华北电力大学

专业硕士学位论文

基于自然语言处理的社交网络数据挖掘研究

Research on Social Network Data Mining Based on Natural Language Processing

张培华

2017年3月

国内图书分类号：
学校代码：10079

国际图书分类号：
密级：公开

专业硕士学位论文

基于自然语言处理的社交网络数据挖掘研究

硕士研究生：张培华
导　　师：翟学明
申请学位：工程硕士
专业领域：计算机技术
所在学院：控制与计算机工程学院
答辩日期：2017年3月
授予学位单位：华北电力大学

华北电力大学硕士学位论文原创性声明

本人郑重声明：此处所提交的硕士学位论文《基于自然语言处理的社交网络数据挖掘研究》，是本人在导师指导下，在华北电力大学攻读硕士学位期间独立进行研究工作所取得的成果。据本人所知，论文中除已注明部分外不包含他人已发表或撰写过的研究成果。对本文的研究工作做出重要贡献的个人和集体，均已在文中以明确方式注明。本声明的法律结果将完全由本人承担。

作者签名：张培华　　日期：2017年3月11日

华北电力大学硕士学位论文使用授权书

《基于自然语言处理的社交网络数据挖掘研究》系本人在华北电力大学攻读硕士学位期间在导师指导下完成的硕士学位论文。本论文的研究成果归华北电力大学所有，本论文的研究内容不得以其它单位的名义发表。本人完全了解华北电力大学关于保存、使用学位论文的规定，同意学校保留并向有关部门送交论文的复印件和电子版本，允许论文被查阅和借阅。本人授权华北电力大学，可以采用影印、缩印或其他复制手段保存论文，可以公布论文的全部或部分内容。

本学位论文属于（请在以下相应方框内打“√”）：

保密□，在　　年解密后适用本授权书

不保密☑

作者签名：张培华　　日期：2017年3月13日

导师签名：[illegible]　　日期：2017年3月13日

Classified Index:
U.D.C:

Thesis for the Master Degree

Research on Social Network Data Mining Based on Natural Language Processing

Candidate： Zhangpeihua

Supervisor： zhaixueming

School： School of Control and Computer Engineering

Date of Defence： March, 2017

Degree-Conferring-Institution： North China Electric Power University

摘要

微博是一种目前非常热门的社交平台，用户以短文本或多媒体信息的方式在平台上实现实时的信息分享与交流。用户发布的文本虽短，但长时间积累下来的数据蕴含着丰富的用户的个性化特征等信息。平台的用户数据中蕴含着丰富的社会信息价值，微博用户数据挖掘对于社交网络发展与社交信息分析具有重要意义。

社交网络数据挖掘完成的主要功能就是通过分析和挖掘用户在微博中的海量短文本，得到用户的个性化特征等信息。其首要工作是从网络中采集大量微博数据，采用特定的格式进行信息存储；然后对获取的微博信息进行分词处理和信息特征表示处理，最后通过数据挖掘方法进行用户识别和用户类型分析。

本文利用网络爬虫技术设计了基于模拟登录的用户数据爬取系统，提供了从网络中获取大量用户微博数据的方法。根据用户数据结构特征，采用基于 JSON 格式的 NOSQL 数据库进行存储。

针对目前分词方法存在的新词发现困难的问题，提出了基于词典匹配与统计标注相融合的中文分词方法。本方法以字典匹配方法为基础，融入 CRF 标注算法，并在分词过程中迭代训练实现算法自学习能力。通过将匹配方法与标注方法相融合，根据汉语语义规律选取分词结果，有效改善了中文分词在分词准确性和未登录词发现等方面的分词效果。在测试语料上实验结果表明，文中提出的方法与最大正向匹配算法相比，F 值提高了 9.6%，且比 CRF 标注算法提高了 2.9%，能更好地满足实际应用需求。

目前的微博数据挖掘中主要采用 One-hot representation 特征表示方法，其缺点是不能表达上下文语义。本文采用基于 word2vec 的用户特征表示方法，在用户特征表示中加入了上下文信息并且降低了用户信息向量维度，提高了后续数据挖掘算法的计算效率。

通过对微博用户数据的分析，发现用户中存在部分垃圾用户会对数据挖掘带来噪声干扰，本文设计了基于 SVM 的垃圾用户识别模型对垃圾用户进行识别，在测试集上 F 值达到 0.94。然后根据微博用户关注内容，利用 K-means 聚类分析算法进行了用户社区划分。由于用户社区划分的不确定性，通过 DB-index 算法计算最优聚类中心数值，提高了聚类结果的类间辨识度和类内相似度。

关键词：微博；分词；SVM 分类器；K-means

Abstract

Micro-blog is a very popular social platform. Users use short text or multimedia information on the platform to achieve real-time information sharing and exchange. Users publish the text is short, but a long time to accumulate the data contains a wealth of personalized features such as user information. The user data of the platform contains rich social information value. The data mining of Micro-blog users is of great significance for the development of social network and the analysis of social information.

The main function of the Social network data mining is to get the user's personalized features and other information through the analysis and mining of micro-blog in the mass of short text. The first work is to collect a large number of micro-blog data from the network and store the information in a specific format. Then, we process the word segmentation and information feature representation. Finally, the data mining method is used to analyze the user identification and user type.

In this paper, we design a user data crawling system based on simulated login using the web crawler technology, and provide a method to get a large number of micro-blog data from the network. According to the characteristics of user data structure, this paper use JSON format and store into NOSQL database.

In order to solve the problem of finding new words in current word segmentation methods, a new Chinese word segmentation method based on lexical matching and statistical annotation is proposed. This method is based on the dictionary matching method and incorporates the CRF annotation algorithm, and iterative training in the word segmentation process to achieve self-learning ability. By combining the matching method with the labeling method, the segmentation results are selected according to the Chinese semantic rule, which effectively improves the segmentation effect of the Chinese word segmentation in terms of word segmentation accuracy and unrecorded word discovery. Experimental results on the test corpus show that the method proposed in this paper improves the F-value by 9.6% more than the matching method and 2.9% more than the CRF algorithm.

One of the main features of micro-blog data mining is the one-hot representation, and its shortcoming is that it can not express the context semantics. In this paper, the user character representation based on word2vec , and the context information is added to the user characteristic representation, and the dimension of the user information vector is reduced, which improves the efficiency of the subsequent data mining algorithm.

Through analyzing the data of micro-blog users, it is found that there are some users who will bring noise interference to the data mining. In this paper, the garbage user identification model based on SVM is designed, and the F value of the garbage user identification test set is 0.94. Then, according to the content of micro-blog users' attention, the K-means clustering algorithm is used to divide the user community. Due to

the uncertainty of user community partitioning, the optimal clustering center values are calculated by DB-index algorithm, which improves the inter-class and extra-class similarity of clustering results.

Key words: micro-blog; segmentation; SVM classifier; k-means

目录

第 1 章 绪论

1.1 选题背景和意义

据中国互联网络信息中心（CNNIC）在国家网信办新闻发布厅公布的第 38 次《中国互联网络发展状况统计报告》显示，到 2016 年 6 月，我国上网用户数量为 7.10 亿，上半年增加上网用户数量为 2132 万，增长率为 3.1%。我国互联网普及率达到 51.7%。社交网络出现后便得到了爆炸式的扩张与发展，国外有脸书、谷歌+、推特，国内有新浪微博、微信、人人网等。到 2016 年 6 月，微博注册量为 2.42 亿，在线率为 34%。微博平台的特点是促进网络用户之间的信息交流，通过用户与用户“关注”和“被关注”的关系来传播信息，用户能够无约束的表达思想与看法。用户数量爆发式增长的同时，用户发布的微博数据更是可以以海量来形容，这样的增长趋势使用户接触到的信息丰富多彩，但同时也给用户对信息的抉择带来困难，甚至由于商业需求，许多企业会制造一些网络机器人，冒充用户去社交平台上发布大量的广告信息，但是在广告投放的目的性不强时，既造成广告效率底下，也会降低社交平台的用户体验。因此通过数据挖掘技术和推荐技术优化用户信息接触面[1,2]，改善用户信息获取和信息分享的效率具有重要意义。

在互联网和社交需求不断扩张，各个社交网站蓬勃发展的过程中，大量涌入的用户和粉丝骤增的背后，也暗藏着隐忧，这就集中体现在垃圾粉丝身上。但是随着微博平台的不断扩大，垃圾用户影响微博正常用户交流的现象越来越明显，垃圾用户的行为造成网络的噪声数据，降低了正常网民的用户体验，给微博用户日常信息交流带来烦恼，甚至极大污染了微博和网络的健康生态，侵害了网民和公众的正常利益。因此针对垃圾粉丝和广告用户的识别与过滤是新浪微博数据挖掘工作中一项意义重大的研究工作。

过去几年中，我们见证了社交媒体领域的快速增长，这主要是由于世界各地的互联网用户和社交平台数量的快速增长。随着越来越多的人加入到社交网络中，人们通过社交平台可以更方便的实现相互之间的联系以及个人观点的表达和分享。伴随着用户规模的快速扩大，社交网络中的内容也变得更加丰富，面对庞大的信息量，用户不可能做到全方位的关注，因此针对微博用户的关注方向进行社区性的用户聚类分析，对于未来进一步的用户数据挖据与推荐动作具有重要意义。

1.2 国内外研究动态

微博（Micro-blog）是最新的 Web 2.0 技术之一，微博的雏形类似于 EvanWiiliams 在 2006 年开发出的 Twitter 网站。我国互联网公司新浪在 2009 年 8 月开放出新浪微博内测版，而之后微博对我国社会环境的影响，得到全球性关注。

国外对于微博的研究中，主要体现在以下几个方面：a.微博数据挖掘，Carter，Simon 主要研究了多语言环境下将微博信息进行语言统一化再进行数据分析的问题[3]，Efron M 总结了微博数据挖据中 使用到的搜索优化、情感挖掘、数据表示等技术[4]；b.交流与推荐，Stieglitz S 对情感对微博信息的交流的影响进行了分析，并提出附有情感信息的微博信息相比于普通微博信息有更快更广的传播能力[5]，Takehara，Takumi 实现了基于个性化推荐的广告推送系统,推送时的关键词提取于相应网站内的信息[6]，c.用户特点与状态，Oulasvirta A 研究了微博用户更新分享信息的频率与用户粉丝量的关系，以及粉丝的互动和反馈与微博用户坚持分享信息持久性之间的关系[7]；d.功能与应用，Lee R 研究了地域范围内的社会热点信息与微博用户的地理标签之间的关系[8]，Ebner M 提出微博平台可作为教育行业中正常课堂教育之外的教学方式以及微博平台在其中的作用力[9]，Marques A 研究了博客与微博的特点，提出前者基于创作用户而后者基于读者认为博客是基于阅读用户的观点，并提出了结合两种信息分享方法的信息交流思路[10]。

国内对微博信息进行数据挖掘研究工作起步较晚，到 2008 年才出现对微博内容、评论内容等信息的分析与研究，王晓兰总结了信息传播学与微博信息分享之间的关系[11]，闫幸对微博平台提供的信息交流方式，以及微博信息交流产生的社会价值等方面进行了研究，并总结了微博信息研究的热点内容[12]，孙晓莹通过对学术文献方面的统计，研究了微博信息在相关文献在期刊、专利和基金等方面的分布[13]。

根据以上调研情况可知，国外关于微博信息的研究起步较早而且在理论与应用方面研究较深入，国内由于微博平台出现稍晚而且微博发展趋势也不等同于国外模式，因此国内相应领域的研究内容从深度和研究广度都有待进一步得到挖掘，尤其对于微博信息数据挖据方面研究内容的直观展示与分析。

1.3 本文主要研究内容

基于上述分析，本文以新浪微博为数据来源，进行对用户数据获取与存储，垃圾用户识别，用户聚类分析三方面进行了深入分析与研究。针对以上问题本论文提出以下几种方法：

1、新浪微博开放了针对研发者的用户数据获取 API 接口，但考虑到众多开发者高频访问可能带来对平台服务的干扰对开放接口添加了许多访问限制。为突破接口访问限制，本文自行设计了基于模拟登录的多用户访问方法，实现了用户数据高频无间断获取的功能。针对微博用户数据和微博数据的结构特征，本文设计 JSON 格式的数据存储结构，以便于后续研究对数据的访问与分析，并最终将数据存储在 MongoDB 数据库中。

2、由于本文针对的是新浪微博平台数据，平台中信息以中文为主要语言，而中文数据挖掘中最基础且非常重要的就是中文分词系统。因此后续数据挖掘工作带来便利性，本文提出了一种基于统计模型和词典匹配相结合的分词算法。

3、针对用户数据中网络机器人会对数据挖掘带来干扰的问题，本文提出基于 word2vec 词嵌入方式的用户信息表示方法。

4、根据对现有机器学习方法中分类算法特点的研究，提出基于 SVM 分类模型的垃圾用户识别方法，并通过引入核函数和松弛变量进一步提高用户识别模型的分类效果。

5、根据用户个人信息以及发布的微博信息实现基于 K-means 的用户聚类，并通过 DB-index 评估算法获取最佳聚类中心数值，同时兼顾了聚类结果中类内聚合度和类间区分度。

1.4 论文结构

论文全文一共分为 6 章。

第一章介绍研究的背景与意义，回顾了国内外相关研究的现状，列出本文研究的关键内容和论文内容安排。

第二章对本文关于微博数据挖掘系统结构以及技术进行论述。

第三章设计了微博数据爬虫的系统，解决微博数据爬取过程中的关键问题，以及微博数据的存储结构设计。

第四章对社交网络数据处理进行研究，主要进行中文分词算法研究与创新，微博用户信息向量表示研究。

第五章研究了社交网络数据挖掘中基于 SVM 分类模型的垃圾用户识别，基于 K-means 用户聚类分析，并通过算法实验验证数据挖掘工作的成果。

第六章综合前述工作，总结本论文的研究内容，并对未来可能的研究方向进行展望。

第 2 章 社交网络微博数据挖掘系统结构

社交数据产生于网络用户的交流，它是指 BBS，脸书，推特，微信，微博等社交平台上的用户数据，如推特的推文，微博消息，评论，转发等内容，以及各社交平台的好友信息，它们均属于社交数据。社交数据包括个人的许多社会化资源，从这些数据中可以分析出用户个人生活特点，社群化好友关系，为精准业务推荐，兴趣推荐等提供了依据。本章对现有微博数据挖掘系统相关研究进行介绍。

2.1 数据挖掘系统结构

数据挖掘系统主要由数据获取与存储模块、数据处理模块、数据挖掘与分析模块三大部分组成。数据挖掘工作的执行流程如图 2-1 所示。

图 2-1 数据挖掘系统工作流程图

数据的获取方式按照数据源有所区别，主要有直接从数据仓库中提取、从大量存储文档中解析、通过传感器收集、从网络数据中解析等方法。由于数据挖掘面向的是大量数据的处理，因此设计适合数据挖掘工作的数据存储结构和选择合适的数据存储工具对后续数据处理工作很有意义。

数据处理工作主要是进行数据挖掘前的数据过滤、数据特征表示等工作。有效的对数据过滤和合理的数据特征表示会提升数据挖掘工作中的计算效率和数据挖掘效果的准确性。数据处理是数据挖掘的重要基础。

数据挖掘与分析工作的目的在于利用现有数据挖掘数据中隐含数据特征，基于大数据分析宏观现象与趋势。

2.2 微博数据获取与存储

微博数据属互联网数据，由于商业安全与保护的目的，微博平台不提供通过数据库访问的获取方法。新浪为数据挖掘用户提供了微博用户数据的 API 接口虽然能够弥补这个不足，但是新浪服务器对不同级别的授权用户，对允许访问接口的频率和可访问内容做了不同的限制，因此无法大量的获取完整的用户数据。本文通过设计网络爬虫进行微博数据的解析与获取。

