基于句法和语义的话题细粒度情感分析的研究

廖纯

2016 年 1 月

中图分类号：TP391
UDC 分类号：004.8

基于句法和语义的话题细粒度情感分析的研究

作 者 姓 名　廖纯
学 院 名 称　计算机学院
指 导 教 师　冯冲　副研究员
答辩委员会主席　廖乐健　教授
申请学位级别　工学硕士
学 科 专 业　计算机科学与技术
学位授予单位　北京理工大学
论文答辩日期　2016 年 1 月

Study on Topical Fine-grained Sentiment Analysis using Syntax and Semantics

Candidate Name:	Chun Liao
School or Department:	Computer Science and Engineering
Faculty Mentor:	Associate FellowChong Feng
Chair, Thesis Committee:	Prof. Lejian Liao
Degree Applied:	Master of Science
Major:	Computer Science and Technology
Degree by:	Beijing Institute of Technology
The Date of Defence:	January，2016

研究成果声明

本人郑重声明：所提交的学位论文是我本人在指导教师的指导下进行的研究工作获得的研究成果。尽我所知，文中除特别标注和致谢的地方外，学位论文中不包含其他人已经发表或撰写过的研究成果，也不包含为获得北京理工大学或其它教育机构的学位或证书所使用过的材料。与我一同工作的合作者对此研究工作所做的任何贡献均已在学位论文中作了明确的说明并表示了谢意。

特此申明。

签　　名：　　日期：

摘 要

海量互联网信息中蕴含着巨大社会价值和商业价值。传统的情感分析技术分析粒度大、准确性不高，并不能满足日益丰富的互联网用户需求。因此，如何采用一种细粒度情感分析技术使计算机能自动地对网络信息进行处理分析，给决策者提供更加详尽的结论性信息，帮助企业了解用户的消费习惯，帮助用户全面了解产品信息，逐渐成为了发展趋势。

目前，针对细粒度情感分析的研究对句法和语义信息的挖掘还不够深入，因此，本文综合情感关键句抽取、评价对象识别、话题相关的情感倾向性分析的研究基础，深度挖掘句法和语义层面的知识，针对互联网语料，对细粒度情感分析进行了深入的研究。主要研究内容和创新点包括：

1. 针对情感关键句抽取工作，提出基于 PMI 的情感词典扩展算法、主题模型与词图模型相结合的关键词抽取算法和依存模板提取算法，来分别获取情感词、关键词和句法依存特征，融合词汇语义和句法依存信息完成情感关键句的抽取工作。

2. 针对评价对象识别工作，提出一种融合词性模板、依存结构分析、语义角色标注和短语结构分析的领域词典构建方法 PDSP，并采用 Word Embedding 方法对该词典进行扩展之后，嵌入到序列标注模型条件随机场 CRF 中进行评价对象识别。

3. 在话题相关的情感倾向性分析中，首先提出了一种融合局部和全局信息的 LTIGT 算法进行关键词特征提取；然后基于 Word Embedding 和依存关系提取话题相关情感词，并对其进行 K-means 聚类，获取情感词特征；最后将以上两种特征，与传统特征一起加入 SVM 进行话题相关的情感倾向性判定，获得每个话题所对应的情感倾向，完成细粒度的情感分析。

关键词：情感分析；句法；语义；情感关键句；评价对象； SVM

ABSTRACT

There is huge social and commercial valuein the massive Internet information. The traditional coarse-grained sentiment analysis technology cannot meet the growing demand of Internet users. How to use a fine-grained sentiment analysis technologyto analyze the network information automatically, to provide more useful information for decision makers to understand the user's consumption habits, to help users understand the more information about a product, to provide the basis for real data, is becoming a new trend.

Considering researches on fine-grained sentiment analysis did not pay much attention on syntaxesand semanticx, this paperfocuses on fine-grained sentiment analysison the basis of sentiment key sentence extraction, opinion targets identification and topic-related sentiment analysis. The main research work and contributions are listed as follows:

1. For sentiment key sentence extraction, an algorithm of sentiment lexicon expansion based on PMI is proposed firstly, and then a keyword lexicon construction algorithm based on topic model and graph model is put forward, and then an extraction algorithm of dependency template is also proposed. Finally, these three sentiment, keyword and dependency features are combine for sentiment key sentence extraction using lexical semantics and syntactic dependencies.

2. For opinion targets identification, a new domain lexicon construction method, which combines POS template, dependency parsing, semantic role labeling and phrase structure analysis, is proposed. And after expansion by Word Embedding, the domain lexicon is embedded in sequence labellingmodel of CRF to identify opinion targets.

3. For the topic-related sentiment orientation analysis, this paper first proposes a feature extraction method which combines local and global information. Then, a sentiment feature extraction approach based on Word Embedding, dependency analysis and K-means clustering is proposed. Lastly, these two features with other basic features are combined for SVM to analyze the sentiment orientation of the each topic, and finally accomplish the fine-grained sentiment analysis.

Key words： Sentiment Analysis;Syntax; Semantics; Sentiment Key Sentence; Opinion Targets; SVM

目录

第 1 章 绪论

1.1 研究背景及意义

近些年来，互联网技术以势不可挡的趋势迅猛发展，越来越多的人开始使用互联网进行交流，互联网已逐渐成为人们生活中必不可少的一员。而与此同时，传统的以报纸、电视、广播等新闻媒介为主的信息传播方式，也逐渐转变为以互联网为核心的信息传播方式。如今，随着博客、微博等社交平台的兴起，网络已不仅仅是获取信息的途径，而是逐渐成为了民众发表意见、交流观点主要方式。同时，用户也不再仅仅作为信息的浏览者，同时也成为信息的发布者，互联网已经从传统的单向传播方式转换为了双向传播方式，为用户提供了一个自由开放的交流平台。例如，在民众对时事政要、娱乐新闻、消费商品、服务等各个层面的信息发表观点看法的同时，这些包含了民众情感倾向的评论资源也势必会成为政府管理、舆情监督、消费品经销商等部门制定策略的重要依据。然而互联网的海量信息纷繁复杂，如何对这些信息进行合理组织，使计算机能自动的对网络信息进行处理分析，给决策者提供有用的结论性信息，帮助企业了解用户的消费习惯，帮助用户全面了解产品信息，也变得越来越重要。情感分析 (sentiment analysis)，一般又称为观点抽取(opinion extraction)、观点挖掘(opinion mining)等，旨在分析海量评论文本中包含的情感倾向，主要涉及到信息检索、数据挖掘、机器学习、文本分类、中文分词、特征提取等相关技术。同时，情感分析的研究不仅富有学术意义，还具有较大的社会和经济价值，可以应用于政治社会科学、商业智能等多个领域。

情感分析对象，主要是互联网上的主观性文本信息，此类信息通常是用户个人对于实体、事件以及它们属性的主观性评价，往往富有个人情感、观点和态度等[1]。主观性的文本信息的重要性主要可以表现为：我们无论做任何决策都需要参考他人的意见[2]。从心理学的角度来说，这是“从众”现象的一种表现，用户更倾向于选择社会评价较好的产品；从人工智能角度来说，这是群体智慧在未来行为预测上的具体表现。卡内基梅隆大学针对十亿条微博消息进行了研究，结果发现 Twitter 上主观性挖掘的结果与公共投票的结果几乎是完全一致的[3]。惠普实验室的一项研究[4]也表明，可以通过对现有 Twitter 上的电影评论进行人工情感分类，通过机器学习的方法预测电影票房，而且这一应用也可以扩展到其他领域，比如产品的销量预测、选举结果预测等。

细粒度的情感分析，不再笼统地以句子为单位进行情感倾向判定，而是以主观性文本中所包含的话题、或评价对象为单位，判定其各自的情感倾向。具体而言，细粒度情感分析的重要性主要体现在两个方面。一方面，在电子商务领域，在线消费者们越来越重视公开的产品评论信息，这些评论信息在很大程度上影响了消费者的购买决策。因此，对这些信息的有效分析整合，或以可视化的方式呈现出来，将在很大程度上给消费者提供便利；同时，对经营者来说，也是把握市场动态的一种有效方式。另一方面，在国家安全领域，细粒度的情感分析是舆情监控中必不可少的一部分，及时分析社会舆论状态，对领导者做出正确决策提供了重要依据。总之，细粒度情感分析提供了一种更加详细、明确的情感分析方式，极大地方便了人们的生活。

近几年来在自然处理语言领域，人们对自然语言文本中的情感关键句识别、评价对象识别与话题相关的情感倾向性分析越来越感兴趣。面向文本的情感分析也是近几年来自然语言处理研究的一个热点[5-6]。尤其是自动识别文本中的各评价对象对应的情感倾向更是近年关注的焦点。在各种国际会议和评测任务中，也不乏相关的问题；颇具影响的国际评测会议 TREC 以及 NTCIR 等，都设置了细粒度情感要素抽取与倾向性分析的任务。

综上所述，细粒度的情感分析是近几年信息处理、自然语言理解领域的一个新的研究方向，已经成为学术界和工业界所关注的焦点。目前，在细粒度情感分析领域的研究还并不是很深入，摆在我们面前的仍有很多亟待解决的难题与挑战。因此，进行细粒度情感分析技术的研究，具有重要的学术价值和实用价值。

1.2 研究现状

细粒度的情感分析主要包含情感关键句抽取、评价对象识别和话题相关的情感倾向性分析，是一项富有挑战性的工作。对文本进行细粒度情感分析，将会对文本从更加细致的角度分析其蕴含的情感信息，并以易于用户理解的方式重新组织和表现，极大程度上提高了人们获取信息的效率。

本文从主要从三个层面进行细粒度的情感分析：情感关键句抽取、评价对象识别和话题相关的情感倾向性分析。

1. 情感关键句抽取

情感关键句又叫主题情感句，主要包含两个要素：主题关键词和情感关键词。主题关键词用来概括篇章的主题；情感关键词用来概括情感倾向。

目前，关于情感关键句抽取方面的研究并不多。最早的是合肥电子工程学院的孙宏纲[7]提出了中文博客主题情感句的自动抽取问题，文中首先设计了一个新颖的基于二元切分的提取算法来获取主题词，然后利用序列标注模型 CRFs 将主题句提取问题转化为中文 chunking 问题；2011 年，杨江[8]提出一种基于主题情感句的汉语评论文倾向性分析方法，该方法充分考虑了评论文的特有属性，通过一种 n 元词语匹配的方法来识别文章主题，然后通过分别计算每个句子与主题的语义相似度，并对相似度高的句子进行主客观分类，最后抽取出主题情感句；2012 年，林政，谭松波等[9]提出了一种情感关键句抽取算法，算法考虑句子的 3 类属性：情感属性、位置属性和关键词属性，并将抽取出的情感关键句分别用于有监督和半监督的情感分类，取得了不错的效果；直到 2014 年，在 COAE(Chinese Opinion Analysis Evaluation)第六届中文倾向性评测的任务一中，提出了面向新闻的情感关键句抽取课题，其要求在给定新闻集合(每篇文章已切成句子)中，判别每篇文章的情感关键句。本文就是在此评测任务的基础上，利用相同的数据集进行实验的。

总的来说，情感关键句抽取的研究尚不系统和成熟，目前还处于起步阶段。而中文语言的灵活性及表达的多样性，也使情感关键句抽取的研究相对更加困难。目前情感关键句抽取的方法大多是基于规则或基于统计的，鲜有两者结合的方法。而且在抽取过程中只进行了浅层语义分析，没有挖掘句子的深层信息。

2. 评价对象识别

评价对象识别是指在一条评论文本中，分析抽取出评论者表达情感所针对的对象，在句法结构上表现为观点句中与情感词具有一定依存关系的实体。评价对象识别是自然语言处理领域中的一个重要的研究方向，近些年来，引起了国内外学者的广泛关注。而且，在多次的中文倾向性分析评测（Chinese Opinion Analysis Evaluation, COAE）和 NTCIR（NII Test Collection for IR Systems）中，都将评价对象识别作为一个重要的评测内容。本文通过调研，发现其研究工作主要分为两大类：无监督学习方法（Unsupervised Machine Learning method）和监督学习方法（Supervised Machine Learning method）。

（1）基于非监督学习的评价对象识别方法

Hu 等[10]最早提出评价对象识别问题，使用了词频等关联规则进行评价对象识别。Li[11]采用词典的方法，分别识别情感词典中的情感词与主题词典中的主题词，抽取出

<情感词，评价对象>的二元组。Popescu 等[12]提出了一种基于 PMI 的产品属性提取方法，通过点间互信息的值来选择特征。刘鸿宇等[13]使用句法分析来获取评价对象，并结合名词剪枝算法确定最终的评价对象。但是，完全基于统计的方法需要依赖大规模的语料集，而在大规模语料集上，此模式的效率往往不足以满足要求。随后，一些研究者注意到，可以将评价对象的识别任务融入到主题模型[14-16]的求解问题中去。Mei 等[17]采用多粒度的话题模型识别产品评论文本中的评价对象，并对其中相似的评价对象进行聚类，获得了不错的效果。

（2）基于监督学习的评价对象识别方法

监督学习是最常用的机器学习方法，主要是通过训练语料训练出模型，进而对测试集进行测试，以此完成识别工作。方法有：决策树（Decision Trees）、隐马尔科夫模型（Hidden Markov Models）、最大熵（Maximum Entropy）、支持向量机（Support Vector Machines）、条件随机场（Conditional Random Fields）等。Zhuang[18]提出了一种监督学习的抽取方法，基于依存关系来识别评价对象评价词，该方法综合考虑了 WordNet 等多种知识库。实验结果表明，监督学习的方法要明显优于 Hu 等[10]的规则方法。Kessler 等[19]将评价对象的识别问题看做用机器学习中的分类问题，并进行试验，再一次证明了监督学习算法的优势。Jakob 等[20]首次将评价对象识别问题看做序列标注问题，通过选取特征，采用条件随机场模型完成了评价对象的识别任务。该方法的最大优势在于其较高的领域适应性，条件随机场模型不依赖于语料的具体领域，在不同语料上都表现出良好的性能。Li 等[21]使用浅层语义分析的方法进行评价对象的识别，通过分析句法树来进行监督学习。综上所述，机器学习方法中，特征选择是一个非常重要的任务，徐冰、王荣洋、郑敏洁等[22-24]提出了考虑词、词性等词法特征，但这种方法都只考虑了浅层的词法信息，并没有深度挖掘深层次的句法和语义特征，而这些特征的评价对象识别中是非常重要的。例如“手机真垃圾，电池一点也不给力!”，“手机”和“垃圾”、“电池”和“给力”之间是存在一定的句法关系的，若能充分考虑此部分信息，对评价对象的识别是非常有益的。因此，本文充分挖掘评价对象与评价词之间的语义联系，对传统模型进行优化，提取出覆盖面更广、准确率更高的评价对象。

