 x 来表示，是一个 n 维向量。w^T 上标中的“T”代表转置，类别用 y 来表示，可以取 1 或者-1，分别代表两个不同的类别。一个线性分类器就是要在 n 维的数据空间中找到一个超平面，其方程如式（3-1）所示为：

$$w^T x + b = 0 \qquad (3\text{-}1)$$

要确定上述分类函数中的两个参数 w 和 b，可以将 w 看成是是法向量，b 是截距。我们寻找两条边界端或极端划分直线中间的最大间隔，继而引入拉格朗日函数

和对偶变量α，化为对单一因数对偶变量α的求解，从而确定最终的最大间隔分类超平面和分类函数。把寻求分类函数的问题转化为对 w、b 的最优化问题，最终化为对偶因子的求解。

SVM 有三种分类模型：SVC(C-Support Vector Classification.)，NuSVC(Nu-Support Vector Classification.)，，LinearSVC (Linear Support Vector Classification)。这三种分类模型，都是输入两个阵列。一个阵列为X，其维度是$[n_samples, n_features]$，作为训练样本。另一个y（列阵）作为类标签，其维度是$[size(n_samples), 1]$。

3.2.2 训练数据预处理

本文选用哈尔滨工业大学语言处理实验室已经标注的微博语料 44552 条，经人工处理，去除不适合本文的部分语料，共计 44026 条。我们将其整理之后存入文本文件中。其中每行的数据格式均相同，表 3-6 中列举了部分原始训练数据：

表 3-6 部分原始训练数据

COAE2014	#手机	15 日，游人正在手持手机录制或拍摄精彩的现场实况。当日，夜幕降临，华灯初放。在冰城中央大街马迭尔门前“冰城凉台音乐秀”拉开帷幕，悠扬的乐曲、美妙的歌声，吸引了无数的游人驻足观赏，许多人用手机记录下这独具魅力的“冰城凉台音乐秀”盛况。	1	1
COAE2011	#电影	今年更是将会有《幽灵信箱》,《步步惊魂》,《最后的王爷》等多部影视剧现身荧屏，星途可谓一片高涨。	1	1
2012NLP&&CC	#新闻	政府为什么不打击这种恶意的炒作？	1	2

其中原始训练数据各部分所代表的意义如表 3-7 所示：

表 3-7 待训练数据各部分意义

数据来源	所属领域	句子正文	主客观	情感极性

其中主客观部分用 1 代表主观，用 2 代表客观。情感极性部分用 1 代表正面，2 代表反面，3 代表中性，4 代表无立场。显然，我们需要的是 1-1 主观正面和 1-2 主观反面的文本。数据来源和所属领域可以去除，只保留句子正文和情感极性。将处理后的训练数据存储在文本文件中。

下一步使用结巴分词器对训练数据进行分词，分词后根据停用词列表去除微博语句中的停用词。去除停用词后的句子保存在文本文件中。

本文使用的停用词如表 3-8 所示：

表 3-8 停用词表

停用词	的、第二、一番、一直、一个、有的是、也就是说、哎哟、俺们、按照、吧哒、本着、比方、比如、鄙人、彼此、别的、别说、并且、不比、不单、不但、不独、不光、不仅、不拘…（共 347 个）

3.2.3 特征选择

特征选择是情感分类中十分重要的一步，对于特征的选择直接关系到最终的实验效果。根据上一节构建的情感词典，我们选择由 8 种积极情感词，8 种消极情感词以及转折词和否定词作为特征。选取的特征及代表意义如表 3-9 所示：

表 3-9 选取的 18 种特征

序号	名称	描述
1	pos_a	包含积极情感形容词的个数
2	pos_b	包含积极情感区别词的个数
3	pos_d	包含积极情感副词的个数
4	pos_i	包含积极情感成语的个数
5	pos_l	包含积极情感习用语的个数
6	pos_n	包含积极情感名词的个数
7	pos_v	包含积极情感动词的个数
8	pos_z	包含积极情感状态词的个数
9	neg_a	包含消极情感形容词的个数
10	neg_b	包含消极情感区别词的个数
11	neg_d	包含消极情感副词的个数
12	neg_i	包含消极情感成语的个数
13	neg_l	包含消极情感习用语的个数
14	neg_n	包含消极情感名词的个数
15	neg_v	包含消极情感动词的个数
16	neg_z	包含消极情感状态词的个数
17	but	包含转折词的个数
18	no	包含否定词的个数

3.2.4 模型训练

第一步以单条微博为处理单位，根据要抽取的特征，在已经构建好的情感词典中统计每种特征词出现的次数，使单条微博向量化成一个 18 维向量。再按照同样的计算方法依次向量化训练数据中的每条微博，从而构成 SVM 算法的第一个输入

矩阵 X 。微博的极性则构成 SVM 算法的另一个输入矩阵 y 。

第二步进行矩阵运算。使用 python 提供的矩阵处理工具 NumPy 可以快速导入矩阵，特征矩阵导入后存为 data，含有情感极性的列矩阵导入后存为 polarity。

第三步使用 TF-IDF 算法对 data 数据进行加权处理。加权计算的原因是如果我们将单纯的计数数据直接输入分类器，那些频繁出现的词会掩盖那些很少出现但是更有意义的词的频率。为了重新计算特征的计数权重，以便转化为适合分类器使用的浮点值，通常都会进行 TF-IDF 转换。TF 代表词频，而 TF-IDF 代表词频乘以逆向文档频率。关于 TF-IDF 模型的相关理论我们将会在 4.1 节中作详细介绍。

第四步训练模型。本文中采用 1 代表正面情感，2 代表反面情感。将已经导入的 data 和 polarity 矩阵作为 SVM 分类器的两个输入进行模型训练，从而得到极性分类器。

3.2.5 极性分类

将要测试的数据逐行读入，进行向量化。然后用已训练好的 SVM 分类器对特征向量的情感极性进行预测。如果输出值为 1，则为正面情感，如果输出值为 2，则为负面情感。

3.3 基于情感词典的情感打分

本节采用基于情感词典的打分策略，对单条微博的情感分值进行计算。第一步进行句子切分，记录分句中存在的否定词、程度副词、感叹号、以及微博表情符号等，并进行相应的分析处理。最后对整条微博作加权计算获得其情感倾向分值。具体步骤如图 3-4 所示：

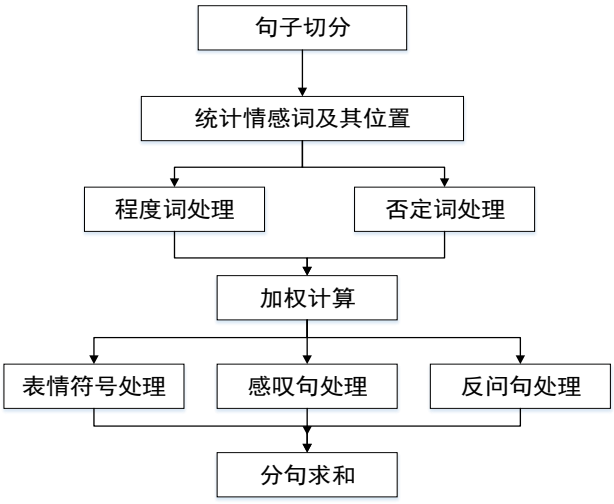

图 3-4 基于情感词典的情感打分算法基本步骤

3.3.1 情感词处理

首先对爬虫爬取到的微博数据进行分词，并按照表 3-8 给出的停用词列表去除停用词。我们根据标点符号所在的位置将微博分成多个分句，以每个分句作为研究对象，根据 3.1 节构建的情感词典，记录每个分句中积极情感词和消极情感词出现的位置和个数。初始将积极情感词的权值设置为 1，消极情感词的权值设置为-1。

我们以一条微博语句为例进行分析：“这手机的画面极好，操作也比较流畅。不过拍照真的太烂了！系统也不好。”我们将其分句后进行情感词处理后的结果如表 3-10 所示：

表 3-10 情感词处理举例

分句	句子分词	积极情感词	消极情感词	积极得分	消极得分
1	这/手机/画面/极/好，	好	-	1	0
2	操作/也/比较/流畅。	流畅	-	1	0
3	不过/拍照/真的/太/烂/了！	-	烂	0	1
4	系统/也/不/好	好	-	1	0

从结果可以看出，很显然这个分值的计算是不合理的，如第 4 个分句，情感词“好”之前有否定词，应该表示的是消极情感，而如果只考虑情感词的话就会把句子归为积极情感。另外第 3 个分句情感词“烂”之前有程度副词，并且句子后有感叹号来加重语气。我们在下一节将针对修饰词进行分析。

3.3.2 修饰词的处理

3.3.2.1 程度副词

程度副词词典构建已经在 3.1 节中做了详细阐述。本文将程度副词的等级分为 5 类，分别设定不同的权值，如表 3-11 所示：

表 3-11 程度副词权值

类别	权值
Most	2.0
Very	1.75
More	1.5
Ish	1.2
insufficiently	0.5

我们仍然以微博语句“这手机的画面极好，操作也比较流畅。不过拍照真的太

烂了！系统也不好。”作分析，分析过程如表 3-12 所示：

表 3-12 程度副词分析举例

分句	句子分词	积极情感修饰词	消极情感修饰词	积极得分	消极得分
1	这/手机/画面/极/好，	极	-	2.0*1	0
2	操作/也/比较/流畅。	比较	-	1.5*1	0
3	不过/拍照/真的/太/烂/了！	-	太	0	2.0*1
4	系统/也/不/好	-	-	1	0

3.3.2.2 否定词

否定词列表也已经在 3.1 节中作了详细说明。否定词是副词的一种，它是表示否定意义的词语，被否定词修饰的情感词往往会改变情感极性。当一个否定词修饰一个正面情感词，则原本表达的正面情感就会转变为负面情感，反之则反。

由于汉语中存在多重否定现象，即当否定词出现奇数次时，表示否定意思；当否定词出现偶数次时，表示肯定意思。若情感词所在位置的前方有奇数个否定词修饰时，须将原分句的情感极性取反，即乘以 -1；若有偶数个否定词修饰，则不做处理。

我们继续处理微博语句“这手机的画面极好，操作也比较流畅。不过拍照真的太烂了！系统也不好。”通过观察可以发现否定词出现在分句 4，否定词的出现使该句原有的积极情感变为消极情感。分析过程如表 3-13 所示：