网络爬虫（以下简称：爬虫）[14,15]即一类网络数据自动获取机器人，主要被应用于像搜索引擎或有相似互联网数据需求的平台，用来抓取或者更新平台中的网络数据以改善搜索效率等实际应用需求。爬虫能够自主的爬取其可以获取到的网页内容并将网页内容解析存储到本地，满足有数据需求的平台后续的数据需要，从而可以使平台本身在数据的全面性和数据实时性方面更好的满足用户需求。

本文主要对微博信息进行数据挖掘，因此爬虫的爬取对象是新浪微博平台。微博数据爬取过程中涉及到用户登录验证的问题，本文通过分析验证原理，本文提出了基于模拟登录的微博用户数据爬虫系统。

2.2.1 微博数据爬取策略

爬虫主要分为两大类：

广度搜索：例如一些知名搜索引擎 Google，百度，必应，雅虎等，它们都属于广度搜索爬虫，原理就是每将一个页面的所有连接都拿下来后，去遍历所有的链接，再按照上述步骤不断抓取页面直到找到相关关键词，之后对页面按相关度进行排序[16-18]。

垂直搜索：即抓取特定的数据，如京东上面的所有书籍信息（包括书名，作者，语种，链接等等），将这些特定的数据进行序列化（Json 或 XML）之后存储到数据库或文件中[19-21]。

本文中的爬虫系统，其主要抓取的是新浪微博中的半结构化特定数据，即垂直搜索方式。

2.2.2 微博数据存储方法

在信息化的时代，用户对于数据的存储有很高的要求，各种类型的数据库应运而生，当前互联网中数据库的类型主要有两种，一种是传统的 SQL 数据库，例如 Microsoft SQL Server，Oracle，MySql 等，另一种是随着当今社会数据规模越来越大产生的 NoSql 型数据库，例如 Cassandra，Hbase，MongoDB 等。

传统 SQL 数据库结构性很强，针对结构化的数据有很好的适应性，SQL 型数据库是基于表的存储方式，而 NoSql 是基于文档，键值，图形数据库等的存储方式，更加适合用来存储非结构化和半结构化数据。

微博数据大多是文本和多媒体数据，数据类型术语半结构化和非结构化数据。针对微博数据的数据类型特点，本文对其存储结构设计和存储工具选择进行了研究。

2.3 微博数据处理中的自然语言处理

微博数据是基于短文本和图片、视频等多媒体数据构成。本文主要对微博数据中的段文本数据进行数据挖掘工作。文本数据处理处要用到的是自然语言处理技术，包括分词技术、知识表示技术等。

2.3.1 中文分词

目前应用比较广泛的有两类中文分词方法，基于词典匹配的方法和基于统计标注的方法。基于词典的方法利用词典作为分词基础，把词作为文本的最小单元，通过匹配方法进行语料分词[22]。该方法思路简单，易于操作，但不能有效解决未登录词以及歧义消解问题。基于统计标注的方法是目前被采用较多的分词方法，其主要思想是把字作为文本的最小单元，由字构词[23]。该方法分词性能较高，但其对训练语料的领域性依赖较强，训练好的模型在变换到不同领域进行分词测试时，效果较差。

最大正向匹配算法与回溯匹配算法（MMSEG）是较典型的词典匹配方法。最大正向匹配算法的算法思路简单，易于实现，且分词词典容易获取。MMSEG 是基于最大正向匹配算法的改进，对相邻词语间关系加入规则，提高了一定程度的分词效果。但是其规则通过人为规定和总结，不能充分考虑分词语料中的词间关系。MMSEG 同最大正向匹配算法一样存在不能有效解决未登录词发现与歧义消解等方面的问题。

CRF（条件随机场）模型是一种统计标注算法。目前普遍采用经人工划分好的分词语料进行模型训练，通过模型对语料进行标注并解码完成分词。但在建立不同领域语料时需要花费大量的人力物力，且分词训练语料的构建存在人为的主观性。

（1）基于字典匹配的方法：较早被采用的字典匹配算法是最大匹配算法。其分词过程依据两项基本规则：分词颗粒度最大化也即分词总体词数最小化。分词颗粒度最大化是指根据对文本语义的分析，使分词结果中分出词语尽可能获得最多的字即最大词长，目标就是让每个被分出来的词语可以更确切的语义，如：“公安局长”可以分为“公安—局长”、“公安局—长”、“公安局长”都认为是正确的 ，但是要用语义分析，则“公安局长”的分词结果最好。由于最大匹配算法在歧义消解方面效果较差，本文在原最大匹配算法的基础上加入了歧义消解规则，并命名为复杂最大匹配算法。

复杂最大匹配算法核心思想是在对句子中某个词进行切分时，需要向后展望两个汉语词，并且找出所有可能的“连续三词语块”。复杂最大匹配算法匹配过程以字为扩展单位，设定最大匹配词长，然后前面词每次向后面词“借取”一个字形成新的匹配词，直到找到最长匹配词且“连续三词语块”不能再扩展。此匹配方法更能体现相邻词之间的关联性，例如对“北京大学生前来应聘”分词匹配过程如图 2-2 中箭头所示。

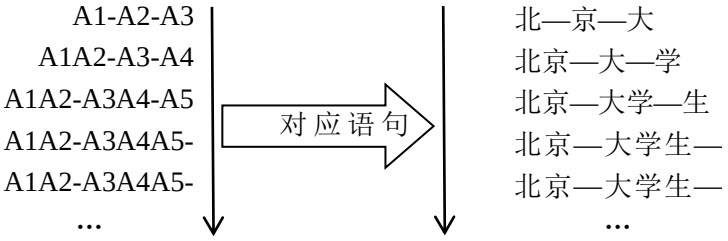

图 2-2 “北京大学生前来应聘”分词匹配过程

歧义消解规则主要用于解决分词过程中出现的语义歧义问题。本文根据语义规则，规定了 3 类歧义消解规则，分别是最大平均词长、最小词长方差、最大单字词语语素自由度之和。

（2）基于统计标注方法：条件随机场(CRF)由 Lafferty 等人[24]在 2001 年提出，其融合了隐马尔可夫模型链接上下文语义和最大熵模型状态生成时全局考虑的特征，属于无向图模型的一种。CRF 的联合概率是无向图中若干最大团的势函数相乘的方式求得的，因此其是一种典型的判别式算法模型，而其应用最广泛的就是基于线性链的条件随机场。薛念文等人[25]在 2003 年提出将中文分词问题通过以字为单位并把字与字之间关系通过序列标注进行表示来解决。

CRF 是一个由无向图表示的联合概率分布统计学习方法，目前被广泛应用在不同领域的预测问题和标注问题方面。本文将其应用在文本序列标注问题上，假设文本序列为 $X=(X_1,X_2,X_3,...,X_n)$，标注序列为 $Y=(Y_1,Y_2,Y_3,...,Y_n)$。在给定文本序列 X 的前提下，标注序列 Y 的条件概率分布 P（Y|X）构成条件随机场，即满足马尔可夫性。算法表达式如式（2-1）所示：

$$P(Y_i \mid X,Y_1,Y_2,...,Y_{i-1},Y_{i-2},...,Y_n)=P(Y_i \mid X,Y_{i-1},Y_{i+1}) \quad \text{i=1,2,3,...,n} \tag{2-1}$$

式（2-1）中当 i=1 和 n 时只考虑单边情况。

在随机变量 X 取值为 x 的条件下，随机变量 Y 取值为 y 的条件概率具有如式（2-2,2-3）：

$$P(y \mid x)=\frac{\exp\left(\sum_{i,k}\lambda_k t_k(y_{i-1},y_i,x,i)+\sum_{i,l}\mu_l s_l(y_i,x,i)\right)}{Z(X)} \tag{2-2}$$

其中，

$$Z(X)=\sum_{y}\exp\left(\sum_{i,k}\lambda_k t_k(y_{i-1},y_i,x,i)+\sum_{i,l}\mu_l s_l(y_i,x,i)\right) \tag{2-3}$$

式（2-2,2-3）中，t_k 和 s_l 是特征函数，λ_k 和 μ_l 是对应的权值。$Z(X)$ 是规范化因子，其对所有的输出标注序列进行求和。在模型中 t_k 代表着相邻结点之间的特征函数，即特征转移方程，其导出值依赖于当前位置状态以及与之相邻的和前一位置的状态，s_l 代表着当前结点上的特征函数，即状态生成方程，其导出值仅依赖于当前位置的输入值和状态值。

以上分词方法有各自的优缺点，基于匹配的分词方法仅需要完备的专业词典就能够实现准确率较高的分词效果，但其新词发现的能力较差；基于统计标注的分词方法，在新词发现和分词准确性等方面表现良好，但是需要大量的人工标注语料作为训练保障，而且有时标注语料会混杂入人的主观因素，导致分词误差。由于微博文本内容多是关于社会热点事件，需要较好的新词发现能力，但是微博的实时更新性和微博数据的巨量增长，对微博数据进行人工标注是一件那已完成的任务。本文对各分词算法的特点进行研究，提出了基于词典匹配与统计标注相结合的分词算法，以更好的完成微博数据的分词任务。

2.3.2 特征表示

为便于文本数据挖掘，常用的解决办法是将此进行词频统计，构造基于词表的词向量，根据向量中词频数据体现文章特征。自然语言理解的问题要转化为机器学习的问题，第一步肯定是要找一种方法把这些符号数学化。NLP 中最直观，也是到目前为止最常用的词表示方法是 One-hot Representation，这种方法把每个词表示为一个很长的向量[26,27]。这个向量的维度是词表大小，其中绝大多数元素为 0，只有一个维度的值为 1，这个维度就代表了当前的词。此种方法由于是根据词表长度构造相应长度的词向量，对于后续的算法分析工作带来维度计算困难的问题。这种词表示有两个缺点：（1）容易受维数灾难的困扰；（2）不能很好地刻画词与词之间的相似性（即，词汇鸿沟）。

由于深度学习模型的兴起，自然语言处理在深度学习方面的研究也有了很多进展。目前比较热门的一项基于神经网络模型的文本表示方法是基于 Distributed representation 的词向量数据表达方式[28-30]。由 Hinton 等[31]于 1986 年提出的基于 Distributed representation 的数据表示方法，其基本思想是通过神经网络模型使每个词都被训练出一个代表自身 K 维实数向量（K 一般为模型中人工指定的超参数），通过词之间的距离（比如 cosine 相似度、欧氏距离等）来判断它们之间的语义相似

度.其采用一个三层的神经网络，输入层-隐层-输出层。此三层神经网络本身是对语言模型进行建模，但也同时获得一种单词在向量空间上的表示。目前应用比较热门的 Distributed representation 模型是 Word2vec。与潜在语义分析（Latent Semantic Index, LSI）[32]、潜在狄立克雷分配（Latent Dirichlet Allocation，LDA）[33]的经典过程相比，Word2vec 考虑到词的上下文语义信息，能够表达的词语含义更加地丰富。

本文经过对两种文本表示方法特点的考量，将 Word2vec 应用于微博文本表示方法中。

2.4 微博用户数据挖掘与分析

微博用户数据量庞大，因此数据挖掘工作主要靠机器学习算法自动学习数据中隐含的规律。近年来微博数据挖掘受到学术界和工业界的重视，代表性工作有话题事件分析、情感分析、信息检索与推荐等。随着商业化信息涌入社交平台，越来越多的广告信息干扰着微博用户正常的社交交流，根据对微博数据的分析，本文对微博垃圾用户识别进行了研究。由于微博用户与日俱增，且用户大多都有自己的特点，而用户关注内容更能充分体现用户在社交网络中所体现的角色，因此本文根据用户关注内容对微博用户进行了社区性划分的研究工作。

2.4.1 微博用户识别

微博用户识别可以归结为数据挖掘工作中的分类问题。本文面对的是微博用户中垃圾用户识别问题，其主要解决的是正常用户与垃圾用户的二分类问题。

分类是数据挖掘、机器学习和模式识别中一个重要的研究领域。常用的分类算法有贝叶斯（Bayes）分类算法[34]、决策树[35]、支持向量机分类算法等。而且在一些分类任务中如果追求预测的准确程度，一般用支持向量机，如果要求模型可以解释，一般用决策树或者贝叶斯。

本文的分类目标是得到高准率的垃圾用户识别模型，因此选择支持向量机作为分类算法进行研究。支持向量机（support vector machine 以下简称：SVM）由 vapnik 等人[36]于 90 年代中期提出的一种基于统计学方法的机器学习算法模型，SVM 的特点是即使训练样本较少也能够通过使结构化风险最小来获得尽可能大的模型泛化能力，从而使算法模型在测试集上有更好的性能指标。SVM 是基于分类超平面的二分类机器学习算法，算法的实现思想是样本在数据表示空间中离超平面的距离尽可能大从而使分类结果达到最大置信度。

2.4.2 微博用户划分

微博用户社区性划分属于无监督分类方法，即在没有标注语料，甚至对用户种类数目也未知的情况下，对用户进行划分归类，因此也成为聚类算法。聚类是指人们事先对分类过程不施加任何的先验知识，而仅凭数据本身进行分类，其结果只是对不同类别达到了区分，但并不能确定类别的属性。常用的聚类算法有基于层次的方法[37]、基于划分的方法[38]、基于密度的方法[39]等。

基于层次的聚类方法将数据对象组成一棵聚类树。根据层次分解是以自底向上还是自顶向下方式，分词聚类方法可以进一步分为凝聚和分裂。该算法计算简单，但是一种纯粹的层次聚类方法的质量受限于：一旦合并或分裂执行，就不能修正。也就是说，如果某个合并或者分裂决策在后来证明是不好的选择，该方法无法退回并修正。

基于密度的方法是为了发现任意形状的聚类簇。该方法将簇看作是数据空间中被低密苏区域分隔开的稠密对象区域，依据基于密度的连通性分析增长聚类。该算法计算复杂，在执行过程中，需要人为设定簇密度值，簇间拓展策略等。其聚类结果受经验参数影响较大。

基于划分的方法是给定 n 个对象的数据集 D，以及要生成的簇的数据 k，划分算法将对象组织为 k 个划分 $k \le n$，每个划分代表一个簇。这些簇的形成旨在优化一个目标划分准则，如基于距离的相似度函数，是的根据数据集的属性，在同一个簇中的对象是“相似的”，而在不同簇中的对象是“相异的”。该算法特点是需要人为参与的参数设定只有 k 值，数据特征的利用率高。缺点是在没有经验的时候需要多次尝试不同的 k 值，已达到最优的聚类效果。

本文聚类分析的数据特征是用户的关注领域，不存在用户分布有层级性和分布不规则性，因此出于充分利用数据特征的目的，选择了基于划分聚类方法中 K-means 聚类方法[40,41]对微博用户划分进行了研究。

2.5 本章小结

本章主要对论文中微博用户数据挖掘系统相关的技术与算法进行了综合描述。根据数据挖掘需求和算法特点，对数据源的获取与存储，用户数据处理工作，以及用户数据挖掘与分析的相关工作进行了介绍。

第 3 章 社交网络微博数据获取

本文利用的是社交网络中最热门的新浪微博平台的数据。由于是网络数据，没有相应的用户数据库公共接口可供使用。新浪为数据挖掘用户提供了微博用户数据的 API 接口虽然能够弥补这个不足，但是新浪服务器对不同级别的授权用户，对允许访问接口的频率和可访问内容做了不同的限制，因此无法大量的获取完整的用户数据。基于以上原因，本文自行设计了可无限制获取微博用户数据的爬虫系统。根据微博用户数据结构特点，本文设计出合理的数据存储结构，通过 NoSql 型数据库进行存储。