3. 话题相关的情感倾向性分析

一条文本往往包含多个话题，中文微博以“＃topic＃”的形式表征话题分类，与 twitter 不同，中文微博的 topic 既可以为一个单独的词语，也可为一个短句。传统的

情感分析大多是与话题无关的，即不必考虑情感所针对的对象，而这样显然是不够精确的。例如：＃三星 galaxy s6＃三星 galaxy s6 真新没什么亮点，华为 P8 就可以秒它了，更不用说 mate8[拜拜]。该句对于话题“三星 galaxy s6”的情感判定是负向，而针对话题“华为 P8”和“mate8”的情感判定为正向。考虑到话题信息对情感判定的重要性，本文研究了话题相关的情感倾向性分析，其研究工作也可以分为无监督和有监督两种方法：

（1）基于无监督方法的情感分析：

Ku[25]利用文本中正负向情感词数目之差来判定情感倾向，其中用到了正负向情感词典；Shen[26]通过为每条微博计算一个情感指数来进行情感分析，通过构建态度词典、权重词典、否定词典、程度词典以及感叹词词典完成情感指数计算。Turney 等[27-28]的工作，主要集中在手机、银行、电影及旅游目的地相关的评论方面，通过计算正负向关联度的差值，来确定情感倾向。而关于正负向关联度，是通过 PMI 方法，计算语料集合中按一定规则提取出的短语与两个基本情感词(正向词: excellent、负向词:poor)的关联度。Quan[29]提出一种使用依存分析和短语结构分析的无监督文本挖掘方法。Wang[30]研究了基于标签的 twitter 情感分析，提出了一种全新的图模型分类算法。

（2）基于有监督方法的情感分析：

这种方法主要是使用机器学习的模型，包括朴素贝叶斯(Naive Bayes)、最大熵(Max Entropy)、支持向量机 (Support Vector Machine)等来对文本进行情感分析。Pang 等[31]主要是对朴素贝叶斯、最大熵、支持向量机方法下电影评论情感极性二分类实验效果进行了比对，实验通过对预处理过的语料进行特征提取，提取出包含一元词特征(unigram)、二元词特征 (bigram)、词性标注、词的位置信息等特征，并将这些特征融入机器学习模型中进行实验，实验结果表明，支持向量机在情感分析二分类实验中表现出良好的性能，且在选用一元词特征时取得了最好的实验效果，准确率为 83%。Sajib[32]提出了一种半监督学习方法，通过生成式和即时学习方法进行情感分类。Li 等[33]对于评论数据，首先提出了用于情感二分类的 Dependency-Sentiment-LDA 模型，它在情感分类的时候不仅考虑了情感词所表达的话题语境，而且还考虑了情感词的局部依赖关系。然后，文中还进一步研究了情感多分类问题，通过调研提出了一种基于 Tensor 的评论分值预测方法，以此完成情感多分类任务。Niblack[34]提出了一种 SA 模型，采用情感词典和数据库挖掘给定话题的引用资源及相应情感。

从目前看来，关于情感关键句抽取的研究刚刚起步，研究方案和方法都不够成熟和系统化，尚需进一步地研究。评价对象识别的研究相对来说，已经比较成熟和全面，文中主要从有监督和无监督两个层面进行论述，从以往工作来看，无监督方法的效果要低于有监督方法，而且对有监督方法的特征选择部分的研究主要集中在词法层面，这显然是远远不够的。因此本文基于此研究现状，着眼于句法和语义层面的信息，采用有监督与无监督方法相结合的方法进行评价对象识别的研究。情感倾向性分析方面，以往工作大多是不针对话题的，而是为给定句子利用一定方法直接判定出一个情感倾向。这种粗粒度的情感分析，虽然在效果上取得了一定进步，但在实际应用中的表现并不乐观。因为实际应用中需要更加细粒度的情感分析，因此，本文对话题相关的情感倾向性分析展开研究，以不同话题为中心，分别确定其所属的情感倾向。情感分析的研究已有较长的时间和较为成熟的方法，但是，具体在细粒度的情感分析方面，对于如何结合多种特定领域文本特征的属性、并综合利用句法语义层面信息的研究仍然有待深入。

1.3 论文主要研究内容

本课题研究基于句法和语义的话题细粒度情感分析技术，首先，采用规则与机器学习相结合的方法，采用自然语言处理基础技术进行分词、词性标注、句法分析、依存分析、语义角色标注，在大规模的语料信息中，抽取出篇章情感关键句，以便获取更加准确的、与话题相关的情感信息；然后采用基于序列标注模型 CRF 的方法，融合规则方法构建的领域评价对象词典，与传统的词、词性、依存信息一起，抽取出每个句子对应的评价对象；最后以每条文本中的评价对象为话题展开话题相关的情感倾向性分析，采用分类方法获得每个评价对象对应的情感倾向，最终得到每个句子对应的<评价对象、情感倾向>的二元组，至此完成细粒度的情感分析。主要研究内容包括：

1. 本文将情感关键句的抽取问题转化为“是否为情感关键句”的二分类问题，在特征选择时，首先采用 PMI 算法对基础情感词典扩充，获取情感词特征；另外采用 LDA 主题模型与 TextRank 词图模型相结合的方法获得关键词特征，然后提出一个面向情感关键句的依存模板提取算法获取依存信息，最后融合情感词特征、关键词特征和句法依存特征，使用 SVM 完成情感关键句抽取工作。

2. 在词图模型构建中，考虑到一个结点对其相邻结点集的影响力主要由位置重要性的影响力（position）、覆盖重要性的影响力（coverage）、频率重要性的影响力（frequency）和共现重要性（co-occurrence）的影响力共同决定，本文还提出一种新的图模型混合加权方法 PCFO；此外，考虑到传统的词图模型没有考虑主题分布，随机游走模型易产生局部最优的情况，本文对文本使用 LDA 模型进行建模，采用每一个词属于特定主题的概率作为该主题下，这个词的随机跳转的概率，然后选取 TextRank 得分较高的词作为关键词特征。

3. 针对评价对象识别问题，提出一种融合词性模板（POS template）、依存结构分析（dependency parsing）、语义角色标注（semantic role labeling）和短语结构分析（phrase structure analysis）的领域词典构建方法，并将其嵌入到序列标注 CRF 模型特征中进行评价对象识别。并采用 Word Embedding 的方法，对上一步中所构建的候选评价对象词典进行扩充，用以提高评价对象识别的召回率。

4. 在话题相关的情感倾向性分析中，依据 TFIDF 算法思想，提出了一个融合主题、位置、共现信息的 LTIGT 算法，该算法分别在局部（句子）和全局（每个 topic 下的所有句子）下构建图模型，按照改进的 TextRank 打分策略，给图中每一个词赋予两个不同的得分：局部下的 TextRank 值（LT）和全局下的 TextRank 值（GT），提取局部重要性高、全局重要性低的词语作为特征词。然后，基于 Word Embedding 对主题词进行扩展，进而根据依存关系提取出评价词特征，与 LTIGT、传统的词法特征一起加入 SVM 进行情感倾向性判定，进而完成细粒度的情感分析。

1.4 论文组织结构

论文共分六章。

第 1 章是绪论，介绍了本文的研究背景，国内外的研究现状、研究内容和主要共工作，以及论文的主要结构。

第 2 章介绍情感关键句抽取的相关背景和算法设计与实现，并以新闻性文章为例，阐述实验流程与结果。

第 3 章介绍评价对象识别的相关背景和算法设计与实现。首先分析评价对象识别的主要方法，然后介绍本文的提取算法，最后通过实验分析算法效果。

第 4 章介绍话题相关的情感倾向性分析的算法设计与实现。首先提出一种考虑局部信息和全局信息的 LTIGT 算法，进而提出一种基于 Word Embedding 的话题词扩展方法，最后将其与传统的 TFDF 及其他词典一起作为特征，采用 SVM 分类器，对特征进行选择实验。

第 5 章介绍基于句法和语义的话题细粒度情感分析系统总体设计说明。

第 6 章总结论文的内容和主要工作，客观分析该系统的优缺点并给出前景展望。

第 2 章 基于词汇语义和句法依存的情感关键句抽取

2.1 概述

随着我国互联网事业的迅速发展，网络作为一种新型媒体逐渐步入人们的生活，成为人们获取和传播信息的主要途径。网络在给人类生活提供了便利的同时，伴随产生的是互联网上的海量信息。这些信息大多是人们发表的意见或看法的集合，有针对性地对这些信息进行分析处理，将极大程度上提高政府、企业等制定决策的有效性。因此，有关网络舆情监测和分析的研究，也引起研究人员的高度重视。但在如今这个大数据时代，海量信息层出不穷，但同时处理这么多信息无疑是件费时费力的事情。因此，我们需要一种网络信息处理技术来帮助我们自动从海量信息中抽取出与主题相关的，并能表述一定情感倾向的句子，即情感关键句，这是一项既有学术意义又有实用意义的研究课题。

情感关键句又叫主题情感句，主要包含两个要素：主题关键词和情感关键词。主题关键词用来概括篇章的主题；情感关键词用来概括情感倾向。目前，关于情感关键句抽取方面的研究并不多。有关情感关键句抽取的研究尚不系统和成熟，目前还处于起步阶段。而中文语言的灵活性及表达的多样性，也使情感关键句抽取的研究相对更加困难。目前情感关键句抽取的方法大多是基于规则或基于统计的，鲜有两者结合的方法。而且在抽取过程中只进行了浅层语义分析，没有挖掘句子的深层信息。因此，本文提出将情感关键句的抽取问题归类为机器学习问题，将情感关键句的抽取看作“是否是情感关键句”的二分类问题，挖掘词汇语义和句法依存特征提取并构造分类器将句子划分为情感关键句与非情感关键句两类。

2.2 面向情感关键句抽取的词汇语义分析

词汇语义是语言的语言的核心组成成分。词汇是表达语义所必需的基本材料，而语义则是词汇组合所要传达的主旨思想。洪堡特先生[35]说，汉语词的语法价值，与词的实体意义、在句中的位置和语境意义息息相关。此外，马庆株先生[36]认为，语义对语法有决定作用，能指导词语之间的搭配规则。由此可知，在汉语研究中，更要注重词汇语义的研究，理解汉语必须从词汇本身的意义出发，深入剖析词汇语义。

由于情感词和主题词是情感关键句的两个重要组成成分，因此我们通过情感词典扩充和关键词词典构建来获取词汇语义信息。

2.2.1 情感词典构建

情感词是能表达一定情感倾向的词语，词性上可以是名词、动词、形容词、副词等，当然也可以是习惯用语或者短语等。情感词典是情感词的集合，即包含情感色彩的词语集合。一般来说，文本内容表达的情感倾向主要通过情感词来体现，而情感词典的全面性和领域相关性将直接影响情感分析的效果，所以情感词典在情感分析中占据着十分重要的地位。始于语义词典构建[37-38]，构建一个覆盖面大、精确率高的情感词典在近些年受到人们的普遍关注。

目前，文本情感分析领域还没有一部完整通用的情感词典，情感词典的覆盖率和领域性也成为情感分析任务中的瓶颈问题。目前用的比较多的是知网（Hownet[1]）提供的情感词典，该词典最大的特点在于作者已经根据词语情感倾向和类别将其划分为6类，分别是“正面评价词”、“负面评价词”、“正面情感词”、“负面情感词”、“主张词”以及“程度级别词”。另外一部比较常见的情感词典是由台湾大学整理和发布的NTUSD[2]。本文采用知网的正负面情感词、评价词加上简体中文的 NTUSD 构成基础情感词典。

由于基础情感词典中包含的情感词是有限的，而中文的表述千变万化，因此，仅靠基础情感词典来进行情感分析是远远不够的。为提高情感分析的准确率和情感词典的覆盖率，必须对情感词典进行扩充。扩展情感词典的方法主要有基于语义相似度[39-40]和基于同义词的方法[41-42]。考虑到互信息方法的准确性与领域适应性，本文采用互信息的方法对基础情感词典进行扩充，构建出一个领域相关的情感词典。

互信息[43]（Mutual Infomation）表示的是两个随机变量之间的关联性，是非常重要的信息量度。在实际情况中，应用最为广泛的是点间互信息[44]（PMI），主要用于计算词语间的语义相似度，通过语义相似程度对情感词典进行扩充。其基本思想就是统计两个词在同一个句子中的出现概率，如果出现概率越大，其相关性就越强，联系越紧密。

两个词语w_1和w_2的 PMI 值计算公式如下：

[1]http://www.keenage.com/html/c_index.html
[2]http://www.datatang.com/data/11837

$$\mathrm{PMI}(w_1, w_2) = \log(\frac{P(w_1 \& w_2)}{P(w_1)P(w_2)}) \quad (2.1)$$

式中$\mathrm{P}(w_1\&w_2)$)表示 w_1和 w_2在同一个句子中共同出现的概率，$\mathrm{P}(w_1)$和$\mathrm{P}(w_2)$分别表示两个词语单独出现的概率。其中两个词语的共现概率和单个词语的出现概率，都可以通过对语料集合的统计得到。

$$\mathrm{P}(w_1\&w_2) = \mathrm{num}_{\mathrm{sen}}(w_1\&w_2)/\mathrm{N} \quad (2.2)$$

$$\mathrm{P}(w_1) = \mathrm{num}_{\mathrm{sen}}(w_1)/\mathrm{N} \quad (2.3)$$

$$\mathrm{P}(w_2) = \mathrm{num}_{\mathrm{sen}}(w_2)/\mathrm{N} \quad (2.4)$$

其中，$\mathrm{num}_{\mathrm{sen}}(w_1\&w_2)$表示即出现$w_1$又出现$w_2$的句子数，$\mathrm{num}_{\mathrm{sen}}(w_1)$表示出现$w_1$的句子数，$\mathrm{num}_{\mathrm{sen}}(w_2)$表示出现 w_2的句子数，N 表示语料集合的全部句子数。