表 3-13 否定词分析举例

分句	句子分词	积极情感否定词	消极情感否定词	积极得分	消极得分
1	这/手机/画面/极/好，	-	-	2.0	0
2	操作/也/比较/流畅。	-	-	1.5	0
3	不过/拍照/真的/太/烂/了！	-	-	0	2.0
4	系统/也/不/好	不	-	0	1

3.3.2.3 感叹号

感叹句是以抒发感情为主的句子，它所抒发的感情有赞美、愉悦、愤慨、叹息、惊讶、哀伤等，句末通常都用感叹号来标识。微博消息中的感叹句多为用户所表达情感的增强。感叹号“！”主要用在感叹句的句末，表示强烈的感情。某种程度上说，它是感叹句存在的标志。若在分句的句末出现感叹号，则如果该分句原本的情感得分为正，则分值加 2；如果该分句原本的情感得分为负，则分值减 2。

微博语句“这手机的画面极好，操作也比较流畅。不过拍照真的太烂了！系统

也不好。”中的分句 3 中出现了感叹词，有加重语气的作用，如表 3-13 所示：

表 3-13 感叹号分析举例

分句	句子分词	积极情感感叹号	消极情感感叹号	积极得分	消极得分
1	这/手机/画面/极/好，	-	-	2.0	0
2	操作/也/比较/流畅。	-	-	1.5	0
3	不过/拍照/真的/太/烂/了！	-	存在	0	2.0+2.0
4	系统/也/不/好	-	-	0	1

3.3.2.4 表情符号

新浪微博提供了多种表情符号供用户选择，表情符号暗含了感情色彩，一些用户常常使用合适的表情符号来直接表达心情。在微博消息中，表情符号的加入不但使文本信息充满了个性化色彩，而且还为分析用户情感倾向带来了帮助。本文整理了新浪微博中相关的表情符号，其中表示正面情感的 37 个，负面情感的 49 个。依据情感程度把正面和负面各分为两个等级。如表 3-14 所示：

表 3-14 新浪微博标签符号及权值设置

类别	个数	权值	内容
正面情感	12	2	[好得意]，[哈哈]，[太开心]，[鼓掌]，[ok]，[good][耶]，[赞]，[给力]，[威武]，[爱你]，[haha]
	25	1	[bobo 抛媚眼]，[红包]，[呵呵]，[嘻嘻]，[可爱]，[亲亲]，[抱抱]，[钱]，[酷]，[心]，[蜡烛]，[蛋糕]，[话筒]，[礼物]，[熊猫]，[兔子]，[奥特曼]，[互粉]，[手套]，[吃饭]，[思考]，[顶]，[握手]，[右抱抱]，[左抱抱]
负面情感	19	-2	[怒火]，[闭嘴]，[鄙视]，[泪]，[生病]，[吐]，[怒]，[悲伤]，[抓狂]，[阴险]，[怒骂]，[伤心]，[失望]，[挖鼻屎]，[愤怒]，[最差]
	30	-1	[可怜]，[吃惊]，[害羞]，[偷笑]，[懒得理你]，[右哼哼]，[左哼哼]，[嘘]，[衰]，[委屈]，[打哈气]，[疑问]，[馋嘴]，[汗]，[困]，[花心]，[哼]，[晕]，[猪头]，[不要]，[弱]，[挤眼]，[睡觉]，[书呆子]，[黑线]，[拜拜]，[感冒]，[拳头]，[围观]，[囧]，[神马]，[浮云]

若在分句的中出现相应的表情符号，则在分句原有情感分值的基础上加上相应的表情权值。

3.3.2.5 疑问句

疑问句通常分为两种，一种是有疑而问，另外一种是反问。第一种疑问句跟本

文情感分析没有多大关系，主要考虑第二种。反问句的目的往往是加强语气，把原本的思想表达更加强烈、鲜明。

在处理时，我们首先在分句末尾查看是否存在“?”标志，如果存在，则向前查找是否存在反问标记词。如果找到，则证明是反问句。本文中将“?”的权值设置为-2，即当存在反问句时，其权值直接为-2。

本文使用的反问句标记词如表 3-15 所示：

表 3-15 反问标记词

权值	反问标记词
-2	为什么、凭什么、难道、何必、怎能、怎么能、怎么会、怎会、哪能、能不、能没、不都、不也、不就、谁叫、谁让、就算、这算、还算、就不、还不、莫非…

3.3.3 情感得分的加权计算

首先对单条微博进行文本预处理，并以标点符号为分割标志，将单条微博分割为 n 个句子 S_1、S_2、S_3 … S_n ，提取每个句子中的情感词 W_i。以下两步的处理均以分句为处理单位。

第二步在情感词表中寻找情感词，以每个情感词为基准，向前依次寻找程度副词、否定词，并作相应分值计算。随后对分句中每个情感词的得分作求和运算。

第三步判断该句是否为感叹句，是否为反问句，以及是否存在表情符号。如果是，则分句在原有分值的基础上加上或减去对应的权值。

最后对该条微博的所有分句的分值进行累加，获得该条微博的最终得分。

具体的分值计算步骤如下：

（1）当出现程度副词 M_{W_a} 修饰情感词 W_i 时，该情感词的情感倾向分值计算公式如式（3-2）所示：

$$O_{W_i} = M_{W_a} * S_{W_i} \tag{3-2}$$

其中 M_{W_a} 表示程度副词的权值，S_{W_i} 是句子中情感词 W_i 的权值。

（2）当出现否定词 W_b 修饰情感词 W_i 时，则该情感词的情感极性取反，如式（3-3）所示：

$$O_{W_i} = M_{W_b} * S_{W_i} \tag{3-3}$$

其中 M_{W_b} 表示否定词的权值，S_{W_i} 是句子中情感词 W_i 的权值。

（3）统计分句 S_i 中包含的 k 个情感词的分值 W_1、W_2、W_3 … W_k，进行累加，

如式（3-4）所示：

$$O_{S_i} = \sum_{i=1}^{k} O_{W_i} \tag{3-4}$$

（4）当在该分句中出现表情符号时，则该分句的情感倾向分值如式（3-5）所示：

$$O_{s_i} = M_{w_c} + O_{s_i} \tag{3-5}$$

其中 M_{w_c} 表示表情符号的权值，O_{s_i} 是分句的情感倾向分值。

（5）当句子为感叹句时，若原来句子情感倾向得分为正，在分值加 2，反之，分值减 2。

（6）当句子为反问句时，该分句的情感倾向分值直接设置为 -2。

（7）含 n 个分句的微博消息 d_i 的最终情感倾向计算公式如式(3-6)所示：

$$O_{d_i} = \sum_{i=1}^{n} O_{S_i} \tag{3-6}$$

经过对修饰词的加权计算和分句的求和运行，我们最终得到的最终情感倾向值 O_{d_i} 作为情感打分的结果。单条微博的情感得分为正代表积极情感，得分为负代表消极情感，分值的绝对值越大表示情感倾向性越强。

3.4 实验与结果分析

3.4.1 基于支持向量机方法

（1）首先进行模型训练

将已标注情感极性的待训练数据存入 data.txt，共计 44026 条。如图 3-5 所示：

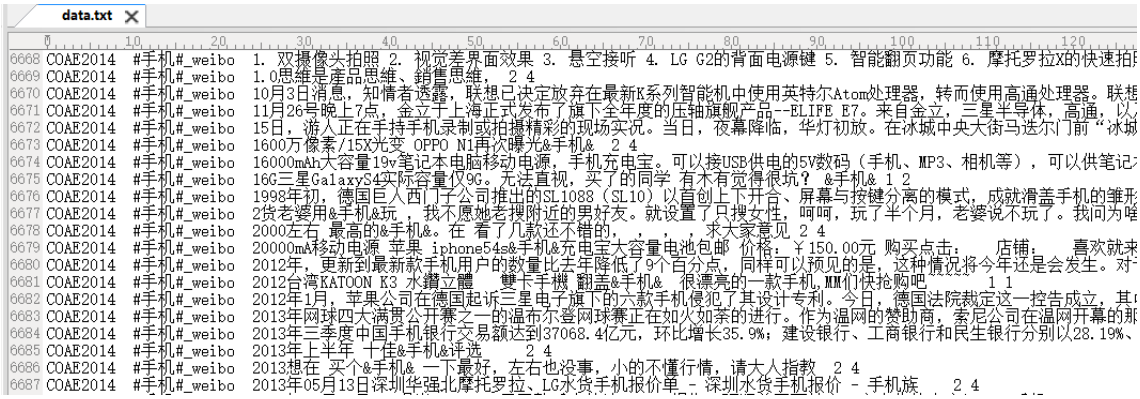

图 3-5 已标注极性的待训练语料

对训练数据进行语料预处理，去除无用成分，保留有情感极性的微博正文，根据停用词列表去除停用词。处理后的文本如图 3-6 所示：

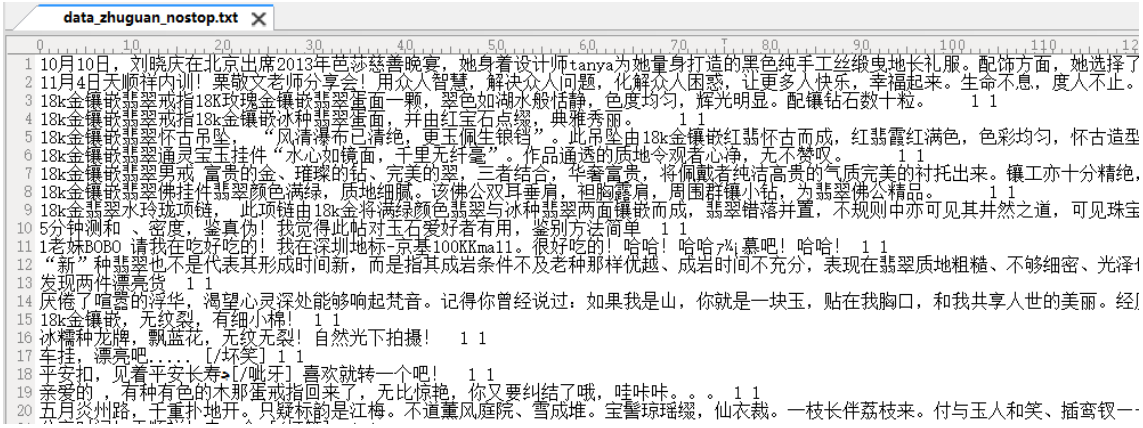

图 3-6 训练数据预处理结果

根据构建好的分类词典，对训练语料特征抽取，每条微博向量化为 18 维向量，向量化后的结果如图 3-7 所示：

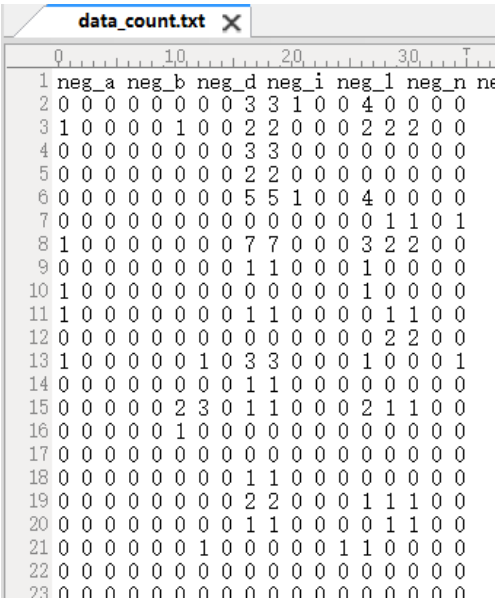

图 3-7 训练数据向量化

（2）将向量化后的数据导入 SVM 分类器中进行模型训练，随后用训练好的模型进行情感极性判别。我们选择微博用户“高晓松”（gaoxiaosong）的微博进行测试，测试结果如图 3-8 所示：