3.1 微博数据爬虫系统设计

本社交网络数据采集系统共分为爬虫模块和数据存储模块两大部分，爬虫模块通过 web 技术和正则匹配技术实现网络页面获取与解析，数据存储模块通过组织合理的数据结构存储到数据库中。社交数据采集系统框架如图 3-1 所示。

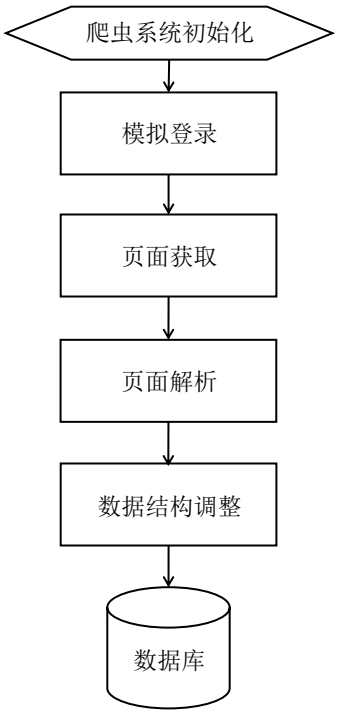

图 3-1 社交数据采集系统框架图

3.2 微博爬虫运行过程

本文微博爬虫执行过程如图 3-2 所示。

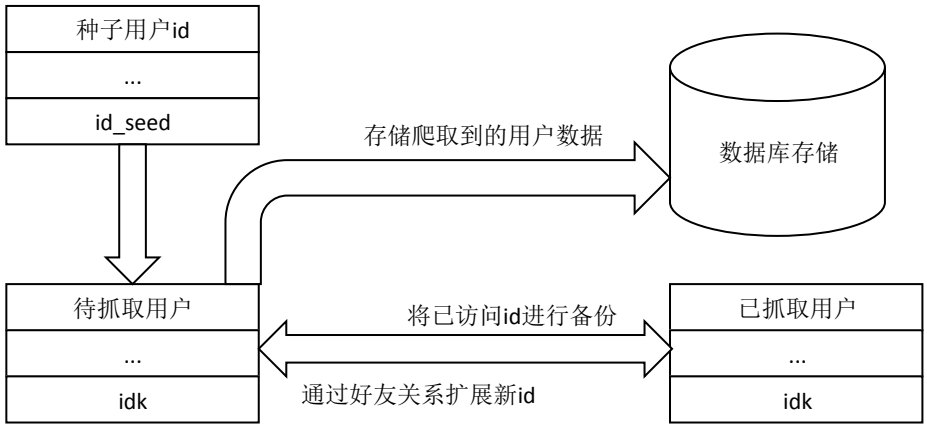

图 3-2 新浪微博爬虫执行过程图

图 3-2 以种子节点为入口进行爬取，种子节点为用户 id 范围中随机产生的 50 个用户 id，将这些 id 放入待爬取队列中，之后运行爬虫系统。

新浪微博爬虫需要解决两个关键问题：

1.用户重复访问：需要避免对同一用户重复访问。

2.新浪微博反爬虫系统的阻碍：新浪微博的反爬虫系统对机器爬虫进行限制，限制主要有禁止同一用户连续快速的访问页面、禁止同一 ip 登录微博过多次数，禁止同一浏览器登录过多用户。

对于问题 1，本文采用了两种数据结构：队列和集合，待爬取用户 id 使用队列存储，已访问用户使用集合存储，系统运行过程中，从队列中进行出栈操作进行 id 获取，之后判断该 id 是否在已抓取用户集合中，当对该用户的访问结束后将该 id 放入集合中，由于集合中元素不重复的特性，保证了集合中元素唯一，也保证了 id 的不重复访问。

对于问题 2，本文通过创建用户池，将所有用户的 cookie 进行了记录，保存，在每次访问用户数据时，从 cookie 池中随机取出一个用于数据访问，在进行访问页面的过程中，如果一个 cookie 访问了多次页面，则再从用户 cookie 池中随机取出一个用于数据访问，以此来应对反爬虫策略。

3.3 用户模拟登录

由于新浪微博是闭源社区，对于闭源社区来说，用户需要登录后才能进行页面的访问，作为自动化爬虫系统，需要首先处理自动化登录的过程，所以本文需要进行登录方式的设计。

在电脑端登录新浪微博，客户端向服务端传递的并非原始的用户名密码信息，而是经过加密处理过的数据。由于新浪微博的安全性策略，其会不定期更新相关请求信息，加密算法等。传统办法是破解加密算法进行本地加密，然后发送加密内容实现模拟登录。

本文设计出基于浏览器测试工具的模拟登陆方法。浏览器测试工具可以通过完全模拟真实用户进行浏览器的开启，密码输入，按钮点击等相关操作，通过自动化测试工具的使用，本文可以完全模拟用户输入新浪微博登录地址，输入账号密码，点击登录按钮的相关操作，并且该测试工具拥有读取 cookie 的功能，可以方便的获取 cookie。

对于模拟浏览器登录方式，它保证了 cookie 获取的便捷性，即无需手动通过查找，复制，又避免了模拟登录加密方式，关键参数解析的复杂过程，而且此过程可以应用于几乎所有的闭源社区，无需针对每个不同的社区重复进行解析加密方式等

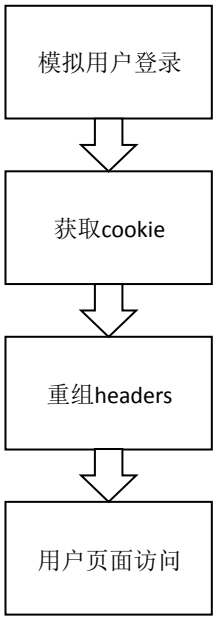

图 3-3 模拟登录流程图

3.4 微博页面解析

由于目前 Web 前端技术越来越成熟，许多网页不再是简单的静态网页，而是动态网页，这些动态网页中许多信息都是通过 Ajax 请求从服务端动态获取的，因此想要抓取那些源代码中不包含的信息，就必须通过一些技术手段得到它们。

新浪微博网页版分为两种，一种是 PC 端的新浪微博页面，一种是手机 web 端新浪微博页面，PC 端新浪微博页面是包含有各种 AJAX+JavaScript 等异步加载的动态页面，而手机 web 端新浪微博界面为无渲染的界面。

由于 PC 端的微博网页使用了很多 ajax 和 javascript 技术进行页面处理，不利于页面爬取，所以在爬虫系统设计中，选择更加简明的手机 web 端站点更利于数据提取，手机 web 端微博页面内容源码仅由简单的 html 标签构成，易于使用正则表达式进行匹配获取。

Web 端新浪微博页面与 PC 端所有内容完全一致，仅仅去除了 ajax 和 javascript 效果。每个用户的微博地址均具有固定格式，均为"http://weibo.cn/" + user_id + "/profile?page="，”page=”后跟页数，通过对隐藏标签的匹配可以获得微博的总页数，以此为标准就能定位到该用户的所有微博页面。

微博页面的分析采用逐层深入分析的方式，解析流程如图 3-4。

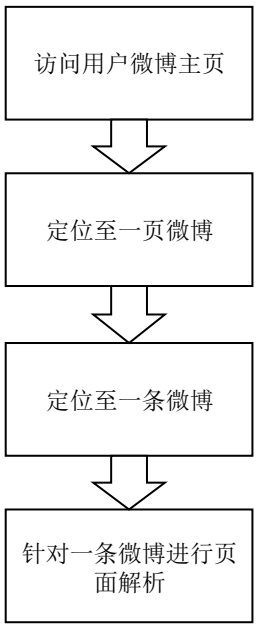

图 3-4 微博分析流程图

（1）微博信息解析：对于每条微博，其内容不仅包括文本信息，发布时间，设备信息，转发和评论的 url 地址内容等，某些微博还包含有以##标识的主题，以@为标识的提及，以及链接内容，在匹配过程中，将这些内容进行分别匹配提取，用于匹配的正则表达式如下：

设备时间信息：""(?<=).*?(?=)""

赞的信息：""<a href="http://weibo.cn/attitude.*?赞\[[0-9]+(?=\])""

转发 url：""<a href="http://weibo.cn/repost.*?转发\[[0-9]+(?=\])""

评论 url：""<a href="http://weibo.cn/comment.*?评论\[[0-9]+(?=\])""

其中，在新浪微博系统中的微博发布时间不仅包括年-月-日 时-分-秒这样准确的时间信息，还包括”刚刚“，”几分钟前“，”几小时前“，”今天“，”昨天“这样模糊的时间概念，对它们的处理方式为以当前系统时间为参考，在此基础上减去这些模糊时间所代表的时间节点，从而使时间的表示标准化。

（2）评论和转发页提取：评论和转发页指用户对某一条微博的评论转发信息，它们属于同一层级的页面，其均包括评论和转发文本信息，主题内容，提及内容，评论或转发时间信息，设备等相关信息。

（3）个人信息页的提取：个人信息页是指个人的资料背景页面。

个人信息页是标签化信息最多的页面，该页面通过"http://weibo.cn/"+userid+"/info"页面可以直接访问，其内容最多由三个 div 标签组成，其分别为用户基本信息，学习经历，工作经历三部分。

用户基本信息中有关个人情况的信息最多，为数据抓取者提供了大量可供分析研究的素材，其数据格式由 key-value 类似结构组成，在数据抓取中，可直接使用关键词进行匹配，匹配使用的部分正则表达式如下：

"(?<=会员等级：).*?(?=)"

"(?<=昵称:).*?(?=
)"

"(?<=认证:).*?(?=
)"

"(?<=性别:).*?(?=
)"

"(?<=地区:).*?(?=
)"

对于学习经历和工作经历来讲，这两部分内容相似，由于用户的学习经历和工作经历均有可能不止一个，所以在匹配时，需要通过以下正则表达式将学习经历和工作经历分为相互独立的两部分，匹配使用的正则表达式如下：

'学习经历 .*?
 '

'工作经历 .*?
 '

之后根据每段学习或工作经历均有相同特征，即均以“.”开头来匹配完成用户完整的工作或学习经历，根据正则表达式的匹配，可以完成对每段学习经历和工作经历的获取。

（4）关注页和粉丝页的提取：微博关注页和粉丝页指用户所关注的人和用户的粉丝页面，它们具有完全相同的页面结构，可以采用相同的模式进行数据获取。

微博粉丝和关注页面主要包括了与用户有互动关系的用户列表，其中包括基本的个人信息如用户名、id 等。用户 id 在微博平台中具有唯一性，因此本文通过用户 id 来获取与用户有互动关系的其他用户的相关信息。在爬虫系统运行过程中，本文就可以通过用户的粉丝和关注者关系，向爬虫队列里添加待爬用户 id。

3.5 数据存储工具选择

对于新浪微博来讲，大部分数据都属于半结构化数据，并且以文档为主。传统SQL 型数据库是以表的形式进行数据存储，其在扩展性方面很不方便。而随着互联网数据规模的扩大，新兴的 NoSql 型数据库在存储非结构化和半结构化数据方面性能良好，而且其在对象属性的扩展性方面具有很高的灵活度。根据微博用户数据结构的半结构化特点，本文选择 NoSql 型数据库作为数据存储的主要工具。

Nosql 自上世纪 60 年代就已经出现，但是直到最近几年才逐渐流行，例如 MongoDB，CouchDB，Redis 和 Apache Cassandra。

传统 SQL 数据库结构性很强，针对结构化的数据有很好的适应性，SQL 型数据库是基于表的存储方式，而 NoSql 是基于文档，键值，图形数据库等的存储方式。这意味着 SQL 数据库表示的数据由表组成，包含有若干列数据，例如，一个电话薄类似于一个文件，它包含一个记录列表，每一个都包含三个字段：名称，地址，电话号码。而 Nosql 型数据库并没有标准的模式定义，是键值对，文档，图形数据库的集合。

过去几十年产生的 web 应用程序由于其用户量，互联网速度，硬件条件等多方面的限制，并没有很大的数据量，并且大多数是标准的结构化数据，所以传统的 SQL 数据库完全可以满足其使用要求。而随着现在互联网应用技术以及社交网络等新型传播媒介的发展，网络数据规模越来越大，而且各式各样的网站层出不穷，面向的群体也各式各样，所以当前互联网中的数据不仅仅是单纯的结构化数据，更多非结构数据，还有半结构化数据。

非结构化数据没有被组织成一种能让其更容易访问和处理固定的格式。在现实生活中，很少有完全的非结构化数据。结构化数据与非结构化数据基本上是对立的：结构化数据被重新格式化，其元素值被组织成固定的数据格式以便元素以不同的组合方式被处理，管理和访问，从而使用户更好地使用这些信息。半结构化数据介于这两者之间，比如微博长文本等。

对于新浪微博来讲，大部分数据都属于半结构化数据，并且以文档为主，所以本课题选择 NoSql 作为数据存储的主要工具。

Mongodb 是一种最流行的基于文档的 NoSql 型数据库，它是一种基于动态模式的非关系型数据库。它是由 DoubleClick 的创始人，用 c++编写而成，目前许多大公司诸如：纽约时报，Craigslist，MTV 网络等都在使用。它具有一下一些优势：

速度快：由于所有的相关数据都在单一文件种，消除了连接操作，对于简单的查询，它拥有良好的性能。

可伸缩性：它是可水平扩展的。例如你可以通过在资源池中增加服务器的数量来减少你的工作量，而非单纯依靠一个独立的资源。

易于管理：无论对于开发者或者管理员来讲，其使用均较为简单，这给了用户碎片化数据的能力。

拥有动态模式：它给了用户无需修改数据就能改进数据模式的灵活性。

基于 MongoDB 以上几大优点，本文将其作为微博用户数据存储的数据库。

3.6 数据存储格式设计

MongoDB 的存储模式为一种类 JSON 的 BSON 存储格式，类似于 Python 语言中的字典，即（key-value）键值对的形式进行存储。爬虫系统抓取的数据主要分为两部分，微博部分和个人信息部分。本文针对微博数据特征设计出基于 JSON 格式的树状图存储结构。其中微博信息数据存储格式如图，个人信息数据存储格式如图 3-5

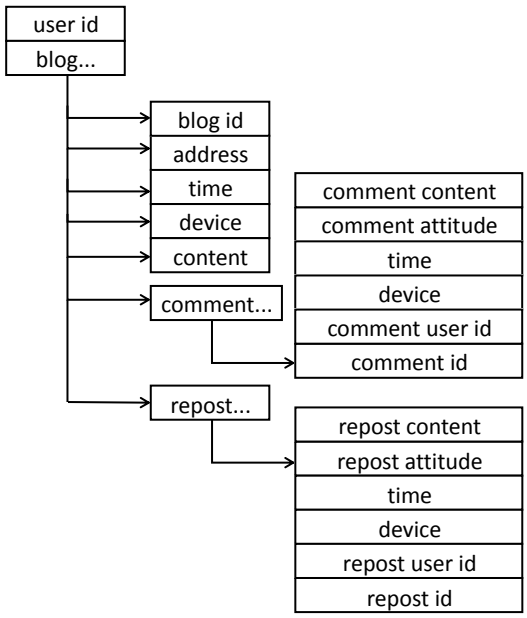

图 3-5 微博信息存储格式图

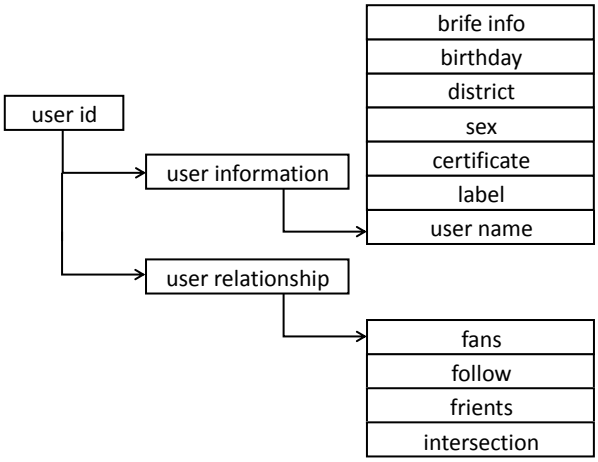

图 3-6 个人信息存储格式图

3.7 本章小结

本章对新浪微博数据获取的方法进行分析，结合微博数据特征采用垂直搜索策略设计了无限制爬取微博用户数据的爬虫系统。接下来利用模拟登录、正则解析等 web 技术进行用户数据获取，根据微博用户数据结构特征，设计相应 JSON 数据存储结构将爬取数据存储在 NoSql 数据库 MongoDB 中。本章主要进行了微博数据的爬取与存储的研究工作，为接下来的用户数据挖掘内容提供了充足的数据源保障。