基于点间互信息 PMI 算法的具体计算过程如下：

（1） 首先对语料进行预处理操作，包括分词、词性标注、去除停用词等。

（2） 然后在预处理过的语料集合中，选择种子词集合，即按照词性筛选出名词、动词、形容词，并将其加入种子词集合，作为候选词。

（3） 接着分别计算上文构建的的基础情感词典中每个词与这些种子词之间的点间互信息，对于情感词典中的每个词，选取至多前 5 个与之互信息最高的种子词作为扩展，同时过滤掉$\mathrm{P}(w_1\&w_2)$、$\mathrm{P}(w_1)$、$\mathrm{P}(w_2)$为零的情况。

（4） 最后将扩展所得词语加入基础情感词典，生成最终的领域相关的情感词典。

通过点间互信息 PMI 算法可以有效的扩展情感词典，不仅弥补了基础情感词典数量不足覆盖率不高的缺点，又在一定程度上丰富了与领域相关的情感词，使之具有更强的领域适应性。

2.2.2 关键词词典构建

关键词是表达文章主旨的最小单元，是文章主题的集中表现。所谓关键词词典 KL（Keywords Lexicon）构建，就是从一篇给定的文本中自动抽取出若干有意义的词语或词组，抽取方法主要分为两种：其中一种是通过机器学习模型算法[45-46]实现；另一种是通过挖掘词语间关系，直接进行抽取。第二种方法由其高效率、无需监督等特点被广泛采纳。关于无监督关键词抽取方法的研究，主流方法可归纳为三种：基于 TFIDF 统计特征、基于主题模型[47-48]和基于词图模型[49-51]的关键词抽取方法。考虑到 TFIDF 统计特征的抽取方法易忽略文档中重要低频词和文档内部的主题分布等语义

特征的缺点，以 TextRank 为代表的文本词图构建排序的方法只关注于单篇文档，不需要事先对多篇文档进行训练和学习，其简单高效的特点使其得到了广泛应用。但传统词图模型构建方法在权值的设定和跳转概率的选择上并没有做深入研究，因此，本文首先基于传统的词图模型构建算法，提出了一种新的加权方法PCFO，然后采用LDA和 TextRank 相结合的词图模型进行关键词抽取，构建关键词词典。接下来将从两个方面来介绍词图模型的构建。

（1）PCFO：一种图模型混合加权方法

我们知道，节点之间的权重值在一定程度上体现了两节点联系的紧密度，即两个词语之间的关联程度。通过观察分析，发现影响结点之间的权重的因素主要有以下四种:

a) 位置重要性：出现在标题、摘要、段落的首句中词语，更有可能成为关键词。
b) 覆盖重要性：指向某词语的不同词语数量越多，被指向的词语越重要，即拥有投票数目越高的词语，其重要性越高，越有可能成为关键词。
c) 频率重要性：出现频率较高的词语，越有可能成为关键词。
d) 共现重要性：如果两个候选关键词在指定窗口大小中共现次数越多，则说明两个词之间的联系越紧密。

对于图中的任一结点v来说，其重要性得分由其相邻结点的贡献组成，而其本身的得分也将被转移到相邻结点。所以，一个节点对其相邻节点的影响力主要可以分解为四个组成部分：位置重要性的影响力（position）、覆盖重要性的影响力（coverage）、频率重要性的影响力（frequency）和共现重要性（co-occurrence）的影响力。因此，本文提出了一种新的图模型混合加权方法PCFO。

对于任意两个结点v_i和v_j，结点v_i对v_j的影响力通过其有向边e=<v_i，v_j>传递，边的权重w_{ij}决定了v_j最终所获得v_i部分的分值大小。令w_{ij}表示结点v_i和v_j的整体影响力权重，α，β，γ，δ分别表示这对相邻节点的四类影响力所占的比重，且α+β+γ+δ=1。则两节点之间的权值可以设为以下形式：

$$w_{ij} = \alpha w_{pos}(v_i,\ v_j) + \beta w_{cov}(v_i,\ v_j) + \gamma w_{freq}(v_i,\ v_j) + \delta w_{co-occur}(v_i,\ v_j) \quad (2.5)$$

1) $w_{pos}(v_i,\ v_j)$表示节点v_i的位置影响力传递到v_j的权重，计算公式如下：

$$w_{pos}(v_i,\ v_j) = \frac{\mathrm{P}(v_j)}{\sum_{v_t \in \mathrm{Out}(v_i)} P(v_t)} \quad (2.6)$$

其中，$\mathrm{P}(v_j)$表示节点v_j的位置重要性得分，根据不同的情况可以设置不同的打分

策略，本文只考虑了标题信息对词语重要性的影响，认为只要是在标题中出现过的词语，则给予更高的得分。具体赋值方式如下：

$$\mathrm{P(v)}=\begin{cases}\lambda，\text{v 所对应的词语在标题中出现}\\ 1，\text{v 所对应的词语在标题中不出现}\end{cases} \tag{2.7}$$

其中，λ是一个比 1 大的数字，实验中，经过验证选择λ = 1.5.

2) $w_{cov}(v_i，v_j)$表示节点v_i的覆盖影响力传递到v_j的权重，计算公式如下：

$$w_{cov}(v_i，v_j)=\frac{1}{|\mathrm{Out}(v_i)|} \tag{2.8}$$

其中，$|\mathrm{Out}(\mathrm{v_i})|$表示节点$\mathrm{v_i}$的出度，即由$\mathrm{v_i}$出发所指向的节点的数目；此公式说明节点$\mathrm{v_i}$的覆盖影响力将被均匀的传递到相邻节点。

3) $w_{freq}(v_i，v_j)$表示节点v_i的频度影响力传递到v_j的权重，计算公式如下：

$$w_{\mathrm{freq}}(v_i，v_j)=\frac{\mathrm{f}(v_j)}{\sum_{v_t\in\mathrm{Out}(v_i)}f(v_t)} \tag{2.9}$$

其中，$\mathrm{f(v_j)}$表示节点$\mathrm{v_j}$所代表的的词语在文章中出现的次数，以上公式则体现出出现次数较高的词语将从连接节点处获得更高的影响力权重。

4) $w_{co-occur}(v_i，v_j)$表示节点v_i的共现影响力传递到v_j的权重，计算公式如下：

$$w_{co-occur}(v_i,v_j)=\frac{\mathrm{Co}(v_i,v_j)}{\sum_{v_t\in\mathrm{Out}(v_i)}\mathrm{Co}(v_i,v_t)} \tag{2.10}$$

其中，$\mathrm{Co(v_i,\ v_j)}$表示节点$\mathrm{v_i,v_j}$所代表的的词语在一定窗口内共现的次数，以上公式则体现出共现次数较高的词语将从连接节点处获得更高的影响力权重，也即这两个词语之间联系更加紧密。

（2）基于主题分布的跳转概率选择

TextRank 模拟 PageRank，以词语单位对文本构建图模型，利用投票机制获取每一个词语的得分进行排序。但传统的词图模型中每个节点是以相等的概率随机跳转的，而且没有考虑文章的主题分布，这种方法极易产生局部最优的情况。因此，本文在随机游走的过程中充分考虑文章的主题分布，分别在每一个主题下构建词图模型，并给每一个词赋予一个在不同主题下的得分，最后按照一定的加权策略，为每一个确定一个最终的得分。

目前在主题模型方面，LDA模型的运用最为广泛。LDA模型假设一个文本由多个隐含主题随机组成，在LDA 模型中也做了“bag ofwords”假设，即单纯地把文章看成词汇的堆积，不考虑任何句法和语法关系。因此，给定一个文档d，文档中的每一

个词w的生成过程可以看成先从文档d的文档-主题分布$\theta^{(d)}$中抽样出一个主题，然后再在主题-词的分布$\varphi^{(z)}$中抽样出一个词；其中$\theta^{(d)}$和$\varphi^{(z)}$是分别从共轭的狄利克雷因子α和β得到的。设主题个数为k，以上过程用以下公式来表示如下：

$$P(w|d;\alpha,\beta)=\sum_{t=1}^{k}P(w|z_t;\beta)*P(z_t|d;\alpha) \quad (2.11)$$

使用LDA我们可以求出P(z|d)和P(w|z)，即文档-主题分布和主题-词的分布；考虑到传统TextRank以相等概率随机跳转，易产生局部最优的缺点，因此，在随机游走的过程中考虑文章的主题分布，把每一个词属于特定主题的概率看做在该主题下，这个词的随机跳转的概率，即把每一个词属于特定主题的概率P(z|w)作为该主题下词的随机跳转概率，$P(z_t|v_i)=P(z|w)$，P(z|w)由LDA模型求得。因此，在考虑主题分布的词图模型中，按照打分策略公式（2.12）给每一个词赋予一个不同主题下的得分：

$$R_{z_t}(v_i)=\lambda\sum_{j:v_j\to v_i}\frac{w(v_j,v_i)}{Out(v_j)}R_{z_t}(v_j)+(1-\lambda)P(z_t|v_i) \quad (2.12)$$

其中，λ是一个阻尼因子，表示每个节点都有$1-\lambda$的概率随机跳转到图中的其他节点。$w(v_j,v_i)$表示节点v_j到v_i的边的权值（由上文PCFO方法求得），$Out(v_j)$表示由v_j出发的所有边的权值之和，$P(z_t|v_i)$表示v_i节点所代表的词属于当前主题的概率，可由LDA模型得到。$R_{z_t}(v_i)$表示节点v_i在主题z_t下的得分，迭代上述式子，直到收敛。

最后按照公式(2.13)对所有主题下的得分加权求和得到一个最终的得分，排序取排名较高的节点作为最终的关键词提取结果。

$$R(v_i)=\sum_{t=1}^{k}R_{z_t}(v_i)\times P(z_t|d) \quad (2.13)$$

其中，$R_{z_t}(v_i)$表示节点v_i在主题z_t下的得分，$P(z_t|d)$表示该篇文档属于主题z_t的概率，$R(v_i)$表示节点v_i的最终得分。

这种方法由于将主题分布考虑在内，将文档中每个词属于某个主题的概率作为跳转概率，并综合考虑了四种影响两节点间权值的因素，运行有偏的词图模型算法，使提取出的关键词能够更好地覆盖文章主题，具备更高的准确率。

因此，关键词提取算法如下：

输入：语料集合 T。

输出：每篇文档对应的关键词词典 KL。

Step 1：首先对语料进行预处理，即分词、词性标注和去除停用词；

Step 2：对语料集合中的每一篇文档，使用 LDA 模型，并利用公式（2.11）求出每个词语属于每个主题的概率$P(z_t|v_i)$和每篇文档下每个主题的概率$\mathrm{P(z_t|d)}$；

Step 3：构建词图模型：选择名词，形容词作为候选词，并以这些词为节点，分别在每一个主题下构建图模型$\mathrm{G=(V,E)}$，节点集合$\mathrm{V=\{v_1,v_2,v_3\ldots\ldots v_n\}}$，其中连接两节点的边$\mathrm{(v_i,v_j)\in E}$。确定两个节点之间是否存在边以及边的方向的方法是，在一个大小为 window 的滑动窗口内，分别按照顺序从第一个词指向窗口内的其他词语。并按上文所提 PCFO 方法为每一条边设置权重。

Step 4：图模型建立完毕之后，利用公式（2.12）代计算每一个节点在特定主题下的得分。由于每个节点得分的初值对最后的排序结果没有影响[51]，所以，本文全部设置为 1。

Step 5：求得每个主题下每个节点的得分之后，最后按照公式（2.13）综合计算每一个节点的最终得分。

Step 6：选取排名靠前的节点，将节点所代表的候选关键词与此节点的最终得分一块加入关键词词典 KL，生成对应于本文章的最终的关键词词典。

Step 6：若语料集合中文档已全部遍历，则停止。否则，重复 Step 2。

2.3 面向情感关键句抽取的句法依存分析

陆俭明先生[52]指出，句法在汉语研究中占据十分重要的地位，汉语研究需要重视和加强词语间相互制约关系和依存关系的研究，因为这些关系对语义表达起着决定性作用。而剖析词语之间制约关系与句法结构，最常用的方法就是依存句法分析。依存句法分析主要通过分析语言单位的各个组成部分，及其之间的相互关系来表达句子中的结构信息。依存关系文法将每个句子的谓语动词作为一句话的中心，认为它可以支配其他成分而它本身是不受其他任何成分的制约，其他所有被支配的成分都附属于其支配者并存在某种依存关系[53]。依存分析结果一般由三部分组成：分词结果、依存弧和依存关系。依存弧表示的是词语之间的相互依赖关系；依存关系指的是词语之间的依存关系类型，如主谓关系、动宾关系等。