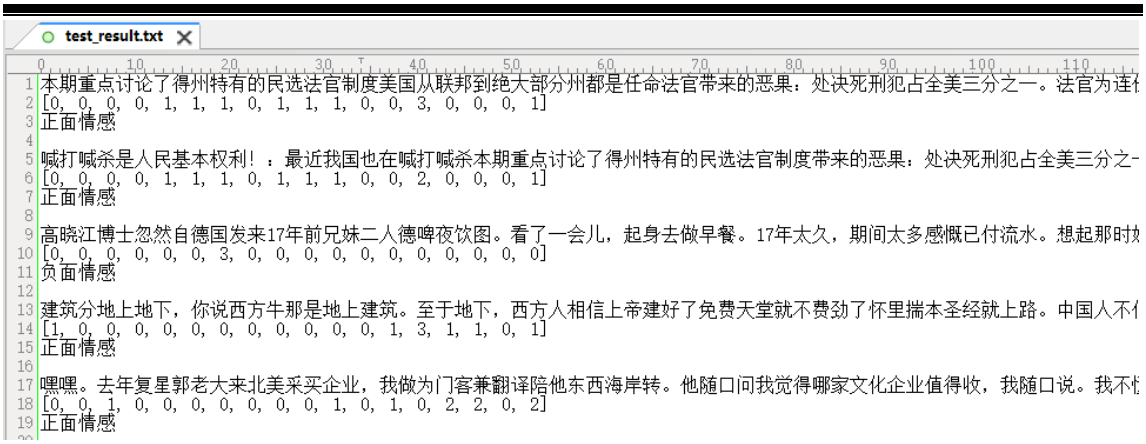

图 3-8 “高晓松”微博的情感极性分析结果

其中数据第一行为微博正文，第二行为该条微博向量化后的结果，第三列为分类器分类得到的结果。

（3）准确率计算。我们随机抓取了新浪微博文本 1000 篇，经过去重、页面内容提取和情感极性的人工标注后，从中筛选出正向文本 309 篇、负向文本 301 篇、中性文本 298 篇。使用基于支持向量机的算法来对抓取到的文本进行极性判断。所得的实验结果如表 3-16 所示：

表 3-16 基于支持向量机算法的情感分析结果

倾向性	准确率	召回率	F-Measure
正向	78.6%	69.5%	73.8%
负向	72.4%	80.9%	76.4%
平均	75.3%	75.3%	75.1%

3.4.2 基于情感词典方法

（1）情感打分。仍然选择第二章爬取到的微博用户“高晓松”近期的微博进行情感打分，测试结果如图 3-9 所示，其中微博前的数字即是计算得到的情感倾向分值。

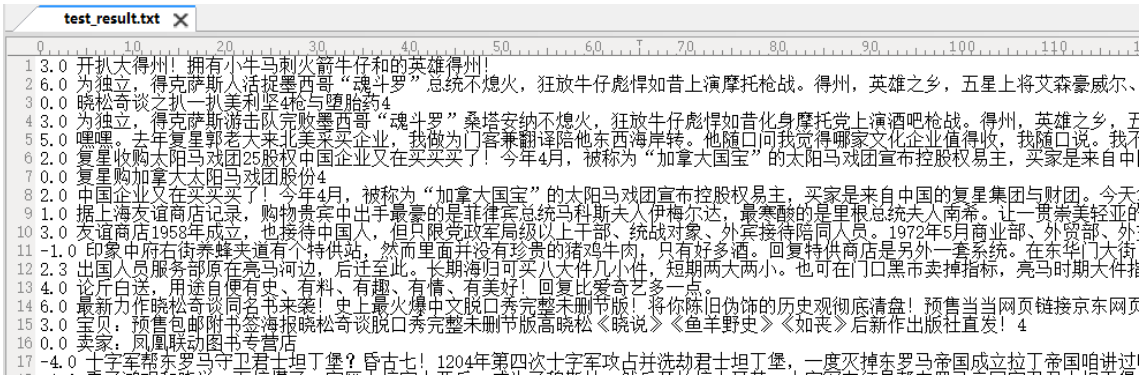

图 3-9 用户“高晓松”微博的情感打分结果

经过情感打分，我们可以得到每条微博的情感分值，情感分值为正表示正面情感，情感分值为负表示负面情感，情感分值为零表示中性情感，绝对值越大表示情感倾向性越强。

（2）统计分析。我们可以由单条微博的情感分值得到博主总体的情感倾向，我们统计所有微博的得分情况，在此基础上做统计。统计指标包括积极、消极、中性微博的占比，情感倾向得分的方差等。随后根据这些数据对该博主的情感倾向和情感波动情况作出一些人为的评价。

对微博用户“高晓松”的 53 条微博的统计评价结果如表 3-17 所示：

表 3-17 用户“高晓松”的总体情感评价

评价指标	数值	意义
积极情感微博数量	25	含积极情感微博 25 条，占比 47%
消极情感微博数量	15	含消极情感微博 15 条，占比 28%
中性情感微博数量	13	含中性情感微博 13 条，占比 25%
积极情感平均得分	3.3	积极情感平均得分较低，情感平和
消极情感平均得分	-4.9	消极情感平均得分中等，情感平和
情感得分方差	21.1	情感值方差较大，情感波动明显
整体情感倾向性	0.2	整体情感倾向接近于 0，情感倾向不明显

（3）准确率计算。我们使用随机抓取了新浪微博文本 1000 篇，经过去重、页面内容提取和情感极性的人工标注后，从中筛选出正向文本 309 篇、负向文本 301 篇、中性文本 298 篇。使用基于情感词典的打分算法来对抓取到的文本进行情感打分。所得的实验结果如表 3-18 所示：

表 3-18 基于情感词典的情感打分分析结果

倾向性	准确率	召回率	F-Measure
正向	72.9%	64.4%	76.6%
负向	68.3%	63.4%	68.9%
平均	71.4%	66.3%	67.1%

3.5 两种方法的对比与分析

针对正负极性的二元分类来说，基于支持向量机的的方法精确度更高，达到了 75.3%，而基于情感词典的方法准确率只有 71.4%。因为词典匹配会由于语义表达的丰富性而出现很大误差，而机器学习方法不会。机器学习方法适用的场合更为广泛，它无需像词典匹配那样要深入到词语、句子、语法这些层面。

而词典方法适用的语料范围更广，无论是手机、电脑这些商品，还是书评、影

评这些语料，都可以适用。并且词典方法可以提供比正负极性这样的描述更加具体的分值，我们可以利用这些分值做相应的统计分析，从而对博主的情感倾向做出更为全面的判断。机器学习的缺点是则极度依赖语料，语料必须经过人工标注，而且对语料的领域有一定的要求，例如使用 IT 相关的微博训练出来的的分类器给艺术相关微博做情感分类，其效果一定不好。

3.6 本章小结

本章首先构建了情感分析所用的情感词典，随后分别使用基于 SVM 的机器学习算法和基于情感词典的权值打分算法对微博进行了情感分析，取得了不错的效果。

第 4 章 兴趣标签提取方法的设计与实现

标签一般是由多个名词组成的短语，用于表示微博博主的兴趣点和关注点。在本章中，我们尝试根据博主所发微博自动提取标签，基本流程为：将微博文本进行预处理获取候选名词，通过某种算法排序得到微博关键词，再基于规则将关键词扩展为用户的标签。本章使用三种方法对博主所发的微博进行标签提取，并对三种算法的效果和适用情况进行简单对比分析。

4.1 基于 TF-IDF 权重的兴趣标签提取

4.1.1 算法介绍

TF-IDF 算法的主要思想是：如果词 w 在一篇文档 d 中出现的频率高，并且在其他文档中很少出现，则认为词 w 具有很好的区分能力，适合用来把文章 d 和其他文章区分开来。该模型主要包含了两个因素：

（1）词 w 在文档 d 中的词频 TF (Term Frequency)，即词 w 在文档 d 中出现次数$count(w,d)$和文档 d 中总词数$size(d)$的比值，计算公式如式（4-1）所示：

$$tf(w,d)=\frac{count(w,d)}{size(d)} \tag{4-1}$$

（2）词 w 在整个文档集合中的逆向文档频率 IDF (Inverse Document Frequency)，即文档总数 n 与词 w 所出现文件数$docs(w,D)$比值的对数，计算公式如式（4-2）所示：

$$idf=\log\frac{n}{docs(w,D)} \tag{4-2}$$

则词 w 在文档 d 中的 TF-IDF 权值如式（4-3）所示：

$$TF-IDF(w,d)=tf(w,d)*idf \tag{4-3}$$

4.1.2 标签提取步骤

在该方案中，我们将借助于 TF-IDF 算法，为微博博主自动生成标签。流程图如图 4-1 所示。

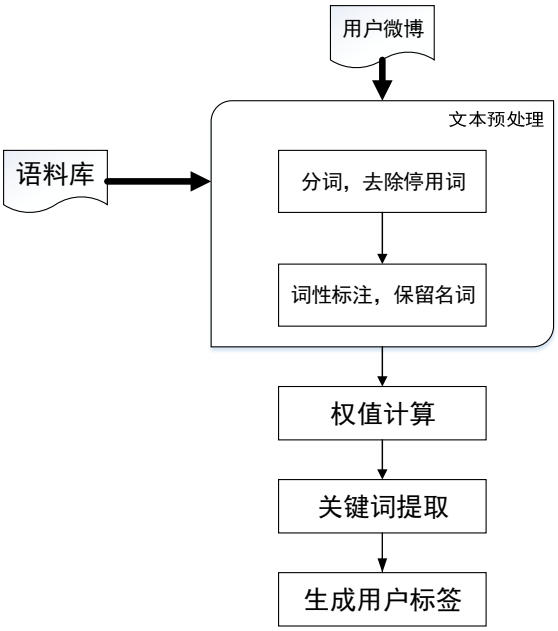

图 4-1 TF-IDF 标签提取基本流程

4.1.2.1 语料库构建

我们在第二章中实现的爬虫随机爬取微博名人榜中 30 位博主的微博，平均每个博主的微博为 700 条，构建算法适用的语料库。语料库以文本文件的形式存储，如图 4-2 所示：