第 4 章 微博数据的自然语言处理研究

数据处理是数据挖掘顺利有效进行的基础工作。原始数据存可能存在数据格式不一致、数据重复、数据含有噪声、数据表示维度高等问题。数据处理的作用是从原始数据里面选择最能代表数据的特征作为数据挖掘的属性，并尽可能表示属性的确切含义、统一多数据源的表示方式、进行数据去重、数据噪声过滤、数据特征权重计算等内容。本章主要对原始数据表示中的中文分词、数据向量表示方法、数据特征权重计算等方面进行研究。

4.1 中文分词算法设计

中文文本信息处理中最基本且重要的工作就是中文分词，因为中文不像字符类语言使用空格作为词与词之间的间隔，中文是以字为单位词与词之间没有明确的分隔符。中文信息处理的基本单位为词语，因此优秀的中文分词性能，会为后续数据挖掘工作带来便利。

针对词典匹配方法不能有效解决未登录词发现和歧义消解，统计标注模型需要大量人工生成训练语料以及模型跨领域分词效果差等问题，本文提出匹配与标注相融合的自学习分词方法。此方法以 MMSEG 为基础，通过 CRF（条件随机场）模型实现未登录词发现，并把未登录词加入训练语料进行迭代训练，不断增强统计标注算法的分词性能。对于提出方法中 CRF 模型首次训练语料不存在、迭代训练频率高时算法工作效率低以及融合过程中分词结果优化选择等问题。本文进一步对训练语料生成、训练迭代阀值设定进行研究，并设计出针对分词结果融合过程的优化方法。

词典匹配与基于统计方法的分词算法在分词领域占据这重要地位。本文中匹配与标注相融合算法核心思想是以词典匹配为分词基础，通过准确分词结果对 CRF 标注模型进行训练，然后通过 CRF 模型进行未登录词发现，增强分词器准确率和新词发现两方面的性能。

算法系统框架如图 4-1 所示。

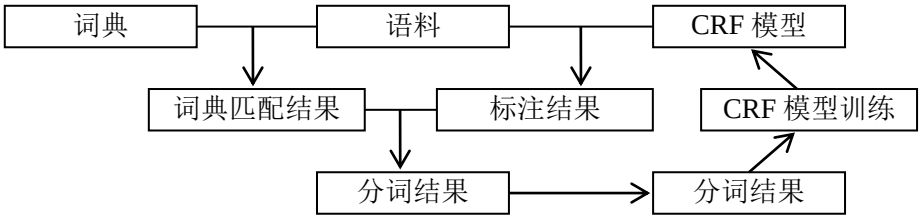

图 4-1 分词算法系统框架

4.2 分词模型训练方法

4.2.1 匹配词典选择

为使不同领域语料都得到较好分词效果，本文算法选择在通用词典基础上加入指定的领域词典。利用组合成的词典和 MMSEG 实现本文算法中字典匹配的分词过程。

采用这种策略的原因是一方面加入领域词典可使词典匹配算法更好的适应特定领域分词任务；另一方面由于 CRF 模型训练语料的获取比较麻烦，需要大量人工参与，而领域词典的获取相对容易很多。因此通过加入领域词典的方法，既能够增强算法分词性能，而且可以解决 CRF 模型训练需要大量训练语料的难题。

图 4-1 中的词典包括两个组成部分，如图 4-2 所示。

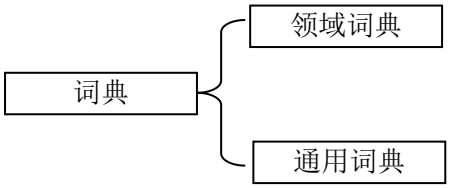

图 4-2 词典组成结构

4.2.2 训练语料生成

在以往应用中训练语料的获取大多是通过人工划分，然后通过划分后的分词结果进行模型训练，此训练语料生成办法在人工划分阶段需要消耗大量人力物力，且分词训练语料的构建存在人为的主观性。本文在训练语料的生成方法上，提出一种新的思路，即训练语料从最终分词结果中获取。但由于分词结果并不是完全准确，在提取训练语料时，需要对分词结果进行过滤处理。

分词结果中未登录词（OOV，Out of vocabulary）是指未在训练语料中出现但是被准确分割出来的词，在此我们定义未登录词的两种情况，真未登录词（ROOV，Real out of vocabulary）指未登录词中不是词典包含词，伪未登录词（FOOV，Fate out of vocabulary）指未登录词中的词典包含词。由于对 ROOV（真未登录词，下同）无法确认是否是正确词语分割结果，因此为能达到更准确的 CRF 标注结果，训练语料仅选取分词结果中包含的 FOOV。

训练语料组成结构如图 4-3 所示：

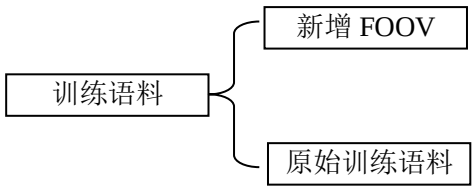

图 4-3 训练语料组成结构

4.2.3 自学习方法

自学习方法（SLSEG,Self-Learning Segment）是本文算法中的重要组成部分。其特点是使由本文算法实现的分词器在分词过程中，不断积累分词经验，从而随着分词任务量的增加，分词器的分词性能逐渐提高。SLSEG 通过分词结果迭代训练 CRF 模型达到自学习效果。

在 CRF 训练过程中，本文算法省略人工划分训练语料，选取分词结果中的 FOOV 部分作为 CRF 训练语料。分词过程中通过不断把每次分词结果中 FOOV 收入到训练语料中，当训练语料增长到预设的量级，开始 CRF 模型迭代训练，以此实现分词系统自学习功能。

CRF 模型训练过程复杂度较高，因此应尽量减少训练次数。由邢富坤调查的中文分词未登录词分布规律可知，在分词阶段随分词语料按倍数增长情况下未登录词增长比例接近 10%左右。在此定义训练语料总词数为 AOW（All of words），新增词数为 GOW（Growth of words），触发 CRF 模型训练的增长阀值为 GR（Growth rate），在算法应用过程中设定 GR=10%。定义当训练语料增长比超过设定的训练语料增长阀值时，即（GOW/AOW）> GR ，重新开始 CRF 训练，获取新训练语料下的 CRF 标注模型。

4.2.4 分词结果融合方法

分词模块实现如图 3-6 所示，主要分为回溯法匹配分词过程、CRF 训练和分词过程、分词结果融合过程。其中 MMSEG 分词过程采用第二节所述算法实现，CRF 分词过程采用训练好的 CRF 模型对语料基于当前字所在词中的位置进行（B/M/E/S）标注，然后根据标注结果进行解码形成 CRF 分词结果。

分词结果融合过程，首先定义分词结果组成，分词结果由 MMSEG 分词结果和 CRF 分词结果组成。融合方法把 MMSEG 分词结果和 CRF 分词结果进行对比，若两种结果相同则判定其为正确分词结果，若不相同则继续向后检索直到再次找到相同的词。

在对比过程中遇到结果不相同时，为保持两种分词结果对比过程的同步性，需要进行特殊处理。分别记录 MMSEG 分词结果的当前匹配长度索引为 LOMMS（Length Of MMSEG）,CRF 模型分词结果的当前长度索引为 LOCRF（Length Of CRF），然后优先扩展 LOMMS 与 LOCRF 中的较小的分词结果。以此来保证两种分词结果对比过程中扩展长度保持一致。

分词结果对比过程，如图 4-4 所示。

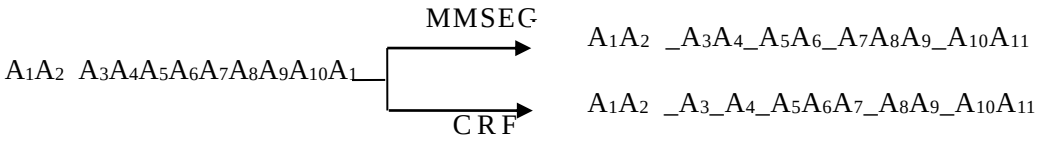

图 4-4 分词结果对比

图 4-4 中 A1 到 An 是语料中字标记，以’_’表示词分割。其中 A3 到 A9 阶段对应有两种分词结果。当融合过程中遇到这种问题时，依据中文分词的语义特点，本文采用了三条语法规则来决定优先选择的分词结果。三条规则分别是，词组的最大平均词长（LAOWL，Largest average of word length），最小词长方差（SVOWL，Smallest variance of word lengths），最大词频之和（MSOWF，Max sum of word frequency）。

（1）最大平均词长

LAOWL 规则解决的问题是，通过 CMM 获取到多个分词结果的歧义问题。根据分词语义中“词数较少分词效果较优”的特点，选取多个结果中平均词长最大的作为优先选择分词结果。LAOWL 计算方法如式（4-1）：

$$AL = \frac{len(R_1) + len(R_2) + len(R_3)}{3} \tag{4-1}$$

式（4-1）中 AL 表示“连续三词语块”平均词长，$len()$ 求词长函数，R1，R2，R3 表示“连续三词语块”中三个匹配词结果。

（2） 最小词长方差

SVOWL 规则的思想是分词结果中词长变化幅度越小分词效果越好。它针对 LAOWL 未解决的分词歧义问题，计算 AL 相等的几种结果的词长方差，然后选取词长方差最小者作为优先选择分词结果。SVOWL 计算方法如式（4-2）：

$$VL = \frac{(len(R_1) - AL)^2 + (len(R_2) - AL)^2 + (len(R_3) - AL)^2}{3} \tag{4-2}$$

式（4-2）中 VL 表示“连续三词语块”词长方差，其他符号含义同式（4-1）。

（3）最大词频之和

针对上述方法尚未解决的分词歧义问题，可以通过 MSOWF 规则获得分词结果，其思想是利用统计学观点，把“连续三词语块”词组中词出现次数总和最大的当作优先选择分词结果，这一规则假设词出现次数越大越能体现文本语义。MSOWF 计算方法如式（4-3）：

$$CL = TF_1 + TF_2 + TF_3 \tag{4-3}$$

式（4-3）中 TF1，TF2，TF3 分别代表，CRF 训练语料中词 R1，R2，R3 的词频数，CL 代表词频总和。

当遇到 MMSEG 与 CRF 分词结果不一致时，根据 LAOWL，SVOWL，MSOWF 规则做出决策，并最终完成分词结果融合过程。

通过上述对分词系统的改进，基础词典采用领域词典加通用词典，CRF 标注模型采用按新增词比例进行迭代训练，可增强分词系统对不同领域的适应能力。当把系统应用到新领域时，只需更换领域词典，经过初次迭代后就能改善分词效果，而且随着 CRF 标注模型的迭代训练能够实现分词系统自学习功能，体现此算法不断增强分词效果的能力。

4.3 中文分词算法实验

本文算法实验验证选用了 SIGHAN CWS BACKOFF 2005 作为实验语料库，它是由中央研究院（AS，Academia Sinica）、香港城市大学（CU，CityU）、北京大学（PKU，Peking University）、微软研究院（MSR，Microsoft Research）机构提供的四大类中文分词语料，其中 AS、CU 语料为繁体中文版，PKU、MSR 语料为简体中文版。通用词典（G_V，General vocabulary）采用搜狗互联网词库，领域词典分别采用 SIGHAN CWS BACKOFF 2005 提供的与四类语料对应的词典。

4.3.1 验证算法的分词效果

为体现本文算法的自学习效果，实验中将 SIGHAN CWS BACKOFF 2005 语料库中四类语料的待分词均分成 5 份，然后利用本文提出的算法对每类语料中的 5 个部分进行迭代分词操作。算法中 GR 为 10%，当 CRF 训练语料中新增词数达到原训练语料总词数的 10%时，设定 CRF 标注模型重新进行训练操作。其中，PKU 语料迭代分词效果如表 4-1 所示。

表 4-1 匹配与标注相融合自学习分词算法在北京大学语料分词效果

分词效果	P	R	F	训练预料新增词数	增长百分比%
PKU_C_1	0.880	0.863	0.871	16703	∞
PKU_C_2	0.899	0.854	0.876	19902	19.15
PKU_C_3	0.901	0.865	0.883	24422	22.71
PKU_C_4	0.924	0.912	0.918	26613	8.97
PKU_C_5	0.950	0.931	0.941	29175	19.46
SMM 算法结果	0.836	0.804	0.820		
CRF 标注算法结果	0.903	0.918	0.910		

从表 4-1 对比结果可总结出，在面对相同分词语料时本文提出的算法自学习能力较强。与单独的 SMM 算法相比较，P，R，F 分别提高了 11.4，12.7，12.1 个百分点，与单独的 CRF 标注算法相比较，P，R，F 分别提高了 4.7，1.3，3.1 个百分点，体现出该算法高效的分词性能。

4.3.2 验证算法的普遍适应性

通过分词算法分别对其他三类语料进行分词操作，并记录各个阶段的分词结果。四类语料 5 个阶段本文 SLSEG 算法与 MMSEG 算法的 P，R，F 结果对比关系如图 4-5,4-6,4-7。