例如句子：即将到来的中美领导人会见，无疑将为未来数年的中美关系奠下第一块基石。对其进行依存分析结果如下图所示：

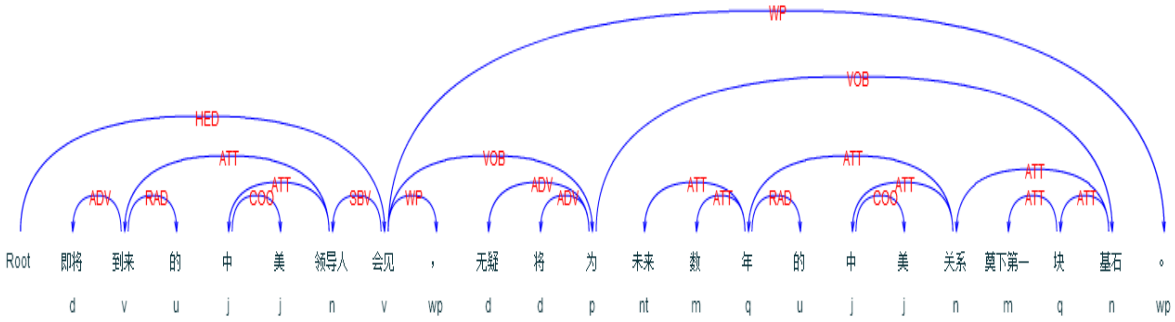

图 2.1 依存关系分析结果示意图

对于情感关键句，本文希望根据词汇与词汇之间的依存关系，挖掘它们隐藏的结构信息。为此我们首先使用哈工大 LTP 的依存句法分析模块来分析待识别的句子，找到该句的中心词 HED，然后与上文扩展后的领域相关的情感词典和 hownet 主张词一起作为中心特征词，并以这些词为起点，对附属或依存于该词的词汇进行关系提取，然后统计它们相互之间的关系，根据统计数据提取出符合要求的依存关系模板。算法步骤如下：

输入：经过预处理得到的语料 T。

输出：依存知识库 D。

Step 1：依次遍历一条语句的所有词语，如果该词语在上文构建的领域相关情感词典或 hownet 主张词中出现，或者依存分析结果中 relate=“HED”，则把它作为 CoreWord。

Step 2：将与 CoreWord 有依存关系的词语存入依存词的集合 dpWords。

Step 3：遍历 dpWords 中的每个词与 CoreWord 的关系，如果其关系为 COO，将它作为 CoreWord 重复 Step2；而若其依存关系为 WP，则将其从 dpWords 中删除。

Step 4：将情感关键句中的所有包括 CoreWord、dpWords，以及 dpWords 中的每一个词与 FatherNode 相互之间的依存关系存入情感关键句模板集合中，并且不改变其出现顺序，如“领导人(SBV)会见(HED)为(VOB)”。

Step 5：从 Step4 中得出的模板集合中按“一个前面的词与中心词的关系+中心词+一个后面的词与中心词的关系”、“一个前面的词与中心词的关系+中心词”、“中心词+一个后面的词与中心词的关系”三种方式作为提取方法，对于同一个中心词去最长模板，并统计其各自在情感关键句、非情感关键句中出现的概率。

Step 6：将候选模板集合中在情感关键句中出现的概率大于在非情感关键句中出

现概率的模板提取出来，与它在非情感关键句中的出现概率一起加入依存关系知识库，形成最终的依存知识库。

2.4 基于 SVM 的情感关键句抽取过程

上文 2.2 和 2.3 节分别分析了情感关键句的词汇语义信息和句法依存信息，在这些工作的基础之上，本节主要研究如何将此信息进行特征集成，融入到机器学习模型支持向量机 SVM 中进行分类，进而完成情感关键句的抽取。

2.4.1SVM 简介

支持向量机 SVM[54]（Support Vector Machine）是一种有监督的机器学习算法，其根据结构化风险最小化归纳原则(Structural RiskMinimization Inductive principle)，通过不断的学习获得期望风险的最小阈值，即从训练集中选出分类能力最好的一组特征函数。SVM 方法通过对有限的样本信息，综合考虑分类能力和模型复杂度，在两者之间寻求最佳折中，提高模型的适用能力。支持向量机 SVM 将分类问题，映射成空间内的点值划分，将其转化成为了一个“求二次函数最优值”的问题，这与其他机器学习方法比如神经网络相比，SVM 趋向于寻求全局最优解，而不是神经网络中的局部最优解。另外，为解决空间向量不可分问题，SVM 通过核函数将低维空间的向量映射到高维，通过对原空间中的非线性判别函数到高维空间中的线性判别函数的转换，巧妙地解决了该问题，使得 SVM 在不增添计算复杂度的情况下，具备了更强的普适性。因此，可以将 SVM 算法简单的概括为，在数据点集合中寻求一个满足分类要求的最优超平面问题。

对于一个二分类问题，如图 2.2 所示，设数据集 A，B，若两者可分，即存在 ω^T, b 使

$$(\omega^T \cdot x_i) + b > 0, \forall x_i \in A \quad (2.14)$$

$$(\omega^T \cdot x_i) + b < 0, \forall x_i \in B \quad (2.15)$$

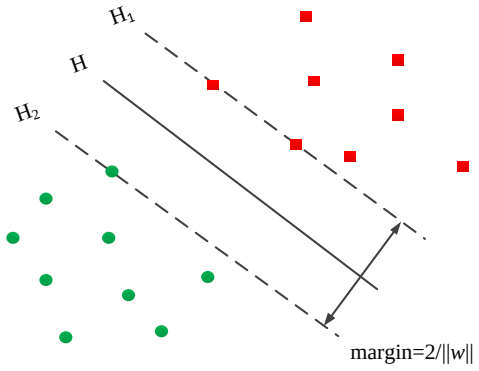

图 2.2 SVM 分类示意图

图 2.2 中，H 指的是分类超平面，H_1、H_2 的含义是指与分类超平面平行且在不同类中与分类超平面的距离最近的样本的平面，它们之间的距离叫做分类间隔(margin)。那么对于若求分类间隔最大的超平面（即 $1/\|\omega\|$最大）等价于二次规划问题：

$$\min \frac{1}{2}\|\omega\|^2, subject\ to\ y_i[(\omega^T \cdot x_i)+b] \geq 1 \qquad (2.16)$$

而对于线性不可分的数据集，除了引入核函数之外，还要表征对错误的容忍程度，因此，在这种情况下，SVM 中引入了“松弛变量 ξ_i”的概念，表示在超平面的确定中，可以容忍有少数点被误分。因此，其约束函数可改写为（2.17）所示：

$$y_i(\omega^T \cdot x_i + b) \geq 1 - \xi_i \qquad (2.17)$$

若 $\xi_i=0$， x_i 则被正确分类；若 $0 < \xi < 1$，x_i 也被正确分类，但它落在边缘中；若 $\xi \geq 1$，x_i 被错误的分类。相应的，在目标函数中也引入惩罚因子，得到带有惩罚因子的 SVM 目标函数如公式（2.18）：

$$\min \frac{1}{2}\|\omega\|^2 + C\sum_{i=1}^{n}\xi_i, subject\ to\ y_i(\omega^T \cdot x_i + b) \geq 1 - \xi_i, \xi_i \geq 0, i = 1,2,...,n \qquad (2.18)$$

SVM 利用最大间隔的思想实现结构风险最小化原则，并提出核函数的思想解决了低维空间不可分问题，在现有的分类问题上表现出了良好的性能。本文采用由台湾大学林智仁教授研发的 libSVM 进行实验。

2.4.2 特征选择

本文提出四种 SVM 的候选特征：情感词特征，关键词特征，依存模板特征和位置特征。针对情感词、关键词和依存模板特征，分别选取领域相关的情感词典 DEL、关键词词典 KL 和依存知识库 DKB 中排名较高的前 n 位的得分，与该类特征的维数一起作为相对应部分的特征。此外，由于中文文章的文章结构不外乎“总-分-总”、“分-总”、“总-分”、“分-分-分”，而上述第四种形式是非常少见的，因此有关作者主观情感及看法的句子，即情感关键句，一般都出现在文章的开头或结尾。因此，针对位置特征[9]，实验选择两种打分函数进行实验，第一种采用改进后的正态分布 Normal 形式，如下：

$$\mathrm{score}_{\mathrm{sen}}\big(\mathrm{pos}(\mathrm{sen})\big)=\frac{1}{\sqrt{2\pi}\sigma}\left(1-\mathrm{e}^{-\frac{(\mathrm{pos}(\mathrm{sen})-\mu)^2}{2\sigma^2}}\right) \tag{2.19}$$

其中$\mu=\frac{\mathrm{n}}{2}$，$\mathrm{pos}(\mathrm{sen})$表示句子在文章中的位置。

第二种采用抛物线的形式，打分函数如下：

$$\mathrm{score}_{\mathrm{sen}}\big(\mathrm{pos}(\mathrm{sen})\big)=\mathrm{a}\times\mathrm{pos}(\mathrm{sen})^2+\mathrm{b}\times\mathrm{pos}(\mathrm{sen})+\mathrm{c} \tag{2.20}$$

其中，$-\frac{\mathrm{b}}{2\mathrm{a}}=\frac{\mathrm{n}}{2}$，$\mathrm{a}>0$，$\mathrm{b}<0$，$\mathrm{pos}(\mathrm{sen})$表示句子在文章中的位置。

2.4.3 系统框架

综上所述，本文情感关键句识别的主要流程如图 2.3 所示：

1. 预处理：分词、词性标注、依存分析、语义角色标注[3]、去除停用词；
2. 分别对句子进行词汇语义和句法依存分析：扩展情感词典、构建关键词词典，构建依存知识库；
3. 根据扩展后的情感词典和构建的关键词词典，按规则的方法对句子进行过滤，获取含有情感词和关键词的候选情感关键句；
4. 生成候选情感关键句的 4 种特征：情感词、关键词、依存模板和位置特征；
5. 使用 SVM 进行分类，判别一个句子是否是情感关键句。

[3]http://www.ltp-cloud.com/

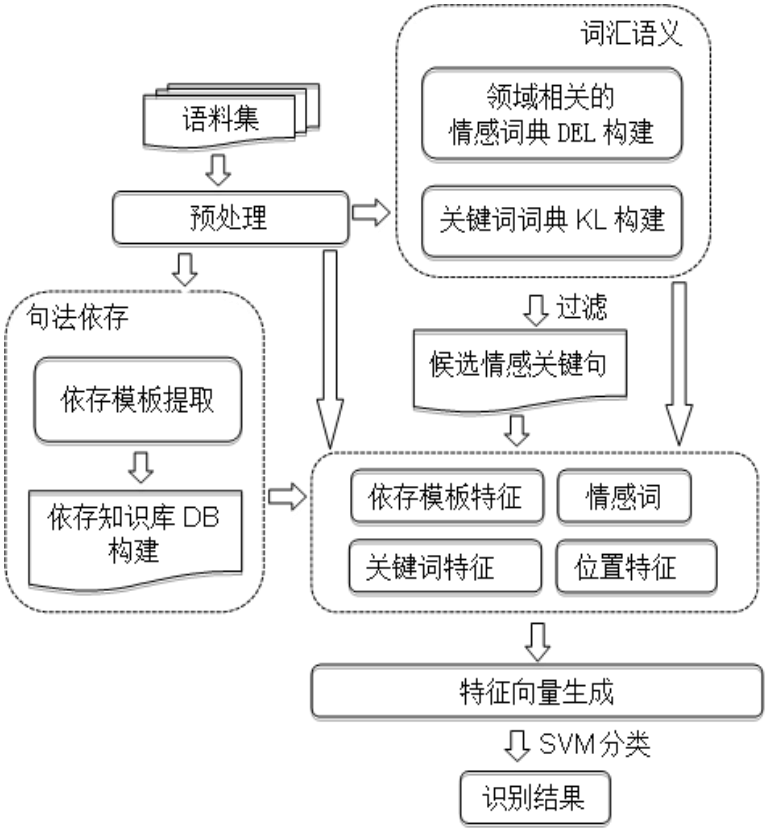

图 2.3 情感关键句识别的方法流程

2.5 实验结果与分析

本节对上文提出的基于词汇语义和句法依存的情感关键句抽取方法进行实验和评估。实验过程是将实验数据集经过上述算法处理后得到其情感关键句，然后，比对其预测结果与标注结果的异同。

2.5.1 实验数据集与评测指标

目前对于情感关键句的研究还不多，没有较多公开的评测数据集。本文从 COAE2014 提供的来自凤凰新闻、新浪博客、搜狐新闻、网易论坛、网易博客、新浪论坛、搜狐博客、腾讯新闻和腾讯博客的新闻性文章入手，这些语料数据以篇章为单位，篇章内部已经分好句子。数据集共包含 1994 篇文档，经领域相关情感词典和关键词表过滤之后，共有 38797 个句子，其中情感关键句 5019 句，非情感关键句 33778 句。使用 SVM 进行分类的时候，使用 4047 句情感关键句，以及非情感关键句 5000 句作为训练集；972 句情感关键句，以及 7325 句非情感关键句进行测试。

目前，关于文本方面实现效果的评估，主要采用的评测指标是准确率（Precision）、召回率(Recall)和 F 值（F-measure）。其主要原理是对预测结果和标注结果进行比对，按一定公式计算出三个指标。对任一实验结果，可以统计得出下表的 4 个值：

表 2.1 分类结果判定表格

	标注属于某类别	标注不属于某类别
预测属于某类别	A	B
预测不属于某类别	C	D

其中，A 表示分类结果中，预测出属于某类别且标注中也属于某类别，即分类器能正确判断属于某类别的个数；B 表示预测出属于某类别但标注中不属于某类别，即分类器错误判断属于某类别的个数；C 表示预测出不属于某类别但标注中属于某类别，即分类器错误判断不属于某类别的个数；D 表示分类结果中，预测出不属于某类别且标注中也不属于某类别，即分类器能正确判断不属于某类别的个数。

由以上指标可得，准确率（Precision）的计算公式（2.21）：

$$\mathrm{P} = \frac{A}{A+B} \times 100\% \tag{2.21}$$

召回率（Recall）的计算公式（2.22）：

$$\mathrm{R} = \frac{A}{A+C} \times 100\% \tag{2.22}$$

F1 值（F-measure）的计算公式（2.23）：

$$\mathrm{F} = \frac{2 \times P \times R}{P+R} \tag{2.23}$$

本节主要采用以上三种指标对实验结果进行评估，其实验效果如下文所述。

2.5.2 情感词典覆盖率实验

情感词典作为情感关键句的重要特征，为了验证其完整性和适应性，我们分别验证了情感词典扩充前后，情感词典在情感关键句中的覆盖率。本文实验数据集在尚未经过规则过滤情况下包含 5119 句情感关键句与 43699 句非情感关键句，在此覆盖率验证实验中，我们对情感关键句进行分析，分别统计情感词典扩充前后，出现情感词的情感关键句的个数，并依次计算扩展前后情感词典的覆盖率。