图 4-2 基于 TF-IDF 的生成方法的语料库

其中每个文件的文本格式均相同，每一行代表用户所发的一条微博。例如 0.txt 的文本格式如图 4-3 所示：

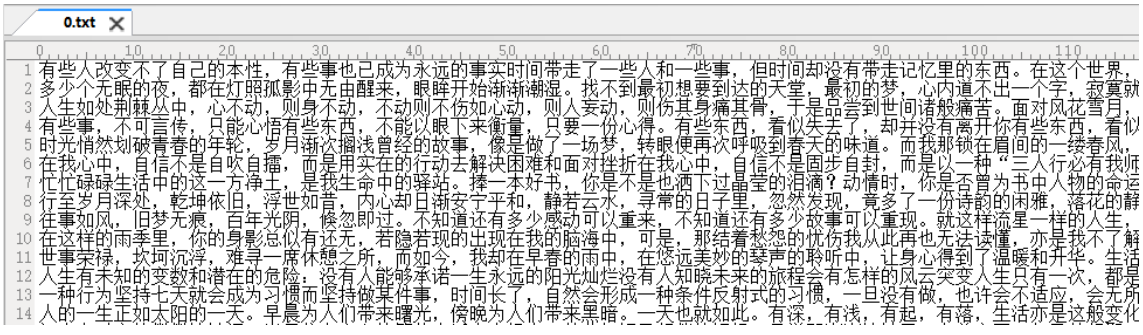

图 4-3 语料库的格式

我们对语料库进行文本预处理。首先剔除无用成分，然后用分词器分词，并对照停用词列表去除停用词。考虑到关键词一般为名词，因此我们将分词后的词语进行词性标注，只保留名词作为候选词。图 4-4 显示了 0.txt 经处理后的结果：

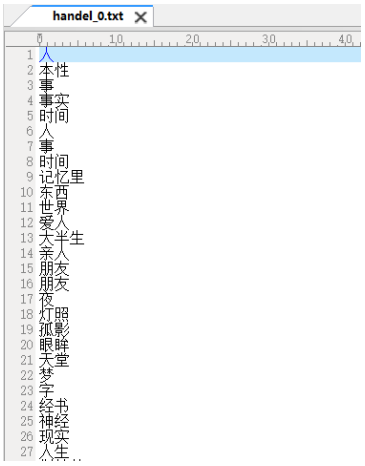

图 4-4 文本预处理后的候选词

4.1.2.2 选取关键词

语料库构建好之后，我们对待分析博主的微博按照同样的方法进行文本预处理，保留名词作为候选词。

将语料库中 30 个微博用户的候选名词存为 30 个文档，加上待分析的博主，文档总数为 31 个。其中第 31 个文档是待提取关键词的用户的微博。

下一步利用公式（4-4）计算待提取博主微博中每个名词的权值：

$$tfidf(t,u) = tf(t,u) * \log(\frac{U}{U_t}) \qquad (4\text{-}4)$$

其中，$tf(t,u)$ 表示用户 u 的微博文本中词 t 的频率，U 表示微博语料中用户的总数，U_t 表示微博文本中包含词语 t 的用户数。

计算所有名词的权值后，我们将其按照权值排序，选取权值排名前 100 的词作为关键词。下一步我们将用提取到的关键词扩展成标签。

4.1.2.3 提取标签

由 TF-IDF 权重算法计算得到的单个名词可能不足以表达用户的兴趣。因此，我们将基于规则，以名词为中心，扩展成关键词串，作为标签。

我们采取的方法是，选取权值排名前 100 的名词作为关键词，如果存在关键词的组合在原文中相邻，则将其组合作为候选标签。考虑到标签的稳定性，我们仅抽取出在原文中出现过 3 次或以上的关键词组合作为标签。标签的权重为组成它的词语的 TF-IDF 权重之和。

4.1.3 实验与结果分析

我们选取微博用户“高晓松”的微博作为实验数据，首先进行分词标注，去除停用词，保留候选名词。处理后的结果如图 4-5 所示：

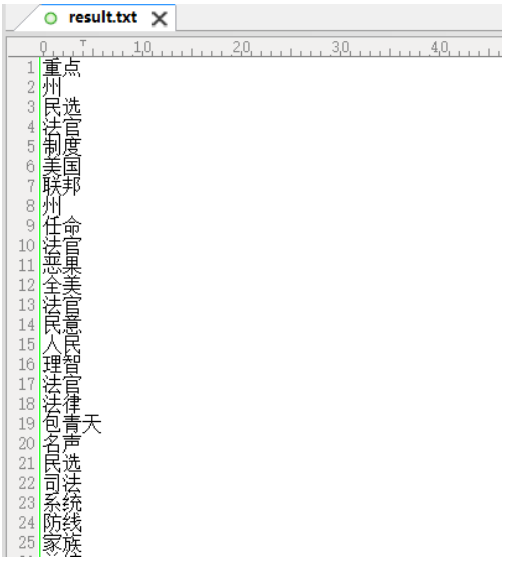

图 4-5 用户“高晓松”的微博预处理后保留的名词

第二步我们将这 31 个用户微博中的候选名词组成 31 个文档，进行 TF-IDF 权值计算，然后我们按照权值进行排序，选取排名前 100 的名词作为关键词。用户“高晓松”微博中提取的关键词如表 4-1 所示：

表 4-1 用户“高晓松”微博中提取出的 100 个关键词

关键词	晓松、高晓松、台湾、战俘、马可波罗、日本、历史、阴影、总统、中国、韦小宝、法官、友谊商店、犹太人、蒋介石、复兴、周记、观感、人民、二战、北平、妄人、战俘营、鱼羊、加州、朝鲜、美国、日军、加西亚、大和族、西语、野史、节目、文盲、法语、马戏团、大伙、民选、美利坚、内政外交、十字军、历史文献、未央、梅厄、满洲、琉球、秘史、财税、金马、铁哥们、民主、战争、伟业、共同点、师弟、情色、血泪、德国、中场、国民党、笔记、上晓松、不同点、两面派、侨民、加利福尼亚、参议院、台独、外宾、嫌隙、德意日、爵位、特准、猛料、茉莉花、莱茵、蒋经国、夫人、惨案、民意、皇帝、解密、东北、启示录、奖杯、细思、谍战、遗孤、彩蛋、灾难、主题曲、同名、安宁、文艺、名将、堪比、天文、扬州、法西斯、趣事

第三步进行关键词扩展，将扩展后的按照权重排序后，抽取排名前 10 的名词词组作为用户标签。如表 4-2 所示：

表 4-2 使用 TF-IDF 算法为“高晓松”提取的标签

标签	TF-IDF 权值
台湾历史	0.356704052
台湾观感	0.3234141449
台湾人民	0.3222189356
台湾民主	0.299123369
东北台湾	0.2888415171
台湾启示录	0.2879995838
日军战俘	0.2372689885
德国战俘	0.2188288689
日本遗孤	0.1376890251
民选法官	0.1336022353
美国犹太人	0.1334810182

4.2 基于 Textrank 权重的兴趣标签提取

4.2.1 算法介绍

Textrank 算法是一种类似于 PageRank 的图模型算法。TextRank 将文档中的词语类比于互联网网页，而词与词之间的联系类比于网页之间的链接关系。也就是说，算法认为文本是一个由词语构成的网络或者说是一个由词语作为节点构成的图，词之间的语义关系构成边。在图中越重要的词，也就越可能是关键词。

形式化地，我们令$G(V,E)$代表文本中由词语构成的有向图，V为词语节点，E为边。对于每一个节点V_i，$In(V_i)$代表指向它的节点集合，$Out(V_i)$代表节点V_i指向的节点集合。W_{ij} 代表V_i和V_j之间边的权重。我们确定一个滑动的文本窗口，窗口中包含 k 个词，倘若两个词语同时出现在这个窗口中，我们可以称它们共现。可以将词对间的共现次数作为连接它们的边的权重。节点V_i的分数计算如式（4-5）所示：

$$S(V_i) = (1-d) + d * \sum_{V_j \in In(V_j)} \frac{W_{ji}}{\sum_{V_k \in Out(V_j)} W_{jk}} S(V_j) \qquad (4\text{-}5)$$

由公式可以看出，Textrank 权值的计算是一个迭代的过程。其中d一般设置 0.85。

可以看出，TextRank 的主要思想是：一个词的重要性由其他与其关联的词决定。算法流程如下：

（1）确定文本的最佳代表形式：词语或者单个字或者其他，并将其作为图中

的节点；

（2）构建节点之间的边，例如共现信息当做权重；

（3）迭代，直到算法收敛；

（4）将节点按照分数排序，得到关键词。

4.2.2 标签提取步骤

在该方案中，我们将借助于 Textrank 权值算法，为微博博主自动生成标签。流程图如图 4-6 所示。

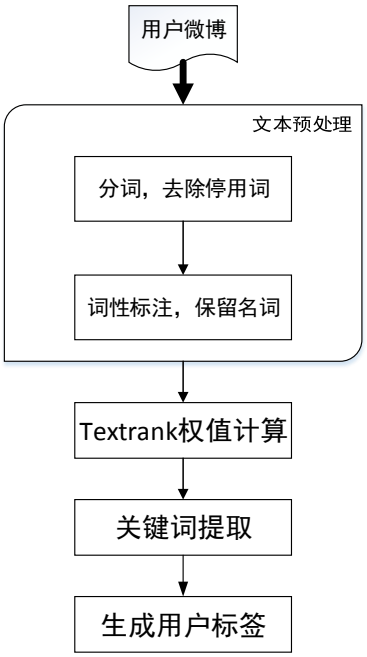

图 4-6 基于 Textrank 自动生成标签流程

4.2.2.1 选取关键词

我们首先将待分析博主的微博进行文本预处理，分词后去除停用词。然后进行词性标注，只保留名词作为候选词。

然后为用户微博文本构建以候选关键词（即选取的名词）为节点的无向图。将共现的窗口定为一条微博的长度，即倘若两个词在同一条微博中出现，我们就认为它们之间存在较强的语义联系，共现次数加 1。我们对每一条微博进行同样的词对共现次数提取。随后，图节点间边的权重记为它们在该用户微博文本中的共现次数。