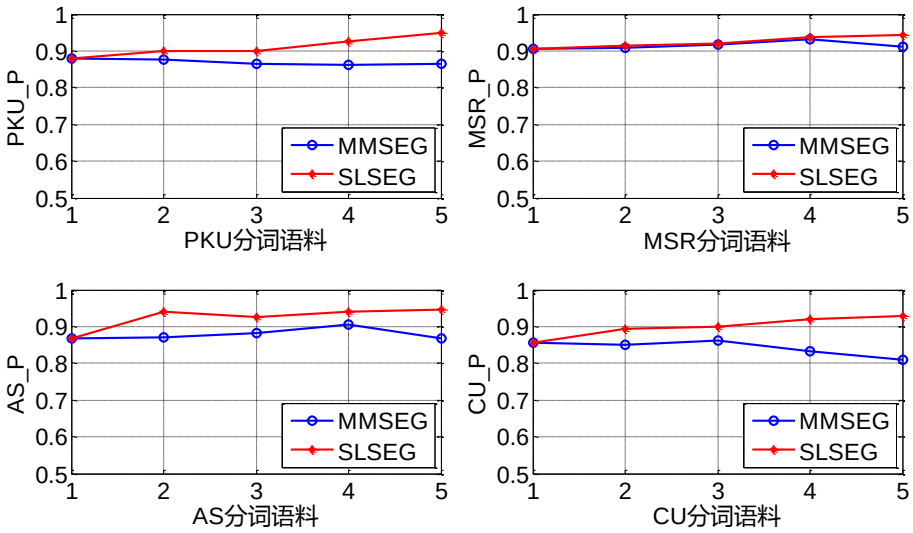

图 4-5　四类语料 5 个阶段的 P 结果关系

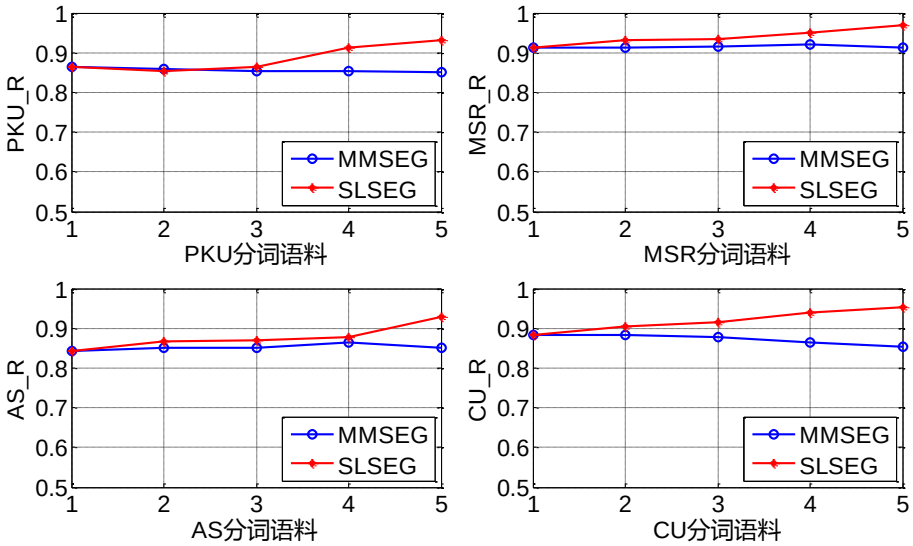

图 4-6　四类语料 5 个阶段的 R 结果关系

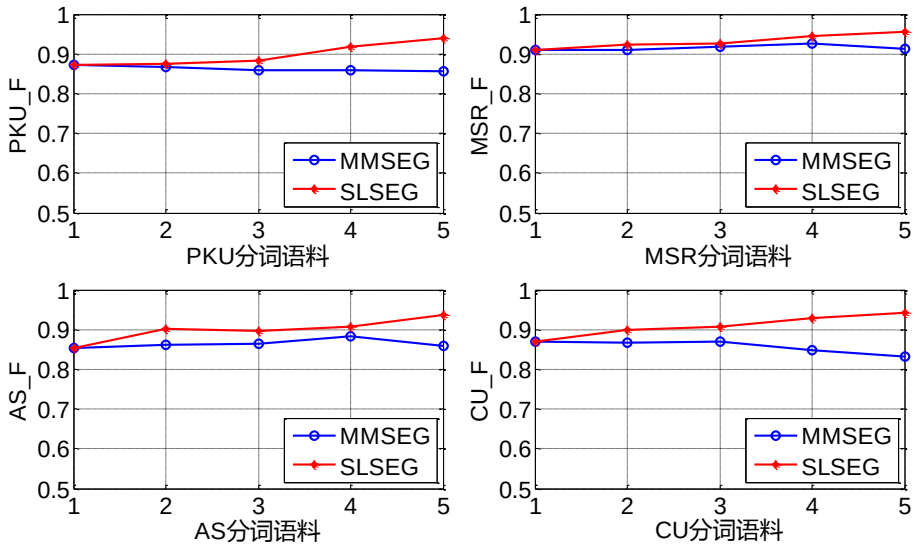

图 4-7　四类语料 5 个阶段的 F 结果关系

通过图 4-5,4-6,4-7 中 P，R，F 各阶段的变化关系可总结出，SLSEG 算法分词结果的评估值与 MMSEG 相比，随着分词实验的进行，逐渐升高，并超过 MMSEG，体现出 SLSEG 的自学习性。且本文算法对四类语料具有相似的分词过程和分词结果，表明其具有普遍适用性。

四类语料最终分词结果对如图 4-8 所示：

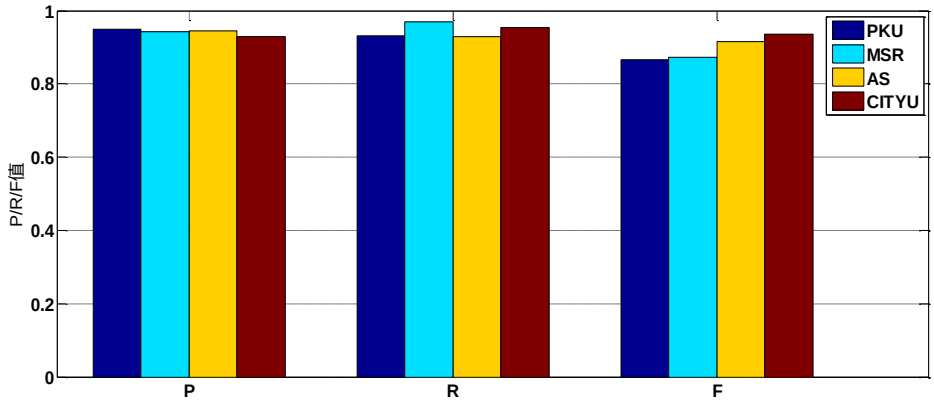

图 4-8 四类语料最终分词结果对比

通过图 4-8 的柱状图可知，本文提出的算法在四类语料分词中都取得了很好的分词效果，体现该算法面对不同领域语料时一样具有高效的分词性能。

本节实验中匹配与标注相融合自学习分词算法所得分词结果表明了该算法相对于普通分词算法在分词性能上的优点，并体现出该算法针对不同领域语料分词时具有普遍适应性。

4.4 用户特征表示

本文采用 Google 开放出的 word2vec 项目，此工具可以将词语表示成固定长度的向量。通过根据转换过后的词向量构造出对于文本内容的向量表示方式，然后通过文本向量表示方式进行文本之间的相似度计算等工作，进而可以将计算结果用于文档的聚类分类等文本数据挖掘工作中。

通过对微博用户信息的进一步分析，发现对用户的描述主要可分为两部分。（1）用户个人信息描述，包括用户名、认证信息、个人简介等，此类数据构成了用户的基本且固定的描述信息；（2）用户所分享的微博信息，此类数据构成了用户实时的且更加详细的描述信息。本文通过分别对个人信息和微博信息进行基于 word2vec 的向量表示，然后将个人信息和微博信息两种词向量表示进行拼接构成用户信息特征表示。

4.4.1 用户个人信息特征描述

针对用户个人信息，本文把各项个人信息看做一篇文档的形式，其转换成向量空

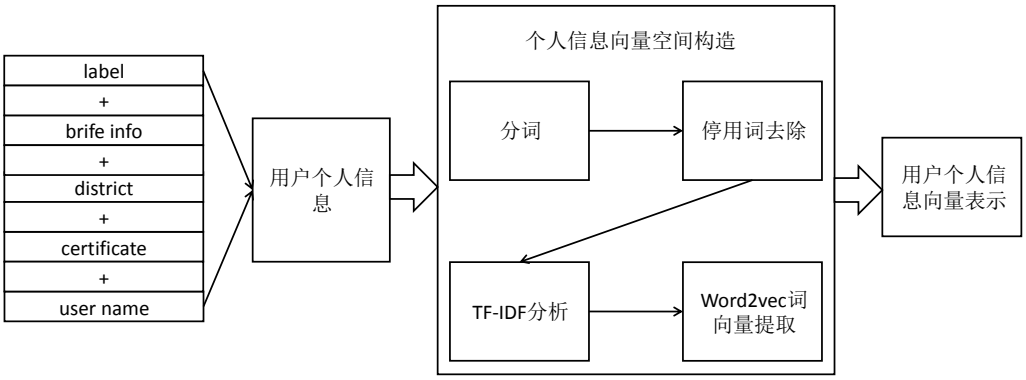

图 4-9 个人信息向量空间表示过程

（1）用户个人信息组成：通过对爬取的用户个人信息中各个属性关键性进行分析，选取出”label”/”brief info”/”district”/”certificate”/”user name”5 项作为个人信息组成。根据个人信息中各属性描述都比较简略，且单一属性所能表述的用户信息较少的缺点。本文提出把各属性描述看做单一的字符串，然后把各属性描述进行字符串拼接，如此形成能够综合性描述形个人信息的文档。拼接效果如图 4-10。

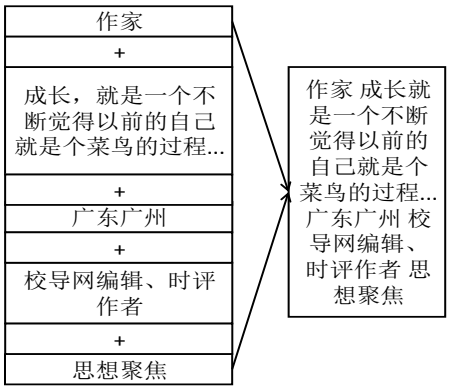

图 4-10 个人信息拼接图

（2）个人信息向量空间构造：从个人信息描述的文本内容转换到个人信息的向量

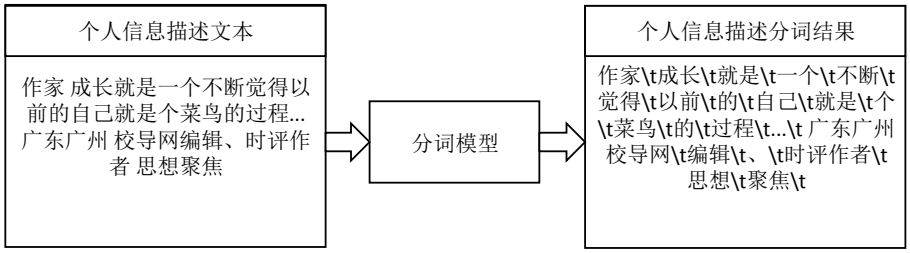

图 4-11 分词效果图

2）去除停用词：

在文本分析中去除停用词是对文本分析效率提升很重要的一个环节。停用词字面意思是没有作用的词语，这些词在大多数文档中都有分布，主要包括介词、语气词、符号等没有实际意义的词语。从信息熵的角度来说明就是这些词的熵值很低，其本身所包含的信息量有限，没有承载与文档相关的语义信息。停用词的存在会对文本分析带来噪声，影响文本分析的准确率，因此在进行文本分析时需要将这些停用词去掉。

本文采用的停用词表是“哈工大停用词表”，将分词结果经过停用词表处理获得最终用于文本分析的过滤后词表。

3）TF-IDF 分析：

词频逆文本频率（term frequency - inverse document frequency 以下简称 TF-IDF）是一种普遍应用在搜索排序和数据挖掘等领域的特征加权算法。TF-IDF 用来为文档中的每一个词进行重要性评价，并给出该词对于该文档的重要性权值。TF-IDF 评估方法是统计待评估词在某一文档中的出现频次以及该词在整个文档库中出现的频

率，然后按照与文档中词频正相关与文档库中词频负相关的原理进行词权重的估计。TF-IDF 的计算表达式是：TF * IDF，TF 指词频(Term Frequency)，IDF 指逆文本频率(Inverse Document Frequency)。TF 是指词语在文档中的频次统计，IDF 是指包含词语的文档数越小则 IDF 越大。

其中 TF、IDF 的计算方法分别如式（4-4）（4-5）所示。

$$tf_{i,j} = \frac{n_{i,j}}{\sum_k n_{k,j}} \tag{4-4}$$

式（4-4）中 $n_{i,j}$ 指词 t_i 的词频。

$$idf_i = \log \frac{|D|}{1+|\{j : t_i \in d_j\}|} \tag{4-5}$$

式（4-5）中 $|D|$ 指语料库中总的文档数，$1+|\{i : t_i \in d_j\}|$ 指包含有词 t_i 的总文档数。

转

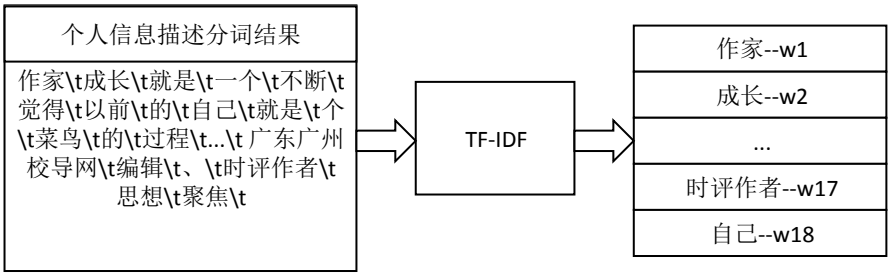

图 4-12　TF-IDF 算法效果图

4）基于 word2vec 的词向量构造：

本文的 word2vec 模型训练语料使用搜狗实验室的互联网媒体新闻语料，把语料中新闻文本的正文选择出来，并且在正文选择过程中对文章进行去重操作，以此提高 word2vec 的词向量表达效率。然后对选择出的正文内容进行分词和词频统计工作，对正文中频率较低的词进行过滤，以提高训练效率。语料准备好之后，通过 word2vec 进行训练，本文设定词向量长度为 200 维，最终处理训练语料后生成 vectors.bin 词向量文件，在实验室的服务器上训练了 2.5 个小时。

然后利用 TF-IDF 处理后的词表和训练好的 vectors.bin 词向量文件构造个人信息向量表示。构造过程如图 4-13。

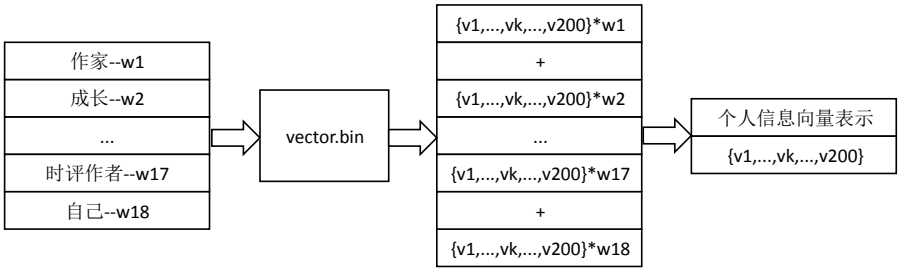

图 4-13 个人信息向量构造

图中{v1,...,vk,...,v200}指词对应的词向量，w 指与词对应的 TF-IDF 权值，最终对文档所包含词列表计算加权平均和，表达式如（4-7）所示。

$$V_{info} = \frac{1}{m} \times \sum_{k} \left(V_k \times w_k \right) \tag{4-7}$$

式（4-7）中V_{info}表示个人信息向量表示，m 表示个人信息中包含词表长度，V_k表示词 k 对应的词向量，w_k指与词 k 对应的 TF-IDF 权值。

4.4.2 用户微博信息特征描述

微博信息是微博用户之间实现信息交流与分享的主要方式，承载着用户爱好、社会角色、性格、专业领域等多方面的个人属性信息。因此用户分析过程中，微博信息的合理利用是十分关键的一部分。本文把用户微博信息按时间顺序进行排序，选取出最近参与的 100 篇作为用户属性信息的载体，然后把微博信息构造成与词向量相同结构的向量形式。

用户微博信息向量构造分为微博信息组成和微博信息向量空间构造两部分。

（1）微博信息组成：由于单篇微博的内容上限是 140 个字符，能包含的个人属性信息有限，因此为便于后续文本处理，本文将多篇微博信息整合为一篇文档。整合形式如图 4-14 所示。