通过实验发现，在 5119 句情感关键句中，对于扩展前的情感词典，出现情感词的情感关键句有 3721 句，覆盖率为 72%；而对于扩展之后的情感词典，出现情感词

的情感关键句有 5019 句，覆盖率高达 98%，由此可见，情感词典的扩充在一定程度上大大提高了情感词的覆盖率，在一定程度上弥补了基础情感词典与领域不相关的不足。但是，仅仅依赖情感词匹配的方法远远不能达到目的，因为出现情感词的句子并不一定都是情感关键句，因此，情感词典要和其他方法相互配合才能达到更好抽取情感关键句的目的。

2.5.3 不同关键词词典构建方法的比较

关键词信息是情感关键句的一个重要元素，因此关键词提取效果将直接影响情感关键句抽取的准确率。因此，本文主要采用了四种关键词提取方法，一种是最基本的 TFIDF 方法，另外三种是基于图模型的方法。对基于图模型的方法，分别尝试了三种加权方法，一种是距离的倒数，一种是共现次数，一种是本文提出的将四种信息融合的 PCFO 方法，利用以下四种关键词提取方法与扩展后的情感词、依存模板信息、位置信息（抛物线形式）一起加入 SVM 分类，并对其进行分析，结果如表 1 所示：

表 2.2 情感关键句抽取中不同关键词词典构建方法的比较

Methods	P/%	R/%	F/%
TFIDF	31.49	48.63	38.22
1/distance	32.51	49.57	39.26
Co-occurrence	33.84	50.10	40.39
PCFO	**36.29**	**52.35**	**42.87**

实验结果表明，本文提出的针对关键词提取的 PCFO 方法大大提升了情感关键句提取的效果。这主要是因为本文采用了 LDA 与 Textrank 相结合的方法，克服了传统图模型中随机游走的缺点，并采用 PCFO 算法综合考虑位置、覆盖、频度、共现四种影响力对图模型的权值进行修正。为使 PCFO 算法达到最优，实验研究了α,β,γ,δ四个参数，即位置、覆盖、频度、共现四种影响力对实验结果的影响。实验采用 5 种不同的(α,β,γ,δ)的组合，前三种只考虑四种影响力的一部分，后两种主要研究四种影响力在不同比重下对实验结果的影响。实验结果如图 2.4 所示，其中 1,2,3,4,5 分别代表(0,1,0,0)，(0.5,0.5,0,0)，(0.3,0.4,0.3,0)，(0.2,0.3,0.2,0.3)，(0.25,0.25,0.25,0.25)五种组合。

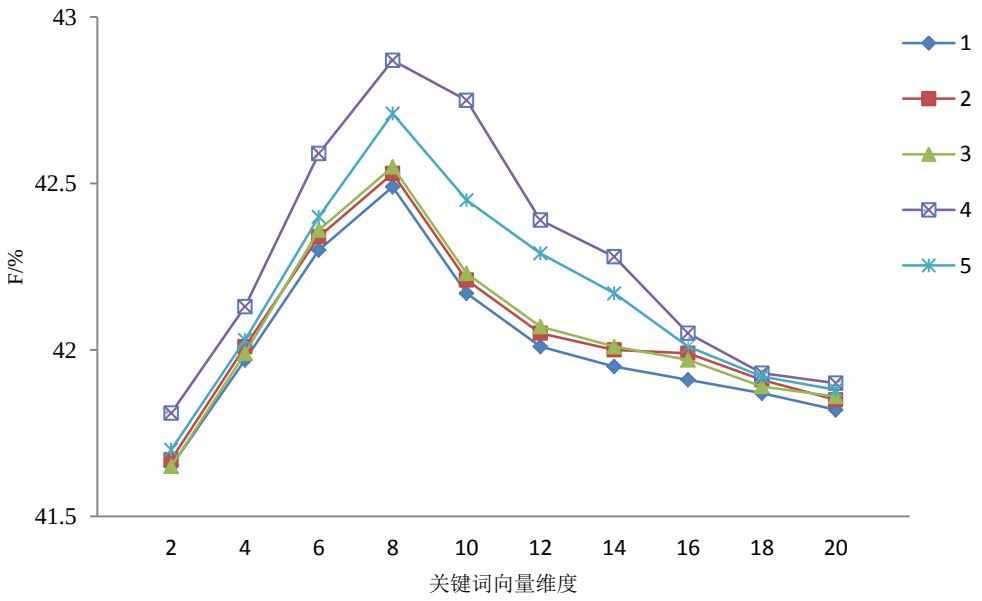

图 2.4 不同(α,β,γ,δ)组合与不同关键词向量维度下的情感关键句抽取的 F 值的比较

从图中可以看到，当选择 8 维作为 SVM 关键词向量维度，并使用第 4 种组合(0.2,0.3,0.2,0.3)时，实验效果最好。在关键词向量维度选择上，过大的维度反而会降低分类能力；而在四种影响力的组合上，综合考虑位置、覆盖、频度、共现四种影响力的组合远比只考虑一部分的效果要好。另外，比较 4,5 两种组合发现，覆盖和共现影响力即β,δ比位置和频度影响力即α,γ更重要，这主要是因为覆盖和共现影响力主要描述的是词与词之间的联系重要性，而不是词本身的重要性。

因此，本实验在一方面验证了 PCFO 算法的高效性，另一方面也证明了关键词提取质量对情感关键句提取的重要性。

2.5.4 不同特征组合的比较

在本小节，实验采用 4 种候选特征的不同组合加入 SVM 进行实验，由于情感关键句的定义是包含情感词和关键词这两个因素，因此，选择以下 4 种特征组合：情感词（Sentiment）与关键词（Keyword）；情感词、关键词与依存模板信息（dp）；情感词、关键词、依存模板信息与采用改进的高斯分布（Pos_(1)），公式（2.19）进行打分的位置信息；情感词、关键词、依存模板信息与采用抛物线（Pos_(2)），公式（2.20）进行打分的位置信息。实验结果如表 2.3 所示：

表 2.3 不同 SVM 特征组合的比较

Methods	P/%	R/%	F/%
Sentiment+Keyword	23.04	50.02	31.54
Sentiment+Keyword+dp	33.24	50.79	40.49
Sentiment+Keyword+dp+Pos_(1)	35.13	51.76	41.85
Sentiment+Keyword+dp+Pos_(2)	**36.29**	**52.35**	**42.87**

实验结果表明，加入依存分析之后，虽然召回率的改变不是很大，但是准确率和 F 值得提升却是显而易见的。这主要是因为依存分析可以深层次的挖掘句子里隐含结构信息，为情感关键句的提取起到了至关重要的作用。另外，加入位置信息后，情感关键句的提取效果也有了很大提升，这主要是由文章结构决定的。同时，实验也对文中提出的两种位置打分函数进行验证，发现使用抛物线形式优于高斯分布形式，这主要是因为高斯分布曲线在篇章首尾部分过于平滑，对篇章首尾部分的句子打分函数值变化不是很大，不能很好地体现出篇章首尾句子重要性。

2.5.5 不同情感关键句识别方法的比较

本节比较了本文融合词汇语义和句法依存的方法，表格中简写为 *Lexicon + Syntax(Rules+Statistics)*，与其他四种基本方法：COAE2014 任务 1 的最好结果（简写为 *COAE*）、基于词汇的方法（简写为 *Lexicon*[9]）、人工标注 500 条数据作为训练集的结果（简写为 *COAE-500labelled*）和去掉本文情感关键句识别流程中第三步，即不预先过滤掉一部分句子获取候选情感关键句的方法（简写为 *Lexicon+Syntax(Statistics)*），实验结果如表 2.4 所示。

表 2.4 不同情感关键句识别方法的比较

Methods	P/%	R/%	F/%
COAE	10.41	38.88	16.42
Lexicon[9]	12.18	29.13	17.19
COAE-500labelled	16.74	39.09	23.44
Lexicon+Syntax(Statistics)	30.70	50.79	38.27
Lexicon + Syntax(Rules+Statistics)	**36.29**	**52.35**	**42.87**

实验结果表明，基于词汇语义和句法依存的方法大大提升了情感关键句识别效果。

这主要是因为本文方法融合了词汇语义和句法依存，很好地挖掘了句子潜在的结构信息。而且，即使仅仅选择 500 条人工标注的句子进行实验，仍然取得了比 COAE 和基于词汇方法更高的效果。另外，当使用情感词典、关键词词典对语料进行规则过滤的时候，其实相当于一个降噪的过程，然后再用统计的方法，分析句法语义信息进行处理，以保证达到更高的准确率 P、召回率 R 和 F 值。

2.6 本章小结

本文针对情感关键句的识别进行了简要的介绍，提出了新的解决思路并进行验证。将情感关键句的识别过程看作一个二分类问题，通过情感词典的扩充与关键词词典的创建，首先对所有文章中的句子进行规则过滤，然后将情感词、关键词、依存模板和位置特征一起加入 SVM 中进行实验。实验结果显示，该方法显示出了比前人更优的抽取效果。

但有些问题还需要进一步的研究，下一步工作中将重点探究如下问题：

1）考虑对句子进行短语结构分析，并将其与依存分析相结合，共同为情感关键句的提取服务。

2）对现有的依存模板进行同义词扩展，并对依存关系提取算法进行改进，更进一步挖掘句法依存关系。

3）尝试使用不同分类算法进行实验，比对分类算法的优劣。

第 3 章 基于领域词典与 Word Embedding 的评价对象识别

3.1 概述

随着互联网新媒体，譬如微博、博客、论坛等的迅速兴起，网页中短文本诸如新闻评论、微博等数据呈海量式增长，其利用价值越来越受关注。这些信息为政府决策、企业经营、用户导购提供了必不可少的信息支撑。微博数量巨大、讨论范围广泛、仅靠人工的方法难以应对海量信息的收集和处理，采用自然语言处理技术对中文微博进行分析，已经成为当前研究的热点。

评价对象（Opinion Targets）是指评论性句子（主要是观点句）中所讨论的主题内容，具体表现为观点句中观点词语所修饰的对象属性。例如“我很喜欢 iphone。”，“iphone”就作为评价对象被提取出来。详细的评论信息挖掘分析对全面掌握互联网用户反馈，具有非常大的价值。而且，通过获取用户反馈来相应改善产品缺陷，方便用户使用，对提高产品的市场占有率也有着极其重要的作用。评价的对象抽取的情感粒度分析更小，所以理论和分析方法同篇章级的算法有所不同。对细粒度的情感要素提取决定了上层次情感文本的表示效果的好坏。评价对象识别的研究任务是为上层次情感分析的必要组件，对评价对象的准确识别是进一步分析的前提，比如：观点问答系统、推荐系统、汇总统计系统，这些系统的一个基本任务就是需要准确的找到观点句的评价属性，即评价对象。本文就是通过设计相关的算法，从句子中抽取所需要的评价对象。

微博语言和其他的媒体语言有所不同，它有着本身的特点[55-56]。

(1) 长度较短，结构差异性大。微博的发布字数被限制在固定的字数内，比如新浪微博限制在 140 字内。

(2) 交互性，不规范文本多。虽然长度短，但是结构差异大，不规则程度高。

(3) 特征关键词数量少，存在大量的变形词的新词。微博的网络结构决定了网络词语多且新词的出现概率高。不同时刻，网络词都会有特定的倾向性。

在产品领域的评价对象识别上有了很多学者的研究，但是目前商业上应用的系统还主要是基于规则的名词或词组的提取上。基于规则的算法虽然得到了一定的应用，但是在准确率等方面还是不如基于统计的方法。而现存的评价对象识别方法大多都只考虑了词法特征，并没有挖掘潜在的句法和语义知识。为了解决这个问题，本文提出

了一种基于句法和语义的融合领域词典与特征组合的评价对象识别方法，并且同时将微博的这些特点考虑在内。本文根据国内外相关研究的分析，基于领域词典构建和 Word Embedding，提出了一种领域词典构建的混合方法 PDSP，并且通过挖掘句法和语义信息获得不同的特征组合，然后把评价对象提取看成一个序列标注的任务，通过选择不同的特征组合进行实验，将 CRF 提取结果和领域词典相结合，完成最终的评价对象提取。实验中，我们在不同的评价标准下，分别对 COAE2014 提供的语料集进行实验，实验结果表明，本文所提方法能在很大程度上提升了评价对象识别的效率。

3.2 PDSP：一种基于句法和语义的领域词典构建的融合方法

所谓领域词典构建，就是使用规则的方法自动地从语料中提取评价对象。由于领域词典构建的好坏与不同的提取方法息息相关，因此我们提出了一种融合词性模板（POS template）、依存结构分析（Dependency parsing）、语义角色标注（Semantic role labeling）和短语结构分析（Phrase structure analysis）的领域词典构建方法。

3.2.1 词性模板

我们知道，评价对象一般都是名词或名词短语，并且与其附近的词性关系紧密。因此，我们首先对分词标注结果进行一些语义处理，然后观察其前后两个词的词性，提取出词性模板如表3.1。其中，*n*,*adj*,*adv*,*aw*,*cmp*和*OT*表示名词、形容词、副词、程度副词、比较词和评价对象，这里我们从Hownet获得副词和程度副词，从文献[57]获得比较词。

表3.1 词性模板

Template	Example
n+adv+adj	屏幕/*OT* 很好
n+adj	外观/*OT* 漂亮
aw+n	认为蒙牛/*OT*
adj+的+n	轻薄的机身/*OT*
n+cmp+n	iphone/*OT* 不如三星/*OT*
n+n	蒙牛牛奶/*OT*

3.2.2 依存结构分析

当我们向一个产品发表评论的时候，常常需要一些评价词来表达观点，而这些评价词往往是和评价对象有着很强的语义联系的。因此，我们将知网 Hownet 和台湾大学的情感词典 NTUSD 做为评价词的集合，同时使用哈工大的 HIT-LTP[58]进行依存分析，主要研究分析结果中评价词和评价对象中的“ATT 定中”、“SBV 主谓”关系，已知评价对象和未知评价对象的“COO 并列”关系。例如“效率和画质都好于一般摄像头。”，它的依存分析结构如图 3.1 所示。从图中可以看到，我们可以从评价词“好”与“效率”的 SBV 关系，得出“效率”为评价对象，然后通过“效率”和“画质”的 COO 并列关系，得出“画质”为评价对象。