图构建完之后，我们开始计算每个名词节点的分数，计算公式（4-5）适用于有向图，我们根据无向图进行适当调整，如式（4-6）所示：

$$S(V_i) = (1-d) + d * \sum_{V_j \in E(V_i)} \frac{W_{ji}}{\sum_{V_k \in E(V_j)} W_{jk}} S(V_j) \quad (4\text{-}6)$$

其中 $E(V_i)$ 表示与节点 V_i 连接的所有节点集合。常数 d 的值通常设置为 0.85。分数计算是个迭代的过程，当所有词排列顺序不变时，我们将停止计算。选取权值排名前 100 的名词作为关键词。

4.2.2.2 提取标签

我们采用与 TF-IDF 算法相同的关键词扩展办法来提取标签。获取 Textrank 权值排名前 100 的关键词排序列表，查看其两两组合是否有原文中出现 3 次或三次以上。如果存在，则作为候选标签。标签的权重为组成它的词语的 TextRank 分数之和。按照权重排序后，抽取权值排名前 10 的标签作为博主标签。

4.2.3 实验与结果分析

我们选取微博用户“高晓松”的微博作为实验数据进行文本预处理，预处理方式与 TF-IDF 方法的相同，结果如图 4-5 所示，此处不再赘述。

下一步进行关键词提取，首先按照词语之间的共现关系构建以候选名词为节点的无向图。然后按照公式（4-6）依次计算每个词的权值，迭代计算完成后，选取权值排名前 100 的词作为关键词。微博用户“高晓松”微博的关键词提取结果如表 4-3 所示：

表 4-3 基于 Textrank 权值的“高晓松”微博关键词

关键词	台湾、晓松、高晓松、日本、中国、历史、人民、战俘、阴影、美国、总统、法官、节目、友谊商店、德国、马可波罗、朝鲜、二战、战争、日军、加州、韦小宝、蒋介石、法律、无法、犹太人、大伙、复星、上海、大师、时候、青春、东北、中文、周记、电影、部分、音乐、鱼羊、时代、故事、人生、国民党、北平、北京、夫人、文艺、加西亚、法语、妄人、世界、名字、野史、民主、感情、战俘营、全球、信息、十字军、中场、大和族、文盲、笔记、满洲、安宁、饭店、琉球、灾难、未央、民选、博士、文化、太阳、马戏团、同志、作品、女人、书单、西语、国家、民众、女性、网络、老师、情色、排行榜、传统、果儿、美利坚、谍战、侨民、特准、遗孤、华夏、两面派、谜团、堪比、百姓、作者、回国

最后进行标签提取，将提取到的关键词进行扩展，将扩展后词组按照权重排序后，抽取排名前 10 的名词词组作为用户标签。实验结果如表 4-4 所示：

表 4-4 使用 Textrank 算法为“高晓松”提取的标签

标签	Textrank 权值
台湾历史	0.0212420137
台湾人民	0.0190969072
东北台湾	0.0158901396
台湾民主	0.0155605392
晓松作品	0.0139010732
高晓松作品	0.013020726
日本遗孤	0.0106359406
中国感情	0.0103177712
德国战俘	0.007705754
朝鲜人民	0.0076714628
日军战俘	0.0072299419

4.3 基于聚类的兴趣标签提取

通过 4.1 节和 4.2 节的两种基于权重的抽取方法可以在一定程度上反映用户的兴趣，但是可以看出，当用户在某一方面的兴趣非常明显时，将会出现某一方面的标签堆积的现象。如 4.2 节的测试结果中“台湾历史”和“台湾人民”意义相近，可以只保留一个标签。因此，我们尝试用聚类分析的方法，抽取博主的标签。

基于聚类分析方法抽取标签的大致流程如下：

（1）将爬取到的制定博主的微博进行预处理，得到名词作为候选关键词

（2）计算词语相似度，进行聚类分析

（3）从每个聚类簇中选取代表词，扩展成用户标签

4.3.1 关键技术与原理

4.3.1.1 同义词词林简介

哈尔滨工业大学信息检索实验室利用现有的词典资源，完成了“同义词词林扩展版”的编写。“同义词词林扩展版”收录词语近 7 万条，按照语义相似性进行编排，并按照树状的层次结构把所有收录的词条组织到一起。

“同义词词林扩展板”将词汇分成大、中、小三类，大类有 12 个，中类有 97 个，小类有 1400 个。每个小类里都有很多的词，这些词有根据词义的远近和相关性分成了若干个词群。每个词群中的词语又进一步分成了若干个行，同一行的词语要么词义相同，要么词义有很强的相关性。小类中的段落可以看作第四级的分类，段落中的行可以看作第五级的分类。

这样，同义词词林就具备了 5 层结构。如图 4-7 所示：

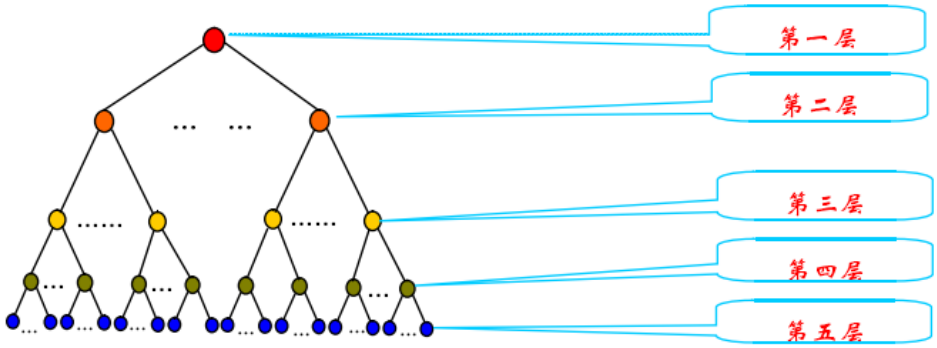

图 4-7 同义词词林的层次结构图

随着级别的递增，词义刻画越来越细，到了第五层，每个分类里词语数量已经不大，很多只有一个词语，已经不可再分，可以称为原子词群。

“同义词词林扩展板”提供了五层编码，第一层用大写英文字母表示，第二层用小写英文字母表示，第三层用二位十进制整数表示，第四层用大写英文字母表示，第五级用二位十进制整数表示。如表 4-5 所示：

表 4-5 同义词词林的编码方式

编码位	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**
符号举例	A	a	1	5	B	0	2	=\#\@
符号性质	大类	中类	小类		词群	原子词群		
级别	第 1 级	第 2 级	第 3 级		第 4 级	第 5 级		

表中的编码位是按照从左到右的顺序排列。第八位的标记有 3 种，分别是“=”、“#”、“@”。其中“=”代表“相等”、“同义”；“#”代表“不等”、“同类”，属于相关词语；“@”代表“自我封闭”、“独立”，它在词典中既没有同义词，也没有相关词。

4.3.1.2 基于同义词词林计算语义相似度

中文的一个词语往往包含了多层意思，也就是说有很多个义项。例如：“骄傲”既可以表示褒义“自豪”，也可以表示贬义“傲慢”。因此计算词语相似度要考虑到所有的义项。本文计算词语相似度的主要思想是：基于同义词词林结构，利用词语中义项的编号，根据两个义项的语义距离，计算出义项相似度。

首先判断在同义词词林中作为叶子节点的两个义项在哪一层分支，即两个义项的编号在哪一层不同。例如 Aa01A01 与 Aa01B01 在第 4 层分支。

随后从第 1 层开始判断，相同则乘 1，否则在分支层乘以相应的系数，然后乘

以调节参数$\cos(n*\frac{\pi}{180})$。其中n是分支层的节点总数，该调节参数的功能是把义项相似度控制在[0，1]之间。然后再乘以一个控制参数$\frac{n-k+1}{n}$，其中n是分支层的节点总数，k 是两个分支间的距离。

两个义项的相似度分数用sim表示，计算方法如下：

（1）若两个义项在第一层分支，计算方法如式（4-7）所示：

$$sim(A,B)=f \tag{4-7}$$

（2）若两个义项在第二层分支，则系数为 a，如式（4-8）所示：

$$sim(A,B)=1*a*\cos(n*\frac{\pi}{180})(\frac{n-k+1}{n}) \tag{4-8}$$

（3）若两个义项在第三层分支，则系数为 b，如式（4-9）所示：

$$sim(A,B)=1*1*b*\cos(n*\frac{\pi}{180})(\frac{n-k+1}{n}) \tag{4-9}$$

（4）若两个义项在第四层分支，则系数为 c，如式（4-10）所示：

$$sim(A,B)=1*1*1*c*\cos(n*\frac{\pi}{180})(\frac{n-k+1}{n}) \tag{4-10}$$

（5）若两个义项在第五层分支，则系数为 d，如式（4-11）所示：

$$sim(A,B)=1*1*1*1*d*\cos(n*\frac{\pi}{180})(\frac{n-k+1}{n}) \tag{4-11}$$

若两个义项的编号相同，即在同一行内时，则考虑末尾的尾号，当尾号为“=”时，为同义词，相似度为 1；当编号的尾号为“@”时，则代表这个词既没有同义词也没有相关词，不予考虑。当尾号为“#”时，直接把定义的系数 e 赋给结果，如式（4-12）所示：

$$sim(A,B)=e \tag{4-12}$$

其中系数：$a=0.65, b=0.8, c=0.9, d=\ 0\ 96, e=0.5, f=0.1$

4.3.1.3 聚类技术

K-means 聚类算法是很典型的基于距离的聚类算法，采用距离作为相似性的评价指标，即认为两个对象的距离越近，其相似度就越大。其中本文中词语之间的距离采用基于同义词词林的语义相似度来衡量。聚类簇是由距离靠近的对象组成的，因此把得到紧凑且独立的簇作为最终目标。