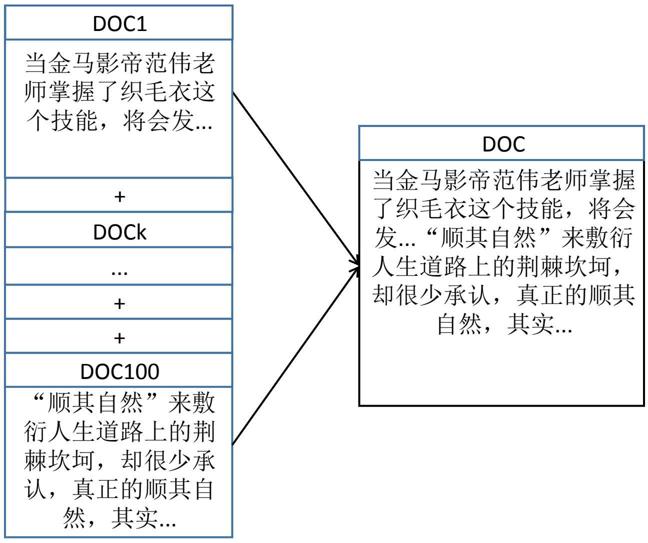

图 4-14 微博信息组成

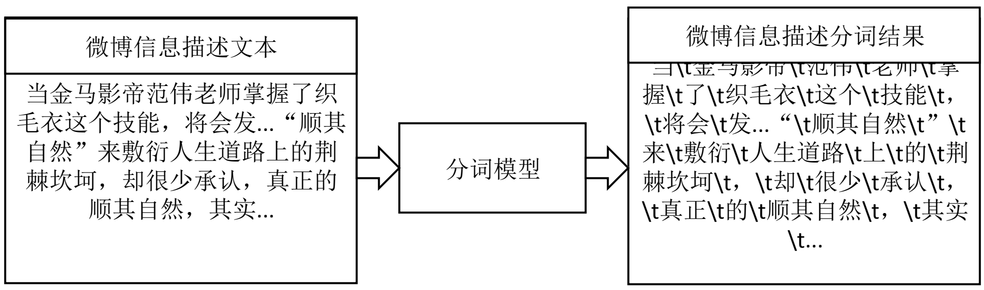

图 4-15 微博信息分词效果图

2）去除停用词：

微博信息分词结果去除停用词采用的词表是上文所述“哈工大停用词表”，将分词结果经过停用词表处理获得最终用于文本分析的过滤后词表。

3）TF-IDF 分析：

经过 TF-IDF 算法分析，词语分散的分词词表转换为“词-权重”的词表形式，转换结果如图 4-16。

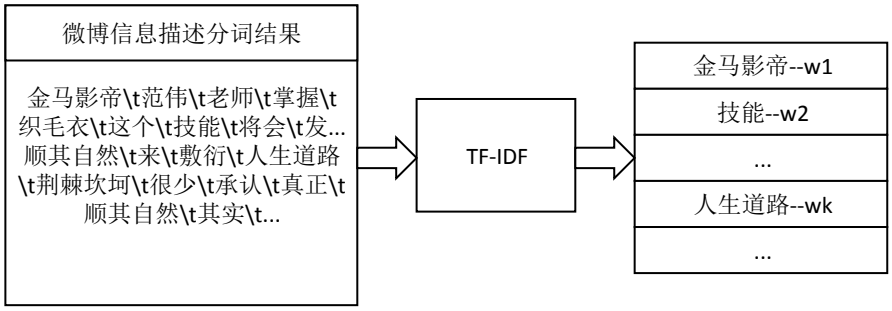

图 4-16　TF-IDF 分析流程

4）word2vec 词向量提取：

利用 TF-IDF 处理后的词表和训练好的 vectors.bin 词向量文件构造微博信息向量表示。构造过程如图 4-17。

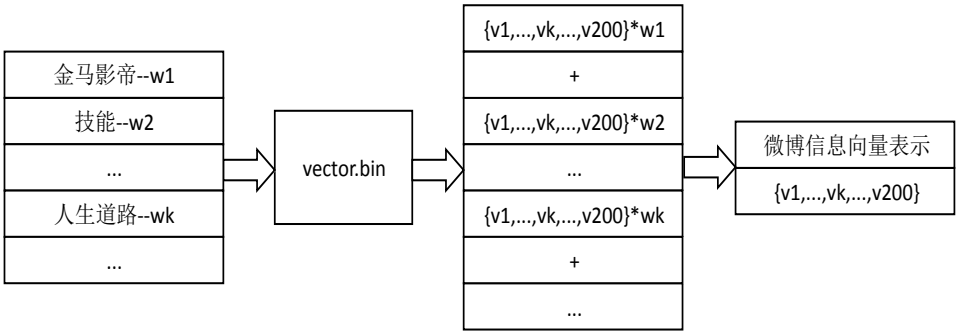

图 4-17 微博信息向量空间构造流程

图中{v1,...,vk,...,v200}指词对应的词向量，w 指与词对应的 TF-IDF 权值，最终对文档所包含词列表计算加权平均和，表达式如（4-8）所示。

$$V_{info} = \frac{1}{m} \times \sum_k \left(V_k \times w_k \right) \qquad (4\text{-}8)$$

式（4-8）中 V_{info} 表示微博信息向量表示，m 表示微博信息中包含词表长度，V_k 表示词 k 对应的词向量，w_k 指与词 k 对应的 TF-IDF 权值。

4.4.3 用户个人信息和微博信息融合

为实现对用户进行更全面的描述，本文采用拼接的策略，即将用户个人信息和微博信息分别用 Word2vec 进行表达，然后将两种表述结果进行向量拼接，构成 400 维描述的信息。通过此拼接方法，用户个人信息和微博信息在进行之后的用户分析时能够起到更多维度的作用，相比将两种维度数据直接相加能达到更丰富的用户表述效果。用户信息描述如图 4-18。

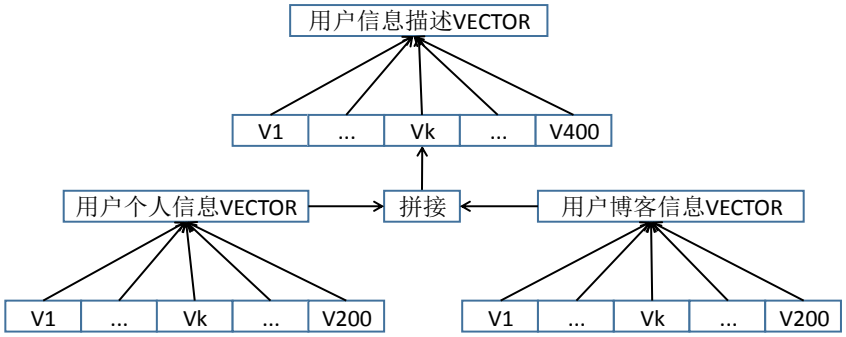

图 4-18 用户信息表述拼接结构图

4.5 本章小结

本章对微博数据挖掘中的数据处理工作进行研究。首先对中文文本信息处理中的中文分词算法进行行了研究，并提出基于词典匹配与统计标注相融合的分词算法，使分词算法的领域适应性更强。然后对用户信息表示方法进行研究，提出基于 word2vec 的词嵌入式用户信息表示方法，将传统的词表向量表示方法优化为使用低维向量表示方法，为后续数据挖掘工作提供了高效的用户信息表示方式。

第 5 章 社交网络微博数据挖掘

新浪微博具有广大的用户群，而且随着移动终端的普及用户数据量也会逐年攀升。针对用户数据量大这一特点，本章主要对大量用户中的垃圾用户识别和用户聚类分析两项工作进行研究。

5.1 基于 SVM 模型的垃圾用户识别

垃圾用户主要包括僵尸粉丝和广告用户两大类，垃圾用户在社交网络中的不良行为会扰乱正常的社交网络秩序，影响社交网络用户的信息分享与接受的体验。本节依据以上章节构造的用户向量表示方法，训练基于 SVM 算法的垃圾用户识别模型。本文采用人工标注的垃圾用户识别语料，将用户标注为正常用户和垃圾用户两类。选取垃圾用户和正常用户各 10000 例用于 SVM 模型的训练与测试，训练与测试语料比例划分采用 8:2 来划分。

5.1.1 基于 SVM 模型的分类算法

SVM 模型的学习目标就是在指定维数的数据向量空间中找到一个分类超平面，超平面表达式可表示为式（5-1）。

$$w^T x+b=0 \tag{5-1}$$

式（5-1）中，w 代表模型参数，x 代表 n 维的数据特征向量，b 代表模型偏置参数。

由式（5-1）可得 SVM 的分类函数表达式如式（5-2）。

$$f(x)=w^T x+b \tag{5-2}$$

由式(5-2)，若 $(x)=0$ 则 x 位于超平面上，若 $f(x)<0$ 则 x 被分为负类即 $y=-1$，若 $f(x)>0$ 则 x 被分为正类即 $y=1$。直观的 SVM 超平面分类效果如图 5-1。

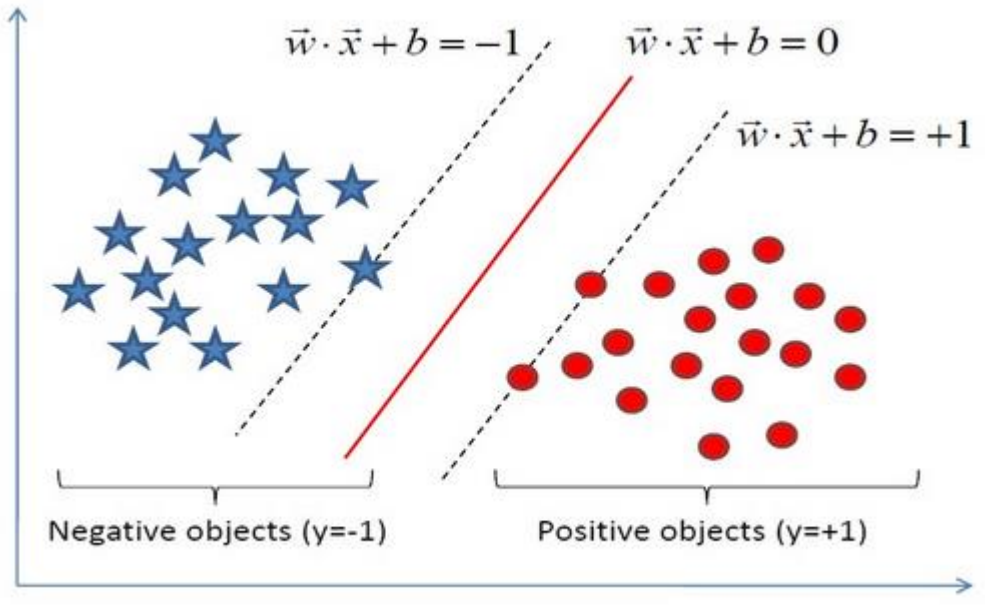

图 5-1 SVM 分类效果图

如图 20 中所示，实线代表分类超平面（$\bar{w}\bullet\bar{x}+b=0$），则数据集中样本点$(x_i,y_i)$到超平面的几何距离如式（5-3）所示。

$$\gamma_i=y_i\left(\frac{w}{\|w\|}\bullet x_i+\frac{b}{\|w\|}\right) \tag{5-3}$$

式（5-3）中，γ_i代表样本到超平面的几何距离，y_i代表样本类别，$\|w\|$为w的L_2范数。

接下来计算数据中样本点到超平面几何距离的最小值γ，并称其为几何间隔。

$$\gamma=\min_{i=1,\ldots,N}\gamma_i \tag{5-4}$$

训练过程中使几何间隔最大化的含义是可以使 SVM 算法模型能够以尽可能大的确信度对样本进行划分，进而当模型面对未知的新样本数据时能够达到良好的分类想能指标。几何间隔最大化可表示为式（5-5）所示的约束最优化问题。

$$\left\{\begin{array}{l}\max\limits_{w,b}\gamma\\ s.t.y_i\left(\dfrac{w}{\|w\|}\bullet x_i+\dfrac{b}{\|w\|}\right)\geq\gamma,i=1,2,\ldots,N\end{array}\right\} \tag{5-5}$$

通过变换并引入拉格朗日乘子，可导出式（5-6）。

$$L(w,b,\alpha)=\frac{1}{2}\|w\|^2-\sum_{i=1}^{n}\alpha_i\left(y_i\left(w^Tx_i+b\right)-1\right) \tag{5-6}$$

因此约束最优化问题可转换成如下目标函数（5-7）。

$$p^*=\min_{w,b}\max_{\alpha_i\geq0}L(w,b,\alpha) \tag{5-7}$$

通过对偶算法求解式（5-7），最终获得目标函数的解，即 SVM 模型参数 w。训练好模型之后，利用式（5-2）便可进行样本的分类任务。

SVM 分类算法在处理非线性可分问题时，能够通过引入核函数的策略将样本特征映射到高维特征空间进行表示，从而使样本变换为高维空间的线性可分问题，既提高计算效率，又能表现出良好的分类性能。而本文的分类目标是新浪微博用户中的垃圾用户与正常用户，其特征选择经过 word2vec 转换后变为固定的 200 维，由于 word2vec 的指定维数并不具有特定含义，因此会造成特征的交错，从而使用户分类线性不可分问题。因此本文采用 SVM 作为垃圾用户识别的分类模型。

5.1.2 SVM 模型中核函数选择

模型结构选择过程中，本文选择高斯径向基核函数进行数据特征映射方法。径向基函数 (Radial Basis Function 以下简称：RBF)，是一类沿径向对称的标量函数。RBF 表达的含义是样本表示空间中某一点 x 到某一中心 c 之间欧几里得距离的单调标量函数，可记作 $k(\|x-c\|)$。高斯径向基核函数是常用的径向基函数中的一种，其表达式如式（5-8）所示。式（5-8）中 c 为核函数中心，σ 为函数的区间度量参数，决定着函数的径向作用区间。而且高斯径向基核函数能够解决特征的非线性映射，并且不会像多项式核函数一样使模型复杂度大幅度提高。

$$k(\|x-c\|)=\exp\left\{-\|x-c\|^{\frac{2}{(2*\sigma^2)}}\right\} \tag{5-8}$$

高斯核函数能将数据映射到无穷维度空间，进而使特征表示维度的提升也是无穷的，即总能找到一个分类面将数据集很好的分开。特征表示维度代表了分类能力，使用 SVM 的时候，只要特征表示维度足够大，就可以保证分类模型拟合的很好，但是不加限制的扩大会导致分类模型对训练数据过拟合。

通过 SVM 模型分类后的用户数据，过滤掉了垃圾用户对后续用户数据挖掘的造成干扰，此部分是整个数据挖掘与分析中十分重要的一项工作。并且通过本节构造的用户向量空间表示，可省去接下来分析工作中特征选择、特征降维处理等工作。

5.1.3 SVM 模型中松弛变量控制

训练数据由人工进行标注，特征不明显的情况下，容易造成数据标注错误，因而容易产生噪声数据。个别噪声数据会造成数据线性不可分，无限增加特征表示维度又会造成模型过于复杂，甚至对训练数据过拟合，导致在测试数据上的分类效果

差等问题。本文利用松弛变量来应对个别标注错误的训练样本点，松弛变量的引入方式如式（5-9）。

$$\begin{Bmatrix} \min \frac{1}{2}\|w\|^2 (1) \\ s.t. y_i\left(w^T x_i + b\right) \geq 1-\xi_i, i=1,2,...,N (2) \\ \xi_i \geq 0 \end{Bmatrix} \tag{5-9}$$

式（5-9）中ξ_i为松弛变量，其是非负的，因此最终的结果是要求间隔可以比 1 小。当样本点出现这种间隔比 1 小的情况时，会利用ξ_i去调节，使式仍然成立。通过ξ_i的调节使分类面不必向噪声数据点的方向移动，在低维空间看来，分类边界也更平滑。