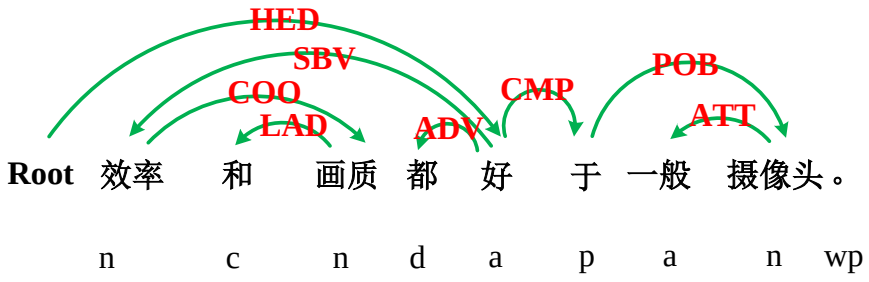

图 3.1 依存分析结果

3. 2. 3 语义角色标注

作为浅层语义分析的一部分，语义角色标注[59]在词汇语义分析中占据着非常重要的地位。观点句中，人们通常通过评价词来表述观点，而形容词和动词则为评价词的两种主要形式。通过观察，我们发现，当评价词为形容词的时候，施事者 A0 为评价对象；当评价词为动词的时候，受事着 A1 为评价对象。图 3.2 为“尼康 D7000 的外观很漂亮”和“我喜欢尼康 D7000”的语义角色标注结果。

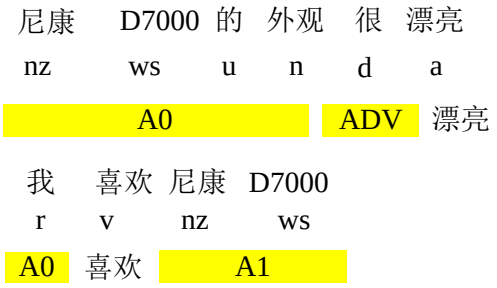

图 3.2 语义角色标注结果

最后，我们对语义角色标注结果进行分析，提出了基于语义角色的评价对象识别算法如下：

输入：分词结果、词性、语义角色标注结果（SRL）；情感词典 SL。

输出：评价对象集合 OT。

Step 1：遍历分词结果中的每一个词，若词在情感词典 SL 中，继续 Step 2；否则循环 Step 1。

Step 2：判断词性：若词性为形容词，继续 Step 3；若词性为动词，继续 Step 4；否则，继续循环 Step 1。

Step 3：将语义角色标注 SRL=A0 的词语加入 OT。

Step 4：将语义角色标注 SRL=A1 的词语加入 OT。。

Step 5：若所有分词都已被遍历，则停止迭代；否则，继续循环 Step 1。

通过以上算法，我们可以提取出“尼康D7000的外观”和“尼康D7000”作为评价对象。

3.2.4 短语结构分析

观察分词结果的时候，我们发现，“奥迪 A4”、“三星 i9300”经常被分为“奥迪，A4”、“三星，i9300”，而这显然是不够准确的。为了评价对象识别的边界识别能力，本文提出了一种基于短语结构分析的方法。该方法首先使用 Stanford Parser[60]对句子进行短语结构分析，采用迭代的规则，合并同一父亲节点的评价对象，并记录两个评价对象的父节点标识。例如图 3 为“好喜欢红色的奥迪 A4，比奔驰 C200 好很多.”的短语结构分析结果，“奥迪”和“A4”并不是作为一个整体的节点“奥迪 A4”，因此我们记录其父节点标识的组合序列“NR+CD”，通过进一步分析，对于尚未识别出的“奔驰”、“C200”进行进一步组合，最终提取出“奥迪 A4”和“奔驰 C200”作为该句对应的评价对象。

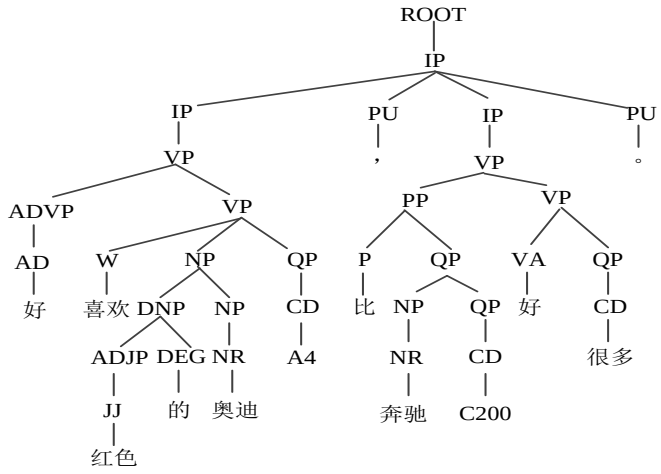

图3.3短语结构树

综上所述，领域词典构建的PDSP算法如下：

输入：预处理之后的语料 C。

输出：基础评价对象领域词典BSL。

Step 1：使用表 3.1 中的 6 个词性模板对语料中的句子进行匹配，选择匹配句子中的名词或名词短语作为评价对象，并加入基础评价对象领域词典 BSL。

Step 2：对句子进行依存分析，通过分析依存结果中的“SBV 主谓”、“ATT 定中”和“COO 并列”关系，提取出另一部分评价对象也加入 BSL。

Step 3：运行基于语义角色的评价对象识别算法，提取出一部分评价对象也加入基础评价对象领域词典 BSL。

Step 4：遍历词典 BSL，合并在语料中相邻的评价对象，并根据短语结构分析，记录已合并评价对象的父节点标识，然后根据标识组合识别出一部分新的评价对象，最终完成领域词典的构建工作。

3.3 基于 Word Embedding 的评价对象词典扩展

在文本分析的向量空间模型中，是用向量来描述一个词的方法有很多，最常见的 One-hot representation。One-hot representation 方法是用一个很长的向量来表示一个词，向量的维度为语料集合中的去重后的总词数，而向量中所有维，只有在该词对应的维度为 1，其他均为 0。该方法简单易用，但其有两个明显的缺点：（1）容易产生维数灾难，尤其当语料级别很大的时候。（2）词与词之间没有建立关联，不能很好的描述词语之间的相似性，会产生“词汇鸿沟”。因此，在深度学习中，一般采用 Distributed Representation 来描述一个词，常被称为“Word Embedding[61-62]”或“Word Representation”，也就是我们俗称的“词向量”。这种方法通过训练，将语料中的每一个词映射成为一个固定长度的向量，然后将所有向量放在一起，即可形成一个词向量空间，并根据向量与空间中点的对应关系，利用距离来描述词语之间的相似性。

Word Embedding 训练语言模型是利用第 m 个词的前 n 个词预测第 m 个词，而训练词向量是用其前后各 n 个词来预测第 m 个词，这样做真正利用了上下文来预测，如图 3.4 所示。

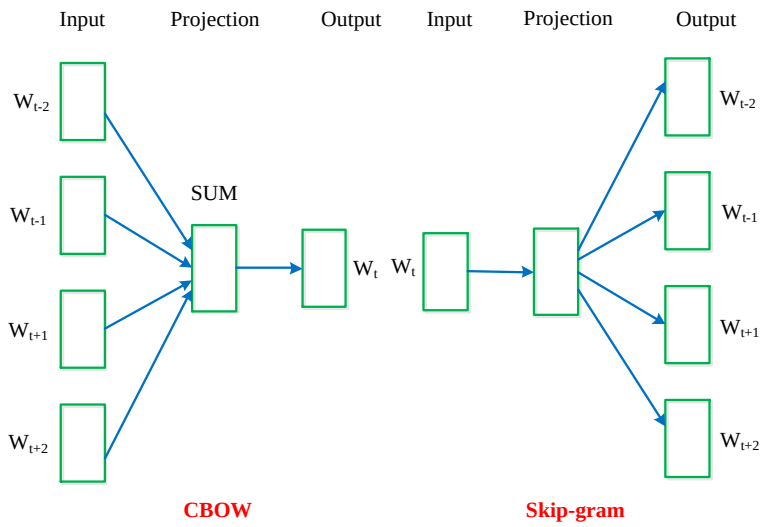

图 3.4CBOW 和 Skip-gram 两种语言模型

图 3.4 是 CBOW(continuous bag-of-words)和 Skip-gram 两种语言模型。在 CBOW 方法里，训练目标是给定一个当前词的上下文，预测当前词的概率；在 Skip-gram 方法里，训练目标则是给定一个当前词，预测当前词上下文的概率。

NNLM 是 Neural Network Language Model 的缩写，即神经网络语言模型。Word Embedding 采用一个三层神经网络来构建语言模型，将每个词被表示为一个浮点向量。其模型图如下：

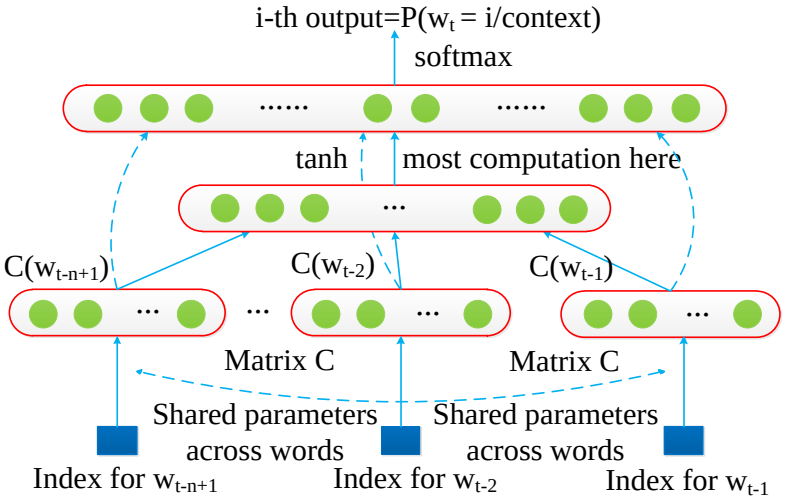

图3.5三层神经网络模型

图 3.5 描述的是根据已知的前$n-1$个词$W_{t-n+1}\dots W_{t-2}W_{t-1}$，来预测下一个词$W_{t-1}$每个是输入词都被映射为一个向量，该映射用C表示，所以$C(w_{t-1})$即为w_{t-1}的词

向量。设|V|表示词表，即语料中词语的总数；m 表示词向量的维度；则针对此三层神经网络，其每一层的任务如下：

第一层（输入层）：将$C(W_{t-n+1}) \ldots C(W_{t-2})\ C(W_{t-1})$这$\mathrm{n}-1$个向量，拼接起来，形成一个$(\mathrm{n}-1)\times \mathrm{m}$维的矩阵，将其记作$X$。

第二层（隐含层）：与传统神经网络一样，使用$\mathrm{d}+\mathrm{HX}$计算，d是一个偏置项。计算完之后，使用tanh作为激活函数。

第三层（输出层）：一共有|V|个节点，每一个节点y_i表示下一个词为 i的未归一化 log 概率，最后再使用 softmax 激活函数，将输出值y归一化成概率，设U表示隐含层到输出层的参数，则其计算公式如（3.1）：

$$y = b + WX + U\tanh(d + HX) \qquad (3.1)$$

最后利用梯度下降等优化算法对模型进行迭代优化，即可获得每个词的词向量。在利用 Word Embedding 获取到每个词的词向量之后，将领域词典中的词语和语料中的词分别计算语义相似度，取领域词典中每个词相似度得分最高的前三个名词作为扩展，加入基础领域词典(BSL)，从而获取到扩展后的评价对象词典(EDL)。

3.4 基于 CRF 的评价对象识别过程

上文3.2和3.3节，分别从词性模板、依存分析、语义角色和句法结构四个方面，构建了基础领域词典（BDL），并基于Word Embedding进行基础领域词典扩展，生成扩展后的领域词典（EDL）。本节主要介绍如何利用上文的基础工作，并将其融入到机器学习模型CRF中，进行评价对象识别。

3.4.1 条件随机场 CRF 简介

条件随机场模型（Conditional random fields，CRF）是Lafferty等[63]在2001年提出了，它是在最大熵模型和隐马尔科夫模型的基础上，提出的一个无向图模型或马尔可夫随机场，是一种用于标记和切分序列化数据的统计模型。该模型是在给定需要标记的观察序列的条件下，定义下一个状态的分布。CRF模型最早是针对序列数据分析提出的，现已成功应用于自然语言处理领域，被广泛地用于分词、词性标注、命名实体识别等任务中。与HMM模型相似，CRF模型也假设模型是符合马尔科夫性的，不同于HMM模型的是，CRF模型在条件转移矩阵之外又增加了词语特征维度。CRF模型定义如下：设无向图$\mathrm{G}=(\mathrm{V},\mathrm{E}),\mathrm{Y}=\{\mathrm{Y_v}|\mathrm{v}\in\mathrm{V}\}$是以G中节点为索引的随机变量$\mathrm{Y_v}$构成的

集合。在给定X的条件下，如果每个随机变量服从马尔科夫属性，即$P(Y_v|X,Y_u,u\neq v)=P(Y_v|X,Y_u,u\sim v)$，其中u~v表示u和v是相邻的边，则(X,Y)就构成一个条件随机场。