假设我们提取到原始数据的集合为(**x**1，**x**2，…，**x***n*)，并且每个 **x**i 为 d 维的向量，K-means 聚类的目的就是：在给定分类组数 k（$k\leq n$）值的条件下，将原始数据分成 k 类 **S** = {*S*1，*S*2，…，*Sk*}，在数值模型上，即对以下表达式求最小值，

如式（4-13）所示：

$$\sum_{i=1}^{k}\sum_{X_j \in S_i}\left\|X_j - \mu_i\right\|^2 \qquad (4\text{-}13)$$

算法步骤如下：

（1）从 D 中随机取 k 个元素，作为 k 个簇的各自的中心。

（2）分别计算剩下的元素到 k 个簇中心的，将这些元素分别划归到相似度最高的簇。

（3）根据聚类结果，重新计算 k 个簇各自的中心。

（4）将 D 中全部元素按照新的中心重新聚类。

（5）重复第 4 步，直到聚类结果不再变化。

4.3.2 标签提取步骤

基于聚类分析的标签自动生成，我们将借助 K-means 聚类算法进行，整个生成标签的流程如图 4-8 所示。

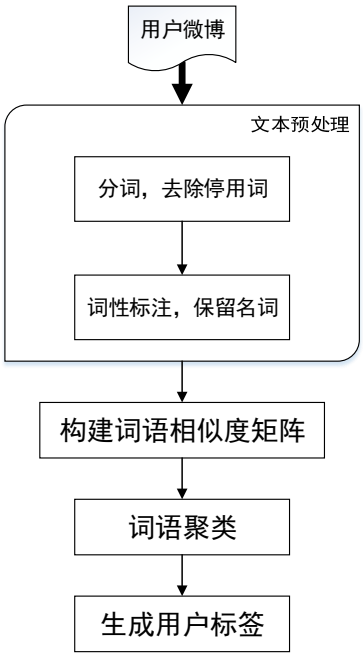

图 4-8 基于聚类技术的标签提取算法流程

（1）文本预处理

预处理过程与基于 TextRank 的生成方法相同，预处理后得到一个候选的关键词集合，词语的词性都为名词。

（2）K-means 聚类

为了减少计算量，我们缩小待聚类的词语集合，选用 TextRank 权重前 500 的

词聚类，我们认为 Top500 的词语集合已经体现了用户的绝大多数兴趣。

在聚类算法开始之前，其实我们需要为词语构造初始距离矩阵。在此处是一个 500*500 的矩阵，矩阵元素是词语之间的相似度。词语相似度的计算方法已经作过相应的阐述，我们采用基于同义词词林的方法依次计算每两个词之间的相似度，构建矩阵。

矩阵构建好之后采用 K-means 聚类算法进行聚类，其中 K 取 10，即将 500 个候选词聚成 10 个簇。

（3）选取簇代表词，扩展

对于聚类形成的每一个簇，选用簇中拥有最高 TextRank 分数的词语作为簇代表词。选取完簇代表词后，按照基于 TextRank 生成方法中同样的策略对词进行扩展，扩展得到名词性词组，注意与代表词合并的词语必须出现在同一个聚类簇中。这样，每一个聚类簇将会提取得到几个名词性词组，我们选用在 Textrank 分数最高的词组作为该簇的标签。将这些标签按照 Textrank 分数排序，从而得到最终标签。注意，并非每个聚类簇都能成功提取出标签。

4.3.3 实验与结果分析

测试用户仍然选择微博用户“高晓松”，我们将 Textrank 分数排名前 500 的词依据语义相似度进行聚类后的效果如图 4-10 所示：

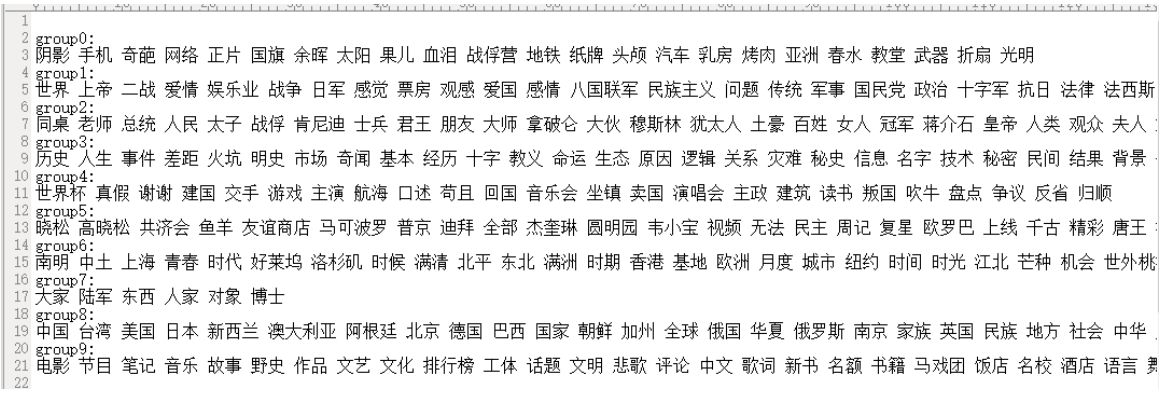

图 4-10 用户“高晓松”微博的 500 个关键词进行词语聚类的结果

聚类之后，在每个簇中提取提取一个标签。考虑到由于将语义相似的词聚到同一类后，他们组成的名词性词组在原文中出现的概率较低，因此有的聚类簇并不能成功提取名词性词组构成的标签。对于不能提取到标签的簇，我们选用簇代表词作为标签。当然我们也可以通过扩大候选名词的数量或者减少簇的个数，来提高标签抽取的效果，但是这样会增加算法的时间复杂度。

最终标签提取的效果如表 4-6 所示：

表 4-6 使用聚类方法提取标签的结果

标签	Textrank 权值
博士	0.00132131355495
晓松高晓松	0.024310028958
台湾	0.014139194604
人民代表	0.00568405850197
历史小事	0.00760562608976
阴影	0.00461891813166
上海	0.00193708939137
法律	0.00204329356285
民选	0.00132838214549
节目	0.00331093793353

4.4 三种方法的对比与分析

就生成标签的总体效果而言，人工评定的结果表明 Textrank 生成方法和聚类生成方法都比基于 TF-IDF 的方法要好，其中基于聚类分析的方法略优于基于 TextRank 的方法。

基于 TF-IDF 的标签提取方法需要额外的微博语料，而其他两种方法则不需要。本文中爬取了另外 30 个微博用户的微博作为语料库。如果能够进一步扩大语料库，则该方法的标签抽取效果将会提高。另外，基于 TF-IDF 的方法提取到的标签侧重于体现该博主与其他博主的不同兴趣点，更能体现博主的个性化。

基于 Textrank 的标签提取方法不需要额外的微博语料，而且算法运行速度快。该算法抽取的标签侧重于体现原文中多次提到或经常关联到的兴趣点，能够较好的体现博主的兴趣，但在体现博主的个性方面不如 TF-IDF 算法。

基于聚类分析的方法基于 Textrank 权值，可以避免同义标签堆积的现象，但就程序运行时间来讲会成倍提升，其时间消耗主要在计算语义相似度构建矩阵方面。通过聚类方法提取的标签能在更多的维度上体现用户的兴趣。

4.5 本章小结

本章为微博博主自动生成标签，介绍了三种方法：第一种是基于 TF-IDF 的方法，第二种是基于 TextRank 的方法，第三种是基于聚类分析的方法。我们分别对这三种方法进行了实验，并分析了实验结果，同时进行了对比分析。

第 5 章 系统集成与测试

本文所涉及到的微博博主情感倾向分析分为三个部分，分别是新浪微博爬虫的设计、情感分析和标签提取。本文分别针对这三个模块在第 2、3、4 章作了详细阐述。本章将以这三部分为基础，构建一个完整的微博博主情感倾向分析系统，并介绍其设计方法和测试效果。

5.1 系统集成

系统首先使用新浪微博爬虫爬取指定博主的微博，随后构建情感词典，对爬取的微博分别使用基于支持向量机的极性分类方法和基于情感词典的情感打分方法进行情感分析，并对分析结果做相关统计评价。随后对该博主的微博进行标签提取，对爬取到的微博分别使用基于 TF-IDF 的生成方法、基于 Textrank 的生成方法和基于聚类的生成方法进行标签提取。

如图 5-1 所示，系统的整体结构被分成三个模块：新浪微博爬虫设计，情感分析和标签提取。

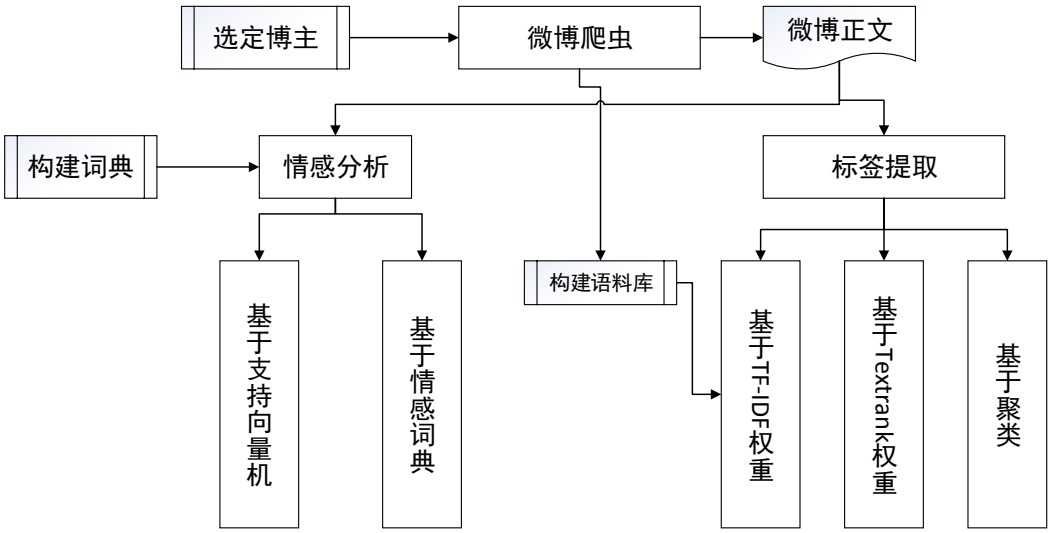

图 5-1 系统设计模块图

5.2 系统测试

5.2.1 新浪微博爬虫模块测试

本文实现了新浪微博的数据爬虫，爬虫运行时由用户指定要爬取的微博博主。为了方便使用，系统预先设置了一些知名的博主可供选择。用户也可以自行输入要爬取博主的 ID。图 5-2 和 5-3 分别显示了爬取微博用户“姚晨”和“高晓松”的界面。