式（5-9）中（1）是 SVM 模型的目标函数，ξ_i会造成目标函数的损失，使式（5-9）中（1）的最值变大。为避免松弛变量在调节过程中使模型目标函数损失太大，造成分类模型效果变差，在次把松弛变量引入到目标函数中，对其调节范围进行约束，最终 SVM 模型如式（5-10）所示。

$$\begin{Bmatrix} \min \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{L}\xi_i^2 \\ s.t. y_i\left(w^T x_i + b\right) \geq 1-\xi_i, i=1,2,...,N \\ \xi_i \geq 0 \end{Bmatrix} \tag{5-10}$$

式（5-10）中，C代表惩罚因子，是可供人工调节的超参数。通过调节C影响松弛变量对噪声数据点的调节范围，也即通过调节C进行数据拟合度的控制。

通过调节惩罚系数，当C值为 8 时最终使 SVM 模型在测试数据集上测试效果达到良好的分类效果，准确率为 0.96、召回率为 0.92、F 值为 0.94。

5.2 基于 SVM 的用户识别实验

本节采用人工标注的垃圾用户分类语料，将用户标注为正常用户和垃圾用户两类。选取垃圾用户和正常用户各 10000 例用于 SVM 模型的训练与测试，训练与测试语料比例划分采用 8:2 来划分。模型结构选择过程中，本节选择高斯径向基核函数进行数据特征映射方法，通过调节惩罚项系数达到最优的测试集分类效果。惩罚项系数调节过程与分类效果关系如图 5-2,5-3。

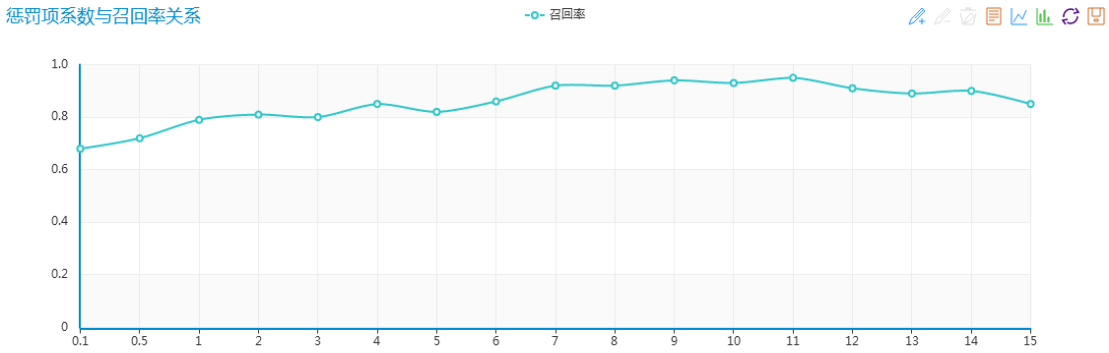

图 5-2 惩罚项系数调节过程与分类召回率关系

图 5-2 中数据如表 5-1 所示：

表 5-1 惩罚项系数与召回率关系数据

惩罚项系数	0.1	0.5	1	2	3	4
召回率	0.64	0.72	0.79	0.81	0.8	0.85
	5	6	7	8	9	10
	0.82	0.86	0.92	0.92	0.94	0.93
	11	12	13	14	15	
	0.95	0.91	0.89	0.9	0.85	

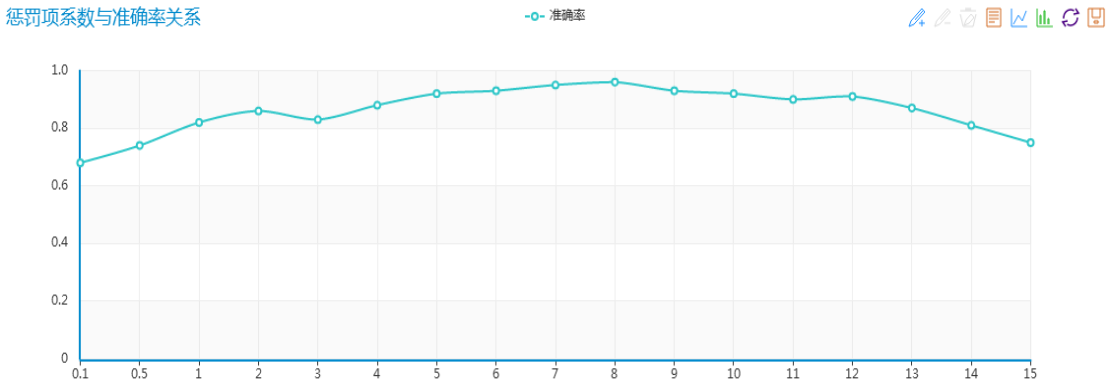

图 5-3 惩罚项系数调节过程与分类准确率关系

图 5-3 中数据如表 5-2 所示

表 5-2 惩罚项系数与准确率关系数据

惩罚项系数	0.1	0.5	1	2	3	4
准确率	0.68	0.74	0.82	0.86	0.83	0.88
	5	6	7	8	9	10
	0.92	0.93	0.95	0.96	0.93	0.92
	11	12	13	14	15	
	0.9	0.91	0.87	0.81	0.75	

本节实验通过设计好的 SVM 模型，并通过调节惩罚项系数，当系数为 8 时，准确率为 0.96、召回率为 0.92、F 值为 0.94，达到最优分类效果。

5.3 基于 K-means 算法的用户聚类分析

新浪微博的使用人群数量基数大，截至 2016 年 6 月微博用户规模为 2.42 亿。而且微博信息更新频繁、信息传播迅速，各个行业、社会阶层、年龄段的人都有参与其中。因此按参与内容进行社区划分对于社会属性数据挖掘以及用户相关内容推荐具有重要意义，本节采用 K-means 聚类算法实现新浪微博用户聚类分析。

5.3.1 K-means 聚类算法

K-means 算法是一种基于距离度量的聚类算法，其把距离度量当作样本间相似性的衡量标准，即假设样本间两两距离越近则相似度就越大。该算法的思想是相似度很高的样本在相应的表示空间中是成簇堆积在一起的，因此算法目标就是获取到簇内紧凑度高且簇间区分度明显的聚类效果。

K-means 聚类算法的聚类中心数值为 k，算法实施首先要选定 k 个中心点的位置，不同的初始中心点选取方案可能会造成不同的聚类结果，因此 k 的选择对于算法的实施具有重要意义。k 个中心点确定后开始进行聚类运算的迭代工作，每次迭代的计算过程是计算某个样本点到每个聚类中心点的距离，，选取距离最小的聚类中心点作为改样本所归属的聚类簇，直到所有的样本点都被计算一遍，计算出新的聚类中心点的向量值，这样一次迭代结束。如果在一次迭代前后，误差准则函数 J 的值没有发生明显变化，说明算法已经收敛。误差准则函数如式（5-11）。

$$J=\sum_{i=1}^{k}\sum_{P\in C_i}\left|p-m_i\right|^2 \tag{5-11}$$

式（5-11）中 J 即误差准则函数值，k 指聚类指定类中心数，p 指新的类中心，m 指本次迭代之前的类中心。

算法过程如下：

1）从 N 个文档文档中随机选取一个文档 a_1 作为一个类中心，然后选取与 a_1 距离最大的 a_2 作为第二个聚类中心点，然后计算与 a_1、a_2 距离最大的 a_3 作为第三个聚类中心点，如此类推直到确定第 k 个类中心点 a_k；

2）计算样本点到每个聚类中心的距离，并把该样本划分到与之距离最近的聚类簇中；

3）重新计算已经得到的各个聚类簇的类中心值；

4）重复迭代 2～3 步，直到迭代次数预先设定的最大迭代次数或者前后两次迭代后类中心的变化指小于预定的阀值，则算法结束。

5.3.2 K-means 用户聚类算法评估

在 K-means 聚类算法中 K 值是人为根据已知条件做出的设定参数，除非有丰富的经验对数据类别有全面的了解否则 K 值的确定是一件比较困难的事情。本文通过研究发现，利用 K-means 聚类算法的评估算法 DB-index 可以进一步获取合适的 K 值。

对聚类效果的评估标准主要可分为类内样本到聚类中心的平均距离和聚类中心两两之间的平均距离两大类。

其中类内样本到聚类中心的平均距离计算方法如式（5-12）、（5-13）所示。

$$\overline{CP_i} = \frac{1}{|\Omega_i|}\sum_{x_i \in \Omega_i}\|x_i - w_i\| \tag{5-12}$$

$$\overline{CP} = \frac{1}{K}\sum_{k=1}^{K}\overline{CP_k} \tag{5-13}$$

式（5-12）、（5-13）中 CP 指类内样本到聚类中心的平均距离，Ω 指样本表示的向量，x 指样本向量的值，w 指类中心向量的值。

聚类中心两两之间的平均距离计算方法如式（5-14）所示。

$$\overline{SP} = \frac{2}{k*(k-1)}\sum_{i=1}^{k}\sum_{j=i+1}^{k}\|w_i - w_j\| \tag{5-14}$$

式（5-14）中 SP 指聚类中心两两之间的平均距离，w 指类中心向量的值。

由于类内样本到聚类中心的平均距离和聚类中心两两之间的平均距离都仅仅考虑了聚类结果的单方面的属性，对于聚类效果的评估具有局限性。本文引入 DB-index 算法，该算法综合考虑了类间距离和类内聚合度两项聚类评价标准，其算法表达式如式（5-15）所示。

$$DB = \frac{1}{k}\sum_{i=1}^{k}\max_{j\neq i}\left(\frac{\overline{C_i}+\overline{C_j}}{\|w_i - w_j\|}\right) \tag{5-15}$$

式（5-15）中 DB 指评估效果值，C 指类内样本到聚类中心的平均距离，w 指类中心向量的值。

由式（5-15）可知，DB-index 算法计算任意两类别的类内样本到聚类中心的平均距离之和然后除以两类中心之间的距离，并求其中的最大值。因此 DB-index 评

估值值越小意味着类内距离越小同时类间距离越大，此评估算法能更全面的作用于聚类效果的评估。

根据 DB-index 的这一特性，本文选定 k 值在一定范围内，并分别对相应的 k 值进行 k-means 聚类运算，然后计算 DB-index 值最小的 k 作为最合适的聚类类中心数。

5.3.3 用户聚类分析

基于上述章节的理论分析，本节实现用户的聚类分析。文本实现的聚类分析并不是针对特定的某一属性的针对性类别划分，而是综合考虑用户多方面属性，找到用户更适合的社交社区归属。其中包含用户的兴趣爱好，职业，社会状态，关注内容等信息，如此设计的目的是挖掘出用户在社交网络中构成的隐性社区，而不是根据已经明确定义的社交圈子进行划分。聚类分析中采用垃圾用户过滤时采用的基于 word2vec 构造的用户向量空间表示。虽然此 400 维向量中的每一维没有指定明确含义，但是其实通过用户个人信息和用户微博信息融合而成。其包含了用户多方面的属性信息在向量中，且充分满足本文聚类分析的目标需求。

（1）计算合理的聚类中心数量：隐性社区的划分种类没有明确的定义，本文利用 DB-index 聚类结果评估算法，获取最佳的社区划分数量。

根据实际聚类实验分析，本文计算了聚类中心数量 2-25 范围内的 DB-index 评估值。根据评估值的后期变化曲线判断其评估值变化接近稳定，并选取了其中的最小评估值作为最适合的类中心数，对应的类中心数为 12。

（2）用户聚类结果分析：本文用户数据挖掘目标是聚类出社交用户的隐性社区划分，划分后的聚类结果没有明确的名称定义。为得到聚类结果的直观认识，对于聚类结果中每一类用户采用词云图的方式进行展示。

通过基于 SVM 的垃圾用户识别模型，本文从爬取的用户数据中，获取 10 万正常用户数据。利用聚类算法分析，获取最终分为 12 类的聚类结果。对每一类的用户数据进行词频统计处理，根据词频统计结果进行词云图构建，展示聚类结果中每一类所体现的用户主要特征。

用户信息词频统计过程如图 5-4。

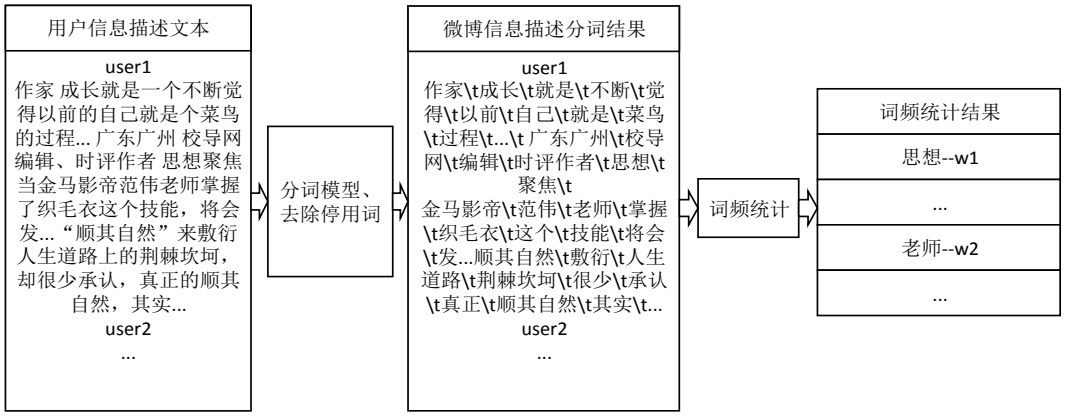

图 5-4 用户信息词频统计过程

通过词云图可直观表达出聚类结果中的每一类具有鲜明的领域性、社会阶层性、群体交互性等特征。本文中基于用户隐性关系进行聚类分析，可对未来社会事件传播、社交热点事件产生、用户个性化推荐等研究课题提供理论与技术支持，对社交网络科学研究具有重要意义。

5.4 基于 K-means 的用户聚类算法实验

本节利用训练好的 SVM 用户分类模型，从已经爬取的用户数据库中提取出 10 万正常用户数据用于聚类分析。聚类分析采用 K-means 算法，并通过 DB-index 评估算法得到最佳聚类中心数值。

其中 DB-index 评估值与聚类中心数值 K 的变化关系如图 5-5。

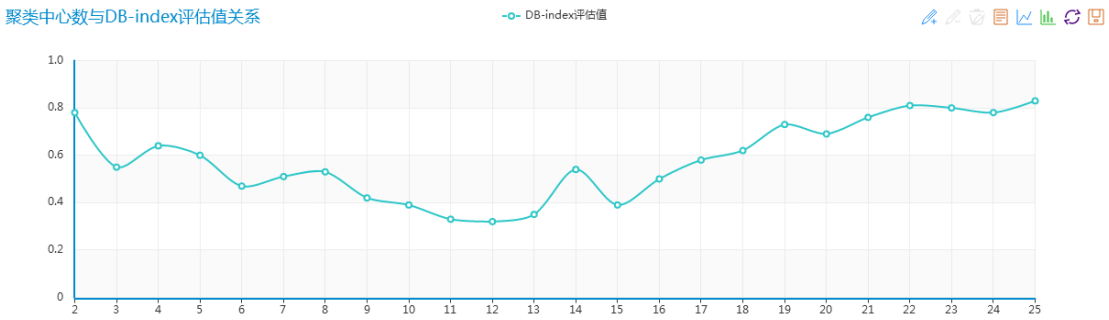

图 5-5 DB-index 评估值与聚类中心数值 K 的变化关系

图 5-5 中数据关系如表 5-3 所示：

表 5-3 聚类中心数值与 DB-index 关系

聚类中心数值	2	3	4	5	6	7
DB-index	0.78	0.55	0.64	0.6	0.47	0.51
	8	9	10	11	12	13
	0.53	0.42	0.39	0.33	0.32	0.35
	14	15	16	17	18	19
	0.54	0.39	0.5	0.58	0.62	0.73
	20	21	22	23	24	25
	0.69	0.76	0.81	0.8	0.78	0.83