理论上讲，图G的可以为人以结构，然而在构造模型时，CRFs采用了最简单和最重要的一阶链式结构，即线性结构（Linear-chain CRFs），如图3.6。

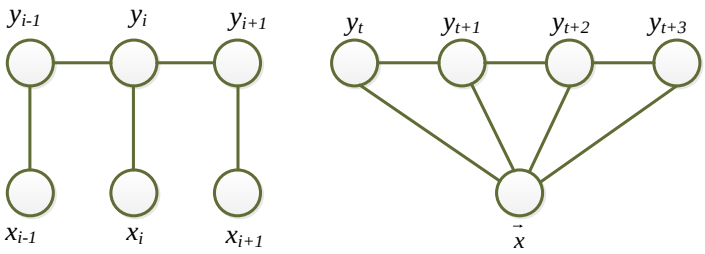

图 3.6 CRFs 一阶链式结构

Lafferty对条件随机场势函数的定义如下：

$$\Phi_{y_c}(y_c)=\exp(\sum_k \lambda_k f_k(c,y|c,x)) \quad (3.2)$$

而在一阶链式结构图$G=(V,E)$中，最大团仅包含图中相邻的边的两个节点。对一个最大团中的无向边$e=(v_{i-1},v_i)$，势函数的表达形式可以扩展为以下公式：

$$\Phi_{y_c}(y_c)=\exp(\sum_k \lambda_k t_k(y_{i-1},y_i,x,i)+\sum_k \mu_k s_k(y_i,x,i)) \quad (3.3)$$

其中，$t_k(y_{i-1},y_i,x,i)$是观察序列和位置i以及i-1所对应标记序列的状态转移特征函数。而$s_k(y_i,x,i)$是在位置i标记和观察序列的特征，它是一个状态函数。联合概率的表达形式可以表示为：

$$p(y|x)=\frac{1}{Z(x)}\exp(\sum_i\sum_k \lambda_k t_k(y_{i-1},y_i,x,i)+\sum_i\sum_k \mu_k s_k(y_i,x,i)) \quad (3.4)$$

为了统一状态函数和转移函数的表达方式，可以将状态函数写为：

$$s_k(y_i,x,i)=s_k(y_{i-1},y_i,x,i) \quad (3.5)$$

并且用 $f_k(y_{i-1},y_i,x,i)$ 统一进行表示，f_k 可能是转移函数 $t_k(y_{i-1},y_i,x,i)$ 或状态函数 $s_k(y_{i-1},y_i,x,i)$，又令：

$$F_k(y,x)=\sum_{i=1}^{T} f_k(y_{i-1},y_i,x,i) \quad (3.6)$$

因此，在给定观察序列x的时候，其对应的标记序列 y 的概率可以写成：

$$p(y|x)=\frac{1}{Z(x)}\exp(\sum_k \lambda_k F_k(y,x)) \quad (3.7)$$

其中 $Z(x)$ 代表归一化因子。

目前，常用的CRFs的实现模型有 CRF++6、CRF Mallet toolkit 等，本文选用CRF++作为评价对象识别的序列标注模型，其使用过程大致分为4个步骤：数据预处理、生成特征模板、训练和测试。

3.4.2 特征选择

在 CRF 模型中，最关键的是特征的选取问题，由于基于条件随机场的中文评价对象识别的研究并不多，我们参考国外基于该算法的英文特征进行对比。在这一章中，我们参考 Jakob[20]在英语中常用的评价对象提取的特征，并且根据中文特殊语法现象引入新的特征，对这些特征进行分类，分组进行试验，比较特征之间的差别，最终选取出最优特征。在本文中，我们根据特征的来源把特征分为以下四类：基本的词法特征、依存关系特征、相对位置特征和语义特征。下面分别对这些特征进行简单的介绍。

采用以下句子为例（其中情感词为“合理”，当前词为“键位”），介绍各个特征：

尼康 D7000 的键位设置合理。

(1) 基本词法特征（lex）

该特征是从基本的词或者词法等来源的，比如词语本身，词语的词性等。

a) 词特征，即选择当前词为特征，例句中选择“键位”为特征。词本身可以很好的对评价对象进行指示，但是词语本身存在一定的局限性，不具有特征的通用性。由于中文的词语多且产生歧义的情况比较频繁，所以词特征对英文评价对象识别的作用更加明显。但是在某一领域中，在一定的训练语料基础下，词特征对结果的影响也非常显著。

b) 词性特征，即选择当前词的词性为特征，例句中选择“键位”的词性即名词词性为特征。对于评价对象的抽取上，评价对象一般都会以名词的形式出现，动词和形容词作为评价对象的概率就会比较小，并且评价对象前后的词性分布相同的概率比较大。所以词性特征的选择，对评价对象的抽取和评价对象边界的确定都有非常重要的影响。

(2) 依存关系特征

依存分析把依存关系反映了句子中词语间的语法依存关系，由于其表示了句子内部逻辑，有助于对句子进行进一步的分析。

a) 最短路径特征，表示当前词同中心词之间的直接依存关系，根据分析，情感词是指示评价对象很好的特征。依存于评价词或中心词的名词等词性一般属于评价对

象。

b) 依存关系，识别当前词与中心词之间依存关系的类型。在评价性句子中，评价词可以很好的表征评价对象的位置信息。在例句中，当前词“键位”和情感词“合理”之间存在依存关系。

c) 父亲词本身：在依存分析中，识别出的当前词的父亲词。

d) 父亲词词性: 在依存分析中，识别出的当前词的父亲词词性。

(3) 相对位置特征

a) 词距离特征：在汉语句子中，词语之间的相对位置可以决定词跟词之间的关系。由于评价词一般对评价对象具有修饰作用，所以在情感词周围的词出现评价对象的概率比较大。通过语料也可以发现这个特征。因此我们选择当前词与中心词的距离是否小于 5 作为该部分特征。

(4) 语义特征

a) 是否是情感词：识别当前词是否是情感词

b) 语义角色名：在句子层面，语义分析在形式上利用其句法结构及动词词义推导出句子释义。作为实现语义分析这一目的的语义角色标注主要利用“谓语动词（或名词）-角色”表示结构，赋予句中的语义角色以特定语义。比如对于句子“他出生在河南”和“他的老家是河南”，在句法分析中，虽得出的结构可能会不尽相同，但所体现的深层次语义却是一致的。语义角色标注的任务即挖掘指定句子中谓语动词的语义角色成分：中心语义角色和附属语义角色。因此，我们选择语义角色标注结果中当前词的语义角色作为该部分特征。

c) 句中情感词词性：在前面我们已经介绍过，评价对象的提取不仅和语义角色有关，而且与句子情感词的词性（形容词、动词）密切相关。因此，我们选择该句中情感词的词性作为该部分特征。

3.4.3 系统框架

综上所述，本章评价对象识别的主要流程如图 3.7 所示：

1. 预处理：分词、词性标注、依存分析、句法分析、语义角色标注；
2. 通过匹配词性序列模板、分析依存结果类型、提取语义角色信息和句法结构，来构建基础评价对象领域词典 BDL；
3. 采用 Word Embedding 方法对 BDL 进行扩展，获得扩展后的领域词典 EDL；
4. 生成候选情感关键句的 5 种特征：词汇特征、语义角色特征、领域词典特征、

相对位置特征和依存特征；

5. 使用 CRFs 进行分类，识别出每个句子中的评价对象。

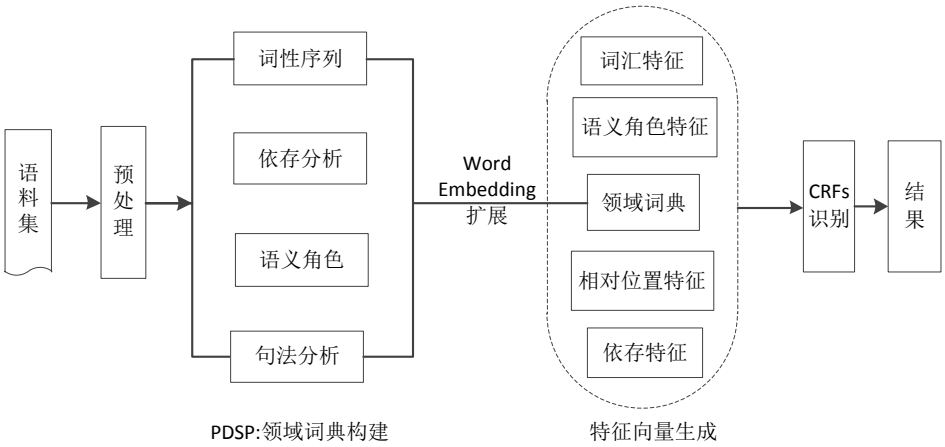

图 3.7 评价对象识别的方法流程

3.5 实验结果与分析

实验中，首先通过上文提出的领域词典构建的 PDSP 方法获得评价对象作为集合 X；然后将次领域词典与第三部分的基本的词法特征、依存关系特征、相对位置特征和语义特征一起加入 CRF 进行实验，获得评价对象集合 Y；最后合并集合 X 和 Y 完成最终的评价对象识别，并将结果与标注集合作对比分析。

3.5.1 语料预处理

考虑到微博语料集中句子的长度相差很大、并且存在转发关系、结构复杂，变化性强、使用不规范，标点符号错误等问题，根据任务，我们定义了以下过滤规则，使句子更加规整：

规则 1：去除纯英文句子（目前主要专注于中文句子的分析）；

规则 2：对句子进行“//”划分，并且使句子顺序翻转。（这样保证句子的转发关系，使后面的句子基于前面的句子进行分析）；

规则 3：对句子中用户名进行删除，即删除“@+用户名”结构，删除“http://...”这样的网址结构；（暂时不考虑用户名作为微博主要内容的情况）；

规则 4：对连续出现多个标点符号，如“。。。。。”，“!!!!”等，采用第一个标点符号进行替换，去除微博中表情标示符；

规则 5：对于微博中特殊的“#内容#”，则把较短内容直接作为候选评价对象，较

长内容作为一个单独的句子另行分析；

规则 6：若句子中包含转折句，则删除转折句前面的结构，并且句子中代词以转折句前面出现的名词或者名词词组代替

规则 7：删除问句句子，主要是以问号结尾的句子（暂时不考虑反问句、疑问句结构）

经过上面的规则，在句子结构上进行了调整。使句子能够更好的进行更加深入的分析。实验中，我们使用 COAE2014 任务五提供的语料集，通过以上八条规则进行过滤，最终得到 5000 条规范化、具有情感倾向的句子。并在此语料集合上进行实验，采用精确评价（strict）和覆盖评价（lenient）下的准确率（Precision）、召回率（Recall）和 F 值（F-measure）进行评价。其中，精确评价是抽取的结果与标注的结果完全一致，覆盖评价是抽取的结果与标注的结果存在交集。

3.5.2 不同领域词典构建方法的比较

由于领域词典的精确与否和不同的构建方法息息相关，因此，本节主要对比了不同的领域词典构建方法。我们主要使用了词性[64]（POS）、句法[65]（Syntactic）、语义角色（semantic role）和 PDSP 方法进行领域词典构建，实验结果如表 3.2 所示：

表 3.2 不同领域词典构建方法的比较

Method	Strict Evaluation			Lenient Evaluation		
	P(%)	R(%)	F(%)	P(%)	R(%)	F(%)
POS[64]	32.38	41.96	36.55	34.36	44.24	38.67
Syntactic[65]	34.52	43.88	38.64	37.50	44.92	40.87
Sematic Role	33.62	42.61	37.58	36.21	45.18	40.20
PDSP	**40.36**	**51.07**	**45.08**	**43.71**	**52.18**	**47.57**

从实验结果中，我们可以看到，本文所提出的 PDSP 方法不论是在精确评价（strict）还是覆盖评价（lenient）下都取得了非常不错的实验效果。究其原因，主要是由于该方法综合考虑了词性模板（POS template）、依存结构分析（dependency parsing）、语义角色标注（semantic role labeling）和短语结构分析（phrase structure analysis）四大类知识，为领域词典构建挖掘了更多有用的信息。

3.5.3 不同特征组合的比较

在本节中，我们采用了五种不同的特征组合使用 CRF 进行实验。其中，*lexi*, *pos*,

dp, srl, bdl 和 *edl* 分别代表基本的词法特征、相对位置特征、依存关系特征、语义特征、基础领域词典和扩展后的领域词典特征。实验结果如表 3.3 所示：

表 3.3 特征组合对比的实验结果

Method	Strict Evaluation			Lenient Evaluation		
	P(%)	R(%)	F(%)	P(%)	R(%)	F(%)
lex	64.70	41.06	50.23	69.32	42.72	52.86
lex+pos	65.38	41.96	51.11	70.40	43.16	53.51
lex+pos+dp	66.73	42.61	52.00	71.34	44.63	54.90
lex+pos+dp+srl	68.24	44.72	54.03	73.74	46.85	57.29
lex+pos+dp+srl+bdl	68.64	45.13	54.46	73.98	47.12	57.57
lex+pos+dp+srl+edl	**69.01**	**46.05**	**55.23**	**74.32**	**48.55**	**58.73**

实验结果表明，当我们加入依存关系特征、语义特征和领域词典特征时，实验结果有了很大的提升，这主要是因为挖掘了潜在的句法和语义信息。例如“好想要啊，三星手机的音质非常赞！”，传统的方法只能提取出“三星手机”作为评价对象，而当我们加入语义特征之后，可以提取“三星手机的音质”作为评价对象。

3.5.4 不同评价对象识别方法的比较

本节我们将本文提出的评价对象识别方法和前人方法做了对比，对比结果如图 3.8 所示，其中，1 表示完全基于词汇的方法，2 表示 Jakob[20]的方法，3 表示 4.3 中 CRF 的最好方法，4 表示本文方法。