图 5-2 爬取“姚晨”的新浪微博示意图

图 5-3 爬取“高晓松”的新浪微博示意图

5.2.2 情感分析模块测试

在爬取到相关用户的微博后，我们进行下一步，即对微博进行情感分析。

5.2.2.1 情感词典的构建

我们首先人工构建情感词典，词典构建好之后进行词性分类，分别提取 18 种词构成分类词典。

未经分类的微博情感词典显示界面如图 5-4 所示：

图 5-4 微博情感词典

词性标注后生成的 8 种积极情感词组成的分类词典如图 5-5 所示：

图 5-5 八种积极情感词分类词典

词性标注后生成的 8 中消极情感词组成的分类词典如图 5-6 所示：

图 5-6 八种消极情感词分类词典

转折词和否定词词典的显示界面如图 5-7 所示：

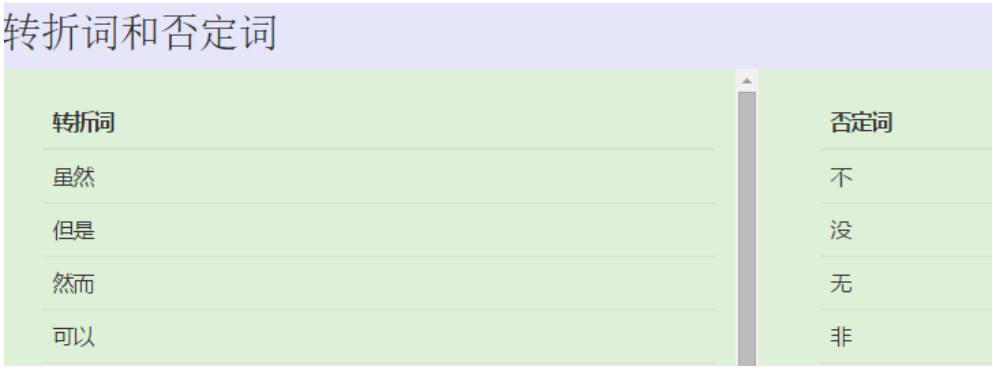

图 5-7 转折词和否定词词典

程度副词词典及其权值设置的显示界面如图 5-8 所示：

图 5-8 程度副词词典

5.2.2.2 基于 SVM 的极性分类

首先进行训练模型，对训练数据进行预处理，并按照选定的特征进行向量化。

图 5-9 和图 5-10 分别显示了经过预处理后的训练数据和训练数据向量化后的结果。

句子	极性
10月10日，刘晓庆在北京出席2013年芭莎慈善晚宴，她身着设计师tanya为她量身打造的黑色纯手工丝缎曳地长礼服。配饰方面，她选择了黑白两色钻石项链镶嵌方形顶级老坑玻璃种 [/呲牙]，设计简约而不失精致，完美衬托出高贵迷人的气质。	1
11月4日天顺祥内训！栗敬文老师分享会！用众人智慧，解决众人问题，化解众人困惑，让更多人快乐，幸福起来。生命不息，度人不止。（图：天顺祥董事长杨俊杰与栗敬文老师合影、分享会天顺祥总经理刘焦萍与伙伴们分享）	1
18k金镶嵌翡翠戒指18K玫瑰金镶嵌翡翠蛋面一颗，翠色如湖水般恬静，色度均匀，辉光明显。配镶钻石数十粒。	1
18k金镶嵌翡翠戒指18k金镶嵌冰种翡翠蛋面，并由红宝石点缀，典雅秀丽。	1
18k金镶嵌翡翠怀古吊坠，“风清瀑布已清绝，更玉佩生锒铛”。此吊坠由18k金镶嵌红翡怀古而成，红翡霞红满色，色彩均匀，怀古造型简约大方，清新雅致，又颇具古风，将中国的传统文化融入到翡翠设计当中，完美传达了东方时尚翡翠文化的独特魅力。	1

图 5-9 预处理后的训练数据

neg_a	neg_b	neg_d	neg_i	neg_l	neg_n	neg_v	neg_z	pos_a	pos_b	pos_d	pos_i	neg_l	pos_n	pos_v	pos_z	but	no
0	0	0	0	0	0	0	0	3	3	1	0	0	4	0	0	0	0
1	0	0	0	0	1	0	0	2	2	0	0	0	2	2	2	0	0
0	0	0	0	0	0	0	0	3	3	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	2	2	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	5	5	1	0	0	4	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	0	1
1	0	0	0	0	0	0	0	7	7	0	0	0	3	2	2	0	0

图 5-10 训练数据向量化结果

模型训练好之后，我们需要对爬取到的微博进行向量化，随后用训练好的模型进行分类，分类结果可以显示单条微博的正负极性。图 5-11 显示了对微博用户“高晓松”的微博进行向量化和正负极性分类的结果：

句子	向量	极性
新奥尔良是我最爱的美国城市，密西西比河蜿蜒，法国风情盎然，尤其美食结合法国西班牙美南精华，小龙虾鳄鱼螃蟹各种海鲜荟萃。吃饱了找个酒馆边喝边听全世界最好的爵士乐，最后坐着帅小伙蹬车的跺而归。每年早春时节狂欢节前后去住几天，排泄一年的苟且。	[0, 0, 0, 0, 0, 1, 0, 0, 1, 1, 0, 0, 0, 0, 1, 1, 0, 0]	正面情感
将在外，三人意外促成路易斯安那购地，美国领土骤然翻倍。再称王，“小强”拿破仑牵制英军突出重围力挺美国，新奥尔良法式风情犹在。大风大浪，飓风来袭路易斯安那因何沉没？小情小调，小龙虾怎样烹饪惹人垂涎？高晓松娓娓道来，浓情开扒色香味俱全的诱人路易斯安那！晓松奇谈之扒一扒美利坚6拿破仑与小龙虾	[1, 0, 0, 0, 0, 0, 2, 0, 0, 0, 0, 0, 0, 1, 2, 2, 0, 0]	正面情感
晓松奇谈之扒一扒美利坚6拿破仑与小龙虾4	[0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0]	负面情感
三兄弟意外促成路易斯安那购地，美国领土骤然翻倍。“小强”拿破仑牵制英军突出重围力挺美国，新奥尔良法式风情延续至今。昨日港口贩奴罪行累累，今朝自由开放吸引黑人常住。大风大浪，飓风来袭路易斯安那能否沉没？小情小调，小龙虾怎样烹饪惹人垂涎？高晓松浓情开扒色香味俱全的诱人路易斯安那！	[0, 0, 0, 0, 0, 0, 2, 0, 1, 1, 0, 0, 0, 1, 4, 4, 0, 0]	正面情感

图 5-11 对“高晓松”的微博进行极性分类

5.2.2.3 基于情感词典的情感打分

基于情感词典对每条微博的情感倾向进行打分，以单条微博中作为分析对象，统计其中存在的否定词、程度副词、感叹号、以及微博表情符号等，并进行相应的分析处理。最后对整条微博作加权计算获得其情感倾向分值。图 5-12 显示了对微博用户“高晓松”的微博进行情感打分的效果：

微博语句	情感得分
新奥尔良是我最爱的美国城市，密西西比河蜿蜒，法国风情盎然，尤其美食结合法国西班牙美南精华，小龙虾鳄鱼螃蟹各种海鲜荟萃。吃饱了找个酒馆边喝边听全世界最好的爵士乐，最后坐着帅小伙蹬车的咏而归。每年早春时节狂欢节前后去住几天，排泄一年的苟且。	3.8
将在外，三人意外促成路易斯安那购地，美国领土骤然翻倍。再称王，“小强”拿破仑牵制英军突出重围力挺美国，新奥尔良法式风情犹在。大风大浪，飓风来袭路易斯安那因何沉没？小情小调，小龙虾怎样烹饪惹人垂涎？高晓松飓尾道来，浓情开扒色香味俱全的诱人路易斯安那！晓松奇谈之扒一扒美利坚6拿破仑与小龙虾	0.0
晓松奇谈之扒一扒美利坚6拿破仑与小龙虾4	0.0
三兄弟意外促成路易斯安那购地，美国领土骤然翻倍。“小强”拿破仑牵制英军突出重围力挺美国，新奥尔良法式风情延续至今。昨日港口贩奴罪行累累，今朝自由开放吸引黑人常住。大风大浪，飓风来袭路易斯安那能否沉没？小情小调，小龙虾怎样烹饪惹人垂涎？高晓松浓情开扒色香味俱全的诱人路易斯安那！	4.0
找了两圈并没找到，在加州也没见过加州牛肉面！一定是我打开方式不对：帮我打包一份新奥尔良烤翅新奥尔良是我最爱的美国城市，密西西比河蜿蜒，法国风情盎然，尤其美食结合法国西班牙美南精华，小龙虾鳄鱼螃蟹及各种海鲜。吃饱找个酒馆听全世界最好的爵士乐。	4.8

图 5-12 对“高晓松”的微博进行情感打分

得到每条微博的情感得分后，我们可以进行简单的统计分析。本文人为设置了 7 项指标，分别为：积极情感微博数量、消极情感微博数量、中性微博数量、消极情感平均的得分、积极情感平均得分、情感得分方差、整体情感倾向等。我们依据统计数值，人为划定范围，可以对博主的情感倾向做出一些评价。

图 5-13 显示了各项指标的统计数值以及人为评价的结果。

指标名称	数值	意义
积极情感微博数量	118	积极微博条数为 118 条，占全部微博比例的 %52.0
消极情感微博数量	66	消极微博条数为 66 条，占全部微博比例的 %29.0
中性情感微博数量	43	中性情感微博条数为 43 条，占全部微博比例的 %19.0
积极情感平均得分	3.3	平和。他感觉到所有的一切都生机勃勃并光芒四射，虽然这人眼里世界却是一个。所以头脑保持长久的沉默，不再人，观照者消融在观照中，成为观照本身。
消极情感平均得分	-3.6	愤怒。如果有人能跳出冷漠和内疚的怪圈，并摆脱恐惧的折感，接着引发愤怒。愤怒常常表现为怨恨和复仇心里望，来自比之更低的能量级。挫败感来自于放大了欲望的侵蚀一个人的心灵。
情感得分方差	16.7	情感波动较大，周围的喜悦或者悲伤都能轻易的感染他
整体情感倾向	0.7	淡定。到达这个能级的能量都变得很活跃了。淡定的能到来这个能级，意味着对结果的超然，一个人不会再经