最终通过对图 5-5 中 DB-index 评估值变化曲线分析，得到最佳聚类中心数值为 12。

本节引入词云图数据展示工具对聚类结果的类簇属性进行词云图展示，以此直观分析类簇属性特征，展示其中四个聚类中心的数据，效果如图 5-6,5-7,5-8,5-9。

图 5-6 聚类效果词云图 1

图 5-7 聚类效果词云图 2

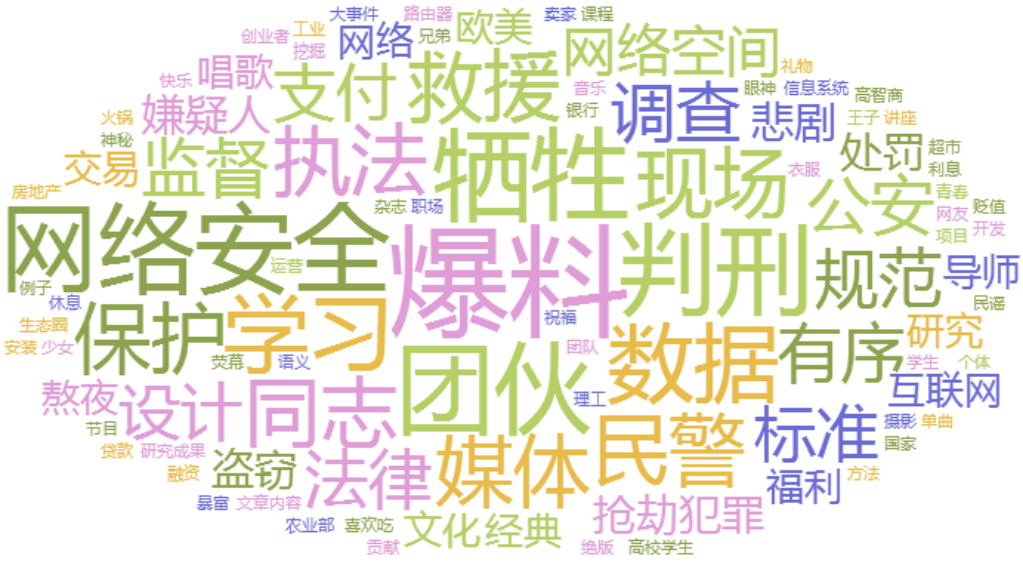

图 5-8 聚类效果词云图 3

图 5-9 聚类效果词云图 4

根据词云图内容我们可以总结出，聚类效果词云图 1 表示用户关注内容是校园生活、休闲娱乐等；聚类效果词云图 2 表示用户关注内容是经济发展、商业数据等；聚类效果词云图 3 表示用户关注内容是安全、救援、治安等；聚类效果词云图 2 表示用户关注内容是科技发展、科学研究等。

通过词云图的直观展示可知聚类结果中每一类都体现了该类中属性特点，并且类之间有明显区别。词云图的展示效果验证了通过 DB-index 评估值获取的聚类中心数值的合理性。

5.5 本章小结

本章在数据准备与数据处理的基础上对微博用户数据进行了数据挖掘工作。根据微博用户中用户类型，利用 SVM 分类模型进行垃圾用户识别研究，并通过引入高斯径向基核函数和松弛变量，最终实现测试集上 F 值为 0.94 的垃圾用户识别模型。针对用户关注方向，利用个人数据和微博数据，对用户进行 K-means 聚类研究，并通过 DB-index 评估值计算最优聚类中心数值，得到最佳聚类效果。最后通过算法实验验证垃圾用户识别模型的有效性，通过数据可视化方法分析用户聚类算法的良好效果。

第 6 章 总结与展望

微博作为社交网络热门平台，拥有着广阔的发展前景和重要的社会信息数据挖掘价值。用户在平台上共享热点事件，关注共同话题，构建关系网络，使微博拥有更实际的存在价值。本文针对微博数据特点，并结合文本数据挖掘技术，对微博用户进行数据爬取与存储、中文分词、用户向量空间表示、垃圾用户识别、用户聚类分析等方面进行研究与分析，以上内容构成了论文的主要内容。

6.1 论文的主要贡献

本论文在如下几方面进行研究与探索：

（1）数据爬取与存储，由于目前官方提供的微博数据获取方法具有限制，本文根据现有网络爬虫技术，通过研究用户访问微博平台网络流程，设计出基于模拟登录的用户数据爬虫系统，实现了微博用户数据无限制的获取功能。由于现有关系型数据库在存储微博数据方面的性能缺陷，本文根据根据用户数据特征，采用 JSON 数据存储格式，将数据存储在 NOSQL 型数据库中，提高了数据存储的灵活性。

（2）中文分词，由于微博数据多是关于社会热点问题，对分词的新词发现和准确率要求较高，本文通过对现有分词方法中分词方法存在新词发现效果差或者需要大量人工标注工作等问题进行研究，提出了基于词典匹配与统计标注相融合的分词算法。该算法充分利用基于词典匹配与基于统计标注分词方法的优势，融合两种分词方法特长，分词效果的 F 值超过单独的词典匹配和统计标注方法。

（3）用户向量空间表示：通过对现有文本表示方法的研究，本文对于用户数据表示采用了基于 word2vec 的词嵌入向量表示方法，使用户数据的特征表示更能体现上下文语义，相对于 one-hot representation 特征表示方法提升计算效率。

（4）垃圾用户识别：由于微博中垃圾用户本会影响社交体验，并且会为社交网络数据挖掘带来噪声干扰。本文通过对现有分类算法进行研究与比较，选择出在较小样本量下分类效果很好的 SVM 模型。并通过在模型中加入核函数和松弛变量进一步提升了本文垃圾用户识别模型的分类效果，使其在测试集上的分类结果的 F 值达到 0.94。

（5）用户聚类分析：针对目前根据用户关注内容进行社区性划分不明确的问题，本文通过对现有无监督聚类方法的特点进行研究。提出基于 K-means 的用户社区性划分方法，并提出了基于 DB-index 评估值选取最优聚类中心数值的聚类算法。

通过此聚类算法对微博用户进行聚类分析，利用词云图的可视化方法对聚类结果进行展示。

6.2 工作展望

微博作为当前火热的社交应用平台，仍具有广阔的发展前景，社会存在价值也会进一步提升。因此，对微博数据进行数据挖掘研究具有重要意义。在本文研究内容之外还有很多方面需要拓展研究。

（1）微博数据中多媒体数据的数据挖掘研究，包括图像，视频等，用于用户的情感分析，对微博流行程度做出预测等。

（2）基于用户聚类数据的个性化推荐，优化用户社交体验。

（3）基于微博数据进行热点社会事件发生以及传播趋势预测等。

参考文献

[1] 彭京, 杨冬青, 唐世渭,等. 一种基于语义内积空间模型的文本聚类算法[J]. 计算机学报, 2007, 30(8):1354-1363.

[2] 陈克寒, 韩盼盼, 吴健. 基于用户聚类的异构社交网络推荐算法[J]. 计算机学报, 2013, 36(2):349-359.

[3] Carter S, Weerkamp W, Tsagkias M. Microblog language identification: overcoming the limitations of short, unedited and idiomatic text[J]. Language Resources and Evaluation, 2013, 47(1):195-215.

[4] Efron M. Information search and retrieval in microblogs.[J]. Journal of the Association for Information Science and Technology, 2011, 62(6):996–1008.

[5] Stefan Stieglitz, Linh DangXuan. Emotions and Information Diffusion in Social Media—Sentiment of Microblogs and Sharing Behavior[J]. Journal of Management Information Systems, 2013, 29(4):217-248.

[6] Takehara T, Miki S, Nitta N, et al. Extracting Context Information from Microblog Based on Analysis of Online Reviews[C]// IEEE International Conference on Multimedia and Expo Workshops. IEEE, 2012:248-253.

[7] Oulasvirta A, Lehtonen E, Kurvinen E, et al. Making the ordinary visible in microblogs[J]. Personal and Ubiquitous Computing, 2010, 14(3):237-249.

[8] Lee R, Wakamiya S, Sumiya K. Discovery of unusual regional social activities using geo-tagged microblogs[J]. World Wide Web, 2011, 14(4):321-349.

[9] Ebner M, Lienhardt C, Rohs M, et al. Microblogs in Higher Education – A chance to facilitate informal and process-oriented learning?[J]. Computers & Education, 2010, 55(1):92-100.

[10] Marques A M, Krejci R, Siqueira S W M, et al. Structuring the discourse on social networks for learning: Case studies on blogs and microblogs[J]. Computers in Human Behavior, 2013, 29(2):395-400.

[11] 王晓兰. 2010 年中国微博客研究综述[J]. 国际新闻界, 2011(1):24-27.

[12] 闫幸, 常亚平. 微博研究综述[J]. 情报杂志, 2011, 30(9):61-65.

[13] 孙晓莹, 李大展, 王水. 国内微博研究的发展与机遇[J]. 情报杂志, 2012, 31(7):25-33.

[14] 周德懋, 李舟军. 高性能网络爬虫:研究综述[J]. 计算机科学, 2009, 36(8):26-29.

[15] 李志义. 网络爬虫的优化策略探略[J]. 现代情报, 2011, 31(10):31-35.

[16] 周庆燕, 何利力, 胡靖枫. 搜索引擎中网络爬虫策略在烟草行业中的应用研究[J]. 工业控制计算机, 2014(12).

[17] Baeza-Yates R, Castillo C, Marin M, et al. Crawling a country: better strategies than breadth-first for web page ordering[C]//Special interest tracks and posters of the 14th international conference on World Wide Web. ACM, 2005: 864-872.

[18] Olston C, Najork M. Web crawling[J]. Foundations and Trends in Information Retrieval, 2010, 4(3): 175-246.

[19] 杜长燕，李祥龙. 基于 WEB 的网络爬虫的设计[J]. 无线互联科技，2015(5):49-50.

[20] Cacheda F, Fernández D, López R. Experiences on a practical course of web information retrieval: Developing a search engine[C]//Second International Workshop on Teaching and Learning of Information Retrieval (TLIR 2008). 2008.

[21] Wei-jiang L, Hua-suo R, Kun H, et al. A New Algorithm of Blog-Oriented Crawler[C]//Computer Science-Technology and Applications, 2009. IFCSTA'09. International Forum on. IEEE, 2009, 1: 428-431.

[22] 罗桂琼，费洪晓，戴弋. 基于反序词典的中文分词技术研究[J]. 计算机技术与发展, 2008, 18(1):80-83.

[23] 迟呈英，于长远，战学刚. 基于条件随机场的中文分词方法[J]. 情报杂志, 2008, 27(5):79-81.

[24] Vail D L, Veloso M M, Lafferty J D. Conditional random fields for activity recognition[C]//Proceedings of the 6th international joint conference on Autonomous agents and multiagent systems. ACM, 2007: 235.

[25] Xue Nianwen. Chinese word segmentation as character tagging[J].The Association for Computational Linguistics and Chinese Language Processing,Vol.8,No.1,February 2003,pp.29-48.

[26] Rocktäschel T, Bosnjak M, Singh S, et al. Low-dimensional embeddings of logic[C]//Proceedings of the ACL 2014 Workshop on Semantic Parsing. 2014: 45-49.

[27] Turian J, Ratinov L, Bengio Y. Word representations: a simple and general method for semi-supervised learning[C]//Proceedings of the 48th annual meeting of the association for computational linguistics. Association for Computational Linguistics, 2010: 384-394.

[28] Goldberg Y, Levy O. word2vec Explained: deriving Mikolov et al.'s negative-sampling word-embedding method[J]. arXiv preprint arXiv:1402.3722, 2014.

[29] Ewbank M P, Schluppeck D, Andrews T J. fMR-adaptation reveals a distributed representation of inanimate objects and places in human visual cortex[J]. Neuroimage, 2005, 28(1): 268-279.

[30] Mikolov T, Sutskever I, Chen K, et al. Distributed representations of words and phrases and their compositionality[C]//Advances in neural information processing systems. 2013: 3111-3119.

[31] Hinton G E. Distributed representations[J]. 1984.

[32] Deerwester S, Dumais S T, Furnas G W, et al. Indexing by latent semantic analysis[J]. Journal of the American society for information science, 1990, 41(6): 391.

[33] Blei D M, Ng A Y, Jordan M I. Latent dirichlet allocation[J]. Journal of machine Learning research, 2003, 3(Jan): 993-1022.

[34] Rish I. An empirical study of the naive Bayes classifier[C]//IJCAI 2001 workshop on empirical methods in artificial intelligence. IBM New York, 2001, 3(22): 41-46.

[35] Sreenivas M K, AlSabti K, Ranka S. Parallel out-of-core decision tree classifiers[J]. Advances in Distributed and Parallel Knowledge Discovery. Kargupta H. and Chan P.(eds.). AAAI, 2000: 317-336.

[36] Schölkopf B, Simard P, Vapnik V, et al. Improving the accuracy and speed of support vector machines[J]. Advances in neural information processing systems, 1997, 9: 375-381.

[37] Zhao Y, Karypis G, Fayyad U. Hierarchical clustering algorithms for document datasets[J]. Data mining and knowledge discovery, 2005, 10(2): 141-168.

[38] Jain A K. Data clustering: 50 years beyond K-means[J]. Pattern recognition letters, 2010, 31(8): 651-666.

[39] Rinaldo A, Wasserman L. Generalized density clustering[J]. The Annals of Statistics, 2010: 2678-2722.

[40] Huang J Z, Ng M K, Rong H, et al. Automated variable weighting in k-means type clustering[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2005, 27(5): 657-668.

[41] Pham D T, Dimov S S, Nguyen C D. Selection of K in K-means clustering[J]. Proceedings of the Institution of Mechanical Engineers, Part C: Journal of Mechanical Engineering Science, 2005, 219(1): 103-119.

攻读硕士学位期间发表的论文及参与科研情况

参与科研情况

输电线路在线监测与诊断系统建设，鄂尔多斯电业局，2015.7-2016.7

致谢

衷心感谢我的导师翟学明教授。翟老师在研究方向上给我很大的自由空间，激发我的兴趣，使我能够能够充分发挥自己的能力去探索并解决问题。与导师多次关于研究方向上问题的探讨，使我感受到翟老师学术视野的广阔和对于学术研究的认真态度。翟老师的指导和教诲是我研究工作顺利进行的关键保障。在学习上翟老师也给我很大的支持，支持我出去见识研究内容相关的先进技术的实际应用场景，使我对自己的研究课题有了更全面的认识。在生活中，翟老师更像是亲人或朋友，营造了融洽的交流氛围。在如此舒适的环境中成长，让我感到很幸运。

感谢实验室的张东阳老师、闫磊老师和王晓辉老师。他们对我的课外实践和科研内容进行了耐心的指导，帮助我完成了研究生阶段的许多学习任务。老师们乐观的生活态度，热心付出的心态使实验室更像一个和睦的大家庭，更使得我和其他同学能够在其中快乐的成长。

感谢实验室的同学和其他帮助过我的朋友。感谢杨泽师弟耐心帮我爬取数据，为我的研究工作提供了很大便利。感谢黑阳和我组队参加了研究生期间的科技型赛事，通过坚持和努力取得了我们满意的成绩。

特别感谢我的家人，他们在我背后的关心鼓励以及他们的默默付出，使我的科研之路走得更踏实更有信心。