图 5-13 对“高晓松”微博的总体情感进行评价

5.2.3 标签提取模块测试

5.2.3.1 基于 TF-IDF 的标签提取方法

我们首先人工构建语料库，语料库越大提取标签也越准确，但计算时间将会越长。依旧针对微博用户“高晓松”所发微博，计算微博中候选名词的 TF-IDF 权值，按照权值排序后提取权值前 100 的词作为候选关键词。图 1 显示了从“高晓松”的

微博中提取出的关键词以及 TF-IDF 权值。

候选词	TF-IDF权值
晓松	0.631438513538
高晓松	0.350799174188
民国	0.15271808134
笙箫	0.149185950618
台湾	0.132826994052
美利坚	0.11741117342
那英	0.11741117342
战俘	0.105239752256

图 5-14 TF-IDF 方法对用户“高晓松”的微博提取关键字

随后我们以关键词为中心，扩展成名词性词组作为标签。选取权值排名前 100 的名词进行组合，如果它们的组合在原文中出现过 3 次或以上，则作为候选标签。图 5-15 显示了最终标签抽取的效果。

标签	TF-IDF权值
晓松笔记	0.6900158498
晓松老师	0.6674154598
高晓松老师	0.3867761205
民国女神	0.2206605348
民国电影	0.2200944517
台湾历史	0.1973960157
民国名媛	0.1877979987
台湾民主	0.1639074005
中国现代舞	0.1173783794
名将竞折腰	0.1138250572
电影音乐	0.1057246794

图 5-15 TF-IDF 方法对“高晓松”的微博进行标签提取

5.2.3.2 基于 Textrank 的标签抽取方法

我们为微博文本构建以候选关键词为节点的无向图，将共现的窗口定为一条微博的长度，即倘若两个词在同一条微博中出现，我们就认为它们之间存在较强的语义联系，共现次数加 1。随后按照公式迭代计算每个候选关键词的 Textrank 权值，

权值排名前 100 的作为关键词。图 5-16 显示了微博用户“高晓松”的微博中排名前 100 的关键词及其 Textrank 权值。

候选词	Textrank权值
晓松	0.0131884082168
高晓松	0.00851716851832
中国	0.00710756498845
台湾	0.00683193864773
美国	0.00569820975144
笙箫	0.00558544409298
电影	0.00523769422965
历史	0.00461817935499

图 5-16 Textrank 方法对“高晓松”的微博进行关键词提取

随后对关键词进行组合，扩展成名词性词组作为标签，方法与基于 TF-IDF 的方法相同。图 5-17 显示了抽取标签的效果。

标签	TextRank权值
晓松老师	0.0158973229
晓松笔记	0.0146232537
台湾历史	0.011450118
高晓松老师	0.0112260832
中国公司	0.0108405231
民国电影	0.0098309903
中国土豪	0.0085916479
中国现代舞	0.008402539
台湾民主	0.0081000915
电影音乐	0.0076509631
民国女神	0.0074965493

图 5-17 Textrank 方法对“高晓松”的微博进行标签提取

5.2.3.2 基于聚类的标签提取方法

我们首先根据上一节计算微博用户“高晓松”微博中候选名词的 Textrank 权值，选取排名前 100 的词聚成 5 个簇，然后在每个簇中提取标签。如果该簇中存在名词性词组在原文中出现过，则选取权值最高的词组作为该簇的标签；如果不存在，

则选取该簇中 Textrank 权值最高的关键词作为该簇的标签。这样每个簇中产生一个标签，图 5-18 显示了聚类的效果和标签提取的结果。

类别	候选词
group0	战俘 总统 法官 蒋介石 犹太人 夫人 妄人 文盲 女人 老师 作者
group1	台湾 日本 中国 美国 节目 德国 朝鲜 二战 战争 日军 加州 中文 文化 马戏团 饭店 传统 排行榜 国家 美利坚 中华 问题
group2	晓松 高晓松 阴影 友谊商店 马可波罗 韦小宝 复星 上海 青春 阳 大和族 满洲 安宁 琉球 未央 西语 情色 网络 回国 金马 猛
group3	历史 法律 人生 名字 信息 灾难 女性 原因
group4	人民 大伙 同志 民众 百姓

标签	TextRank权值
战俘	0.0048728385243
台湾	0.0140970489874
晓松高晓松	0.024499176239
历史	0.0071314520303
人民	0.005044512358

图 5-18 聚类方法对“高晓松”的微博进行标签提取

5.3 本章小结

本章通过爬取特定博主的微博进行情感分析和兴趣标签提取，实现一个完整的基于微博博主的情感倾向分析系统。其中情感分析对单条微博的情感进行了判断，并对整体情感进行了统计和评价。兴趣标签提取则在整体上体现了博主的兴趣点和关注点。综合以上两点，构成了较为完整的博主情感倾向分析系统。

结　论

本文针对新浪微博博主构建了一个较为完整的情感倾向分析系统。本文首先依次阐述了三项主要工作的算法原理及实现，然后综合应用，构建完整系统。

主要研究工作可概括如下：

（1）完成了新浪微博爬虫。本文的情感倾向分析基于微博博主，因此准确获取博主的微博是前提。本文构建了基于微博博主的爬虫，可以实时、准确的爬取指定用户的微博。

（2）使用两种方法对博主进行情感分析。两种方法都要使用情感词典，因此本文首先构建了微博情感词典。第一种方法基于支持向量机模型，首先抽取特征，对微博进行向量化，随后训练模型，对待分析文本进行极性分析。第二种方法基于情感词典，通过对微博中存在的否定词、程度副词、感叹号、以及微博表情符号等进行相应分析处理，从而计算整条微博的情感分值。经测试，两种方法的准确率分别达到了 75.3%和 71.4%。

（3）使用三种方法对博主进行兴趣标签提取。前两种方法分别基于 IF-IDF 权重算法和 Textrank 权重算法。其中 TF-IDF 权重算法需要额外构建语料库，Textrank 权重算法则根据词语之间的共现关系构建无向图，这两种算法都可能会产生语义堆积的现象。第三种方法基于聚类技术，本文首先提出了基于同义词词林的词语相似度的计算方法，随后根据词语相似度进行聚类，聚类后在每个簇中进行标签提取。

情感分析通过分析单条微博的情感，基于统计来分析博主整体的情感。标签提取则针对博主近期的所有微博提取标签，体现了博主的兴趣点和关注点。综合以上两点，本文对基于微博博主的情感倾向分析作了初步探索。

参考文献

1. Long Jiang,MoYu,Ming Zhou,et al.Target-dependent Twitter Sentiment Classification.in:Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics:Human Language Technologies.Oregon,USA:2011
2. Dmitry Davidiv,Oren Tsur,Ari Rappoport.Enhanced Sentiment Learning Using Twitter Hash-tags and Smileys.In Coling 2010
3. Luciano Barbosa,Junlan Feng.Robust Sentiment Detection on Twitter from Biasedand Noisy Data.In Coling 2010
4. Alec Go,Richa Bhayani,Lei Huang. Twitter Sentiment Classification using Dista Supervision.Technical report.Stanford Digital Library Technologies Project.2009
5. Matthew Michelson, Sofus A. Macskassy. Discovering Users' Topics of Interest on Twitter: a First Look[C]. AND '10 Proceedings of the fourth workshop on Analytics for noisy unstructured text data. 2010.
6. Yegin Genc, Yasuaki Sakamoto, Jeffrey V. Nickerson. Discovering Context:Classifying Tweets through a Semantic Transform based on Wikipedia [C].HCII. 2011.
7. Wei Wu, Bin Zhang, Mari Ostendorf. Automatic Generation of Personalized Annotation Tags for Twitter Users[C]. ACL. 2010.
8. 徐琳宏,林鸿飞.基于语义特征和本体的语篇情感计算.计算机研究与发展.2007.
9. 闻彬,何婷婷,罗乐等.基于语义理解的文本情感分类方法研究.计算机科学.2010.
10. 赵妍妍,秦兵等.基于句法路径的情感评价单元识别.计算机研究与发展.2011.
11. 陈晓东. 基于情感词典的中文微博情感倾向分析研究:[D].华中科技大学.2012.
12. 唐慧丰,谭松波,程学旗.基于监督学习的中文情感分类技术比较研究.中文信息学报.2007.
13. 谢丽星.基于 SVM 的中文微博情感分析的研究:[D].北京:清华大学.2011.
14. 方维.微博兴趣识别与推送系统的研究与实现[D].华中科技大学.2012.
15. 谢毓彬. 面向微博用户的标签自动生成技术研究[D].哈尔滨工业大学.2012.
16. HowNet[R/OL].HowNets Home Page.http//www.keenage.com.2011.
17. 张伟,刘缙,郭先珍.学生褒贬义词典.北京:中国大百科全书出版社.2004.
18. 史继林,朱英贵.褒义词词典.成都:四川辞书出版社.2005.
19. 杨玲,朱英贵.贬义词词典.成都:四川辞书出版社.2005.
20. 同义词词林扩展版.http://www.ir-lab.org.2011.
21. 台湾大学中文通用情感词典.http://www.datatang.com/data/43460.2012.

致　谢

毕业设计得以顺利完成，首先要感谢我的导师黄俊恒副教授的悉心指导。在导师的关怀下，我不仅较好的完成了毕业设计，更向老师学到了很多学习和工作的方法，尤其是导师在学术研究上一丝不苟的态度深深打动了我。藉此论文完成之际向老师致以由衷的感谢。

另外感谢验收组的闫健恩、张小东、王伟、徐芳老师，感谢老师四年来在我的学习、生活和工作方面对我的关心和指导，为我开启了知识的大门，提供了大量的实践机会，提高了我的实践能力和论文写作水平。

感谢论文中参考的参考文献的作者，并对论文中隐含的理论支持者和设计思想开拓者表示感谢。

特别感谢实验室老师和师兄、师姐对我论文的完成提供了许多帮助。感谢我的同学和朋友对我提供的支持和帮助！

特别感谢我的家人，在 4 年的求学生涯中，您们一直关心着我，鼓励着我，教育开导我，为我提供了最好的成长环境，给予了我最大的关爱和支持，在此要特别向我的父母说一声谢谢！

最后衷心的感谢在百忙之中评阅论文和参加答辩的各位专家、教授和老师！