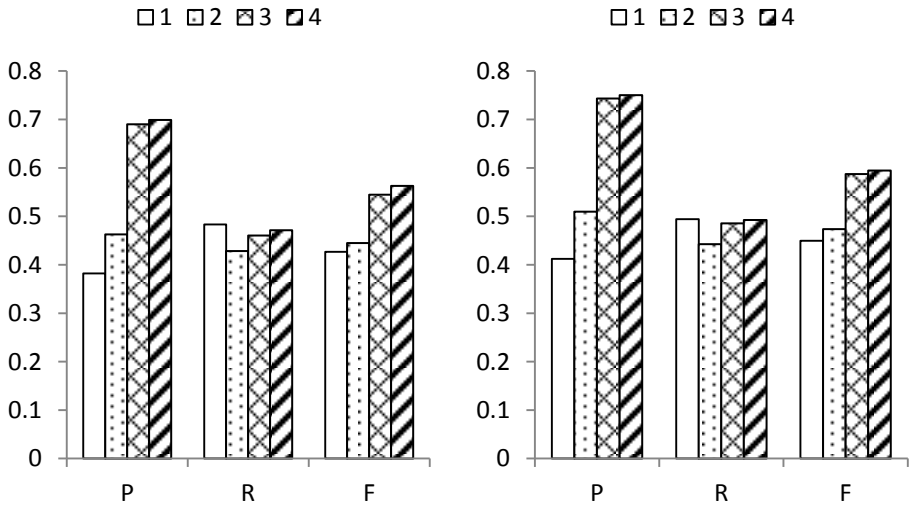

图 3.8 不同评价对象方法的对比实验结果

从实验结果中我们可以看到，本文提出的基于句法和语义的融合领域词典与特征

组合的方法，在很大程度上提高了实验效果。这主要是因为，该方法不仅使用领域词典构建方法PDSP获得一部分评价对象，而且还是用CRF的方法去弥补完全基于规则方法的不足，以达到更高的准确率、召回率、F值的目的。所以，该实验非常有力地证明了本文所提出的评价对象识别方法的有效性。

3.6 本章小结

本文主要提出了一种基于句法和语义的评价对象识别方法，其中包含一种融合词性模板、依存结构分析、语义角色标注和短语结构分析的领域词典构建的 PDSP 方法进行基础领域词典构建，并利用 WordEmbedding 对基础领域词典进行扩展；然后，融合基本的词法特征、相对位置特征、依存关系特征、语义特征和领域词典特征，选择不同的特征组合使用 CRF 进行实验，并将其识别结果与领域词典相结合，完成最终的评价对象识别工作。实验结果表明，在不同的领域词典构建方法和特征组合的情况下，本文方法都优于其他基准方法。这也进一步说明了，深层次的处理对传统浅层处理任务有着非常重要的影响。

在未来工作中，我们将主要研究以下三点：

1. 考虑到中文微博语言表达的多样性，我们准备挖掘更多的规则或提取出核心句进行分析。
2. 在为 CRF 选取特征的部分，我们将进一步研究，挖掘出更多有用的特征来进行评价对象提取。
3. 本文主要研究的是句子级别的评价对象提取，未来我们考虑研究语料级别的评价对象提取的工作。

第 4 章 多特征融合的话题相关情感倾向性分析

4.1 概述

微博(microblog)是近几年兴起的新的网络社交媒体，与 Twitter 类似，是一个基于用户关系的信息分享、传播以及获取平台，具有内容简洁，交流便利的优势，促进了虚拟社会间的交流。情感分析是指分析信息文本中所蕴含的情感倾向，对文本所表达的意见、看法进行分析评估。情感分析在微博海量数据上的应用，将有助于完善互联网的舆情监控系统，是自然语言处理中较前沿的研究领域。微博本身作为新生事物，其参与人数庞大，话题涉及广泛，已经成为了大众日常生活中的重要部分，因此对其进行情感分析的研究有着十分重要的意义。

中文微博以“＃topic＃”的形式表征话题分类，与 twitter 不同，中文微博的 topic 既可以为一个单独的词语，也可为一个短句。传统的情感分析大多是与话题无关的，即不必考虑情感所针对的对象，而这种分析方法在话题相关的情感分析中，显然是不够精确的。经分析，传统的情感倾向性分析方法在对以下两种类型微博进行分析的时候，往往性能不佳：

(1)包含多个话题的微博

例如,“＃三星 galaxy s6＃＃华为 P8＃＃mate8＃三星 galaxy s6 真新没什么亮点，华为 P8 就可以秒它了，更不用说 mate8[拜拜]”，该句对于话题“三星 galaxy s6”的情感判定是负向，而针对话题“华为 P8”和“mate8”的情感判定为正向。

(2) 含有话题相关情感词的微博

例如，“＃股票＃前天刚入手一支股票，一直在升，股价越来越高。”和“＃三星＃三星手机电量明显不够用，耗能高。”同样的一个情感词“高”，在第一句中，针对话题“股票”来说，被判定为正向；而在第二句中，针对话题“三星”来说，被判定为负向。

因此，为解决以上两种问题，必须细化情感分析的粒度，不再单纯地只为一条微博确定一个情感倾向，而是根据一条微博中的不同话题，赋予不同的情感倾向，同时还要注意情感词的领域性，即其与话题的相关性。考虑到话题信息对情感倾向性分析的重要性，本章主要研究基于话题的微博情感倾向性判定，提出了一种全新的多特征融合的话题相关情感倾向性分析，首先提出一种基于不同话题下局部和全局信息的词

图模型构建方法，然后通过 Word Embedding 方法对话题词扩展，根据话题词与情感词之间的依存关系来提取话题相关的情感词，最后结合前人工作对特征进行整合，融合多种特征并采用 SVM 分类模型完成最终的话题相关情感倾向性分析工作。

4.2LTIGT：一种基于局部和全局词图模型构建的特征提取方法

话题相关的情感倾向性分析，区别于传统情感分析技术，着重强调话题在情感倾向性分析中的重要性。因此，如何提取出适应所属话题的关键词特征，成为一个研究重点。本节主要提出了一种基于局部和全局下词图模型构建的分类特征提取算法，进行话题相关的关键词特征提取。

4.2.1 基本思想

与传统的文本分类相似，进行话题相关情感倾向性分析的一个重要任务，就是选择分类能力较高的词或者短语。传统的常用方法如TFIDF（Term Frequency-Inverse Document Frequency）方法等。Salton在1973年提出了 TFIDF 算法[65]，此后又论证了TFIDF方法在信息检索领域的有效性，后逐渐被融入向量空间模型中，应用到文本分类等自然语言处理任务中。TFIDF方法的计算公式如（4.1）：

$$TFIDF = TF \times IDF = \frac{n_{w,d}}{N_{w,d}} \times log\frac{N_d}{n_{d,w}+1} \quad (4.1)$$

其中，TF(Term Frequency)为词频、IDF (Inverse Document Frequency)反文档频率，$n_{w,d}$表示在文档d中词w的出现次数，$n_{d,w}$为出现词w的文档d的个数，$N_{w,d}$文档d的总词数，N_d为总文档数。

TFIDF 的主要思想是：如果一个词在特定的文本中出现的频率越高，$n_{w,d}$越大，即TF值越大，说明它在区分该文本内容属性方面的能力越强；如果一个词在文本中出现的范围越广，即$n_{d,w}$越大，IDF越小，说明该词区分文本内容的属性越低。总的来说，在一篇文章中的出现频率高，而在语料集合中的出现频率低的词语，拥有较好的分类能力。

而针对本文基于话题的情感倾向性分析，仅仅依靠TFIDF信息是远远不够的。主要原因归纳为以下两点：

(1) TFIDF方法没有考虑话题相关信息，仅根据出现频率对语料集进行统计得出的词，在话题相关任务中，并不具备同等分类能力，也不能很好地提升话题相关情感倾向性分析任务的性能。

(2) TFIDF方法没有考虑词的位置和共现信息，仅根据词频与逆文档频率并不能全面地表征一个词的分类能力，对本章所研究的话题相关情感倾向性分析来说，是不够完善的。

因此，本节根据TFIDF的基本思想，提出了一种融合局部和全局词图模型构建的LTIGT(Local Textrank-Inversed Global Textrank)算法，充分考虑了话题、位置、共现信息在话题相关情感倾向性分析任务中的重要性。该算法分别在局部（句子）和全局（每个topic下的所有句子）下构建词图模型，按照改进的TextRank[51]打分策略，给图中每一个词赋予两个不同的得分：局部下的TextRank值（LT）和全局下的TextRank值（GT）。

根据TFIDF算法思想，为了提取出更具有分类能力的特征，需提取局部重要性高、全局重要性低的词语作为特征词。故而，本节提出了一种融合局部和全局信息的LTIGT算法，用局部下的TextRank得分和全局下的TextRank得分，分别来衡量某个词在局部和全局下的重要性，其公式如（4.2）所示。

$$LTIGT = LT \times IGT = TR_{lt}(v_i) \times \frac{1}{TR_{gt}(v_i)} \qquad (4.2)$$

其中，$TR_{lt}(v_i)$和$TR_{gt}(v_i)$分别表示节点v_i在局部和全局下的TextRank得分。

4.2.2 话题相关的词图模型构建

考虑的话题信息在话题相关情感倾向性分析中的重要性，本节基于上文工作，主要介绍融合局部和全局信息的LTIGT算法中，话题相关的词图模型构建。如4.3.1中所述，LTIGT算法用于提取局部较重要，而全局相对不重要的词作为分类特征。显然，其中的两个核心部分便是局部下的TextRank值（LT）和全局下的TextRank值（GT），而如何通过话题相关的词图模型构建来得到这两个值，乃是本节的研究重点。

话题相关的情感倾向性分析中，一个词的分类能力不仅表现为其在单句中的重要程度，更与其在整个话题下的重要程度息息相关。而如何衡量一个词的重要性呢？本节采用词图模型构建的方法，按照TextRank基本思想进行话题相关的词图模型构建。

在构建话题相关的词图模型$G = (V, E)$时，需要考虑的无非节点、边、权重和跳转概率四个要点，本节图模型的构建与2.2.2节的构建方法类似，但主要对公式（2.12）的权值$w(v_j, v_i)$和跳转概率$P(z_t|v_i)$进行改进。

（1） 结点：图模型节点集合$V = \{v_1, v_2, v_3 \ldots\ldots v_n\}$，由文档中的名词、形容词组成；

（2） 边：连接两节点的边$(v_i, v_j) \in E$，采用滑动窗口的方法确定两个节点之间是否存在边以及边的方向。该方法在一个大小为window个词的滑动窗口内，分别按照顺序从第一个词指向窗口内的其他词语，然后依次向后滑动。经过试验，设定window大小为10。

（3） 权值：权值$w(v_j, v_i)$与上文所述PCF0方法基本一致，但考虑到两节所做任务不同，本节将$w(v_j, v_i)$的位置重要性公式（2.7）中的“是否在标题中出现”更改为“是否为话题词”；

（4） 跳转概率：采用一种全新的跳转概率计算方法$p_{rj}(w_i)$，将当前词与话题词的点间互信息（PMI）值作为当前词的随机跳转概率，其计算公式如（4.3）所示。

$$p_{rj}(w_i) = \frac{PMI(v_i, \text{topic})}{\sum_{j=1}^{|V|} PMI(v_j, topic)} \tag{4.3}$$

式中$PMI(v_i, \text{topic})$表示当前节点v_i和话题词topic的互信息值，|V|为图模型中节点个数。其中，在全局下的TextRank时，以句子为单位计算共现和单独出现的概率；在局部下的TextRank时，是以窗口为单位计算，经过实验，设置窗口大小为5。

根据以上方法，分别构建局部和全局下话题相关的词图模型，如图4.1所示。局部TextRank是指以句子为单位，分别为每个句子建图；全局TextRank是以话题为单位，分别为每个话题下的所有句子建图；进而分别得到每个词的LT和GT得分。

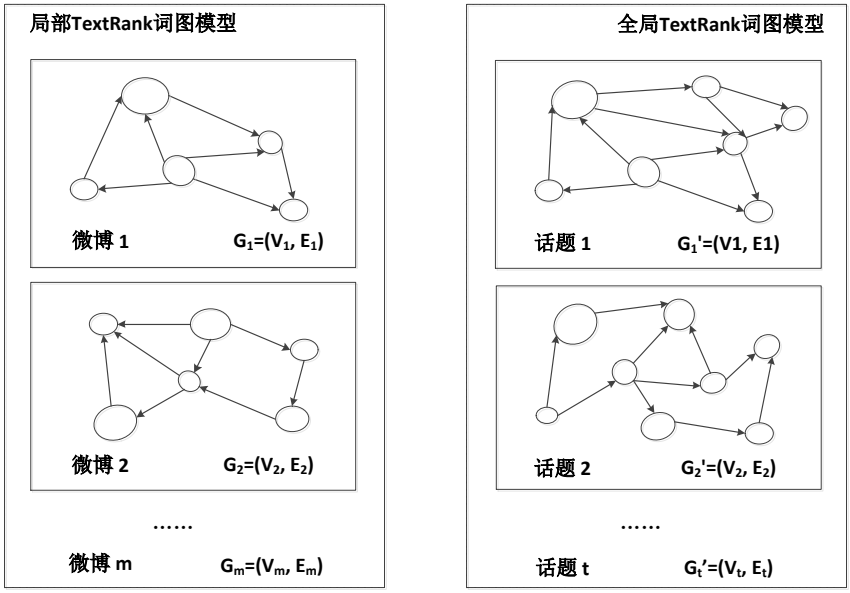

图4.1 话题相关的词图模型

在图模型建立完毕之后，分别利用公式（2.12）迭代计算每一个节点的得分，最后，对于每一个节点，我们分别可以得到一个局部和全局下的TextRank得分；然后使用公式（4.2）求得每一个词的最终LTIGT得分，用来表示每个词的分类能力。

4.3 基于 Word Embedding 和依存分析的情感词特征提取

情感词是文本情感分析过程中重要的情感特征，其具体是指在文本内容中具有情感倾向性的词语，一般是名词、形容词、副词、动词以及一些习惯用语等。通常情况下，文本表达情感倾向的主要方式是运用情感词表现，所以情感词在情感倾向性判定中占据十分重要的地位。基于本章所研究的话题相关情感倾向性分析任务，传统的情感词典不能解决4.1节中提到的“含有话题相关情感词的微博”的情况，因此，必须提取与话题相关的情感词进行情感倾向性分析。因此，本章提出了一种基于Word Embedding和依存分析的话题相关情感词特征提取方法。该方法首先通过word Embedding对话题词进行扩展，获取扩展后的话题词ETW，再以话题词为中心进行依存分析，继而根据依存关系提取话题相关的情感词TSW，最后对话题相关的情感词进行聚类，将每句话中属于某一类别的的情感词个数作为每一维的值，生成最终特征。

4.3.1 话题词扩展

例如句子：“三星S6的做工很赞，屏幕不错，价格也给力。”其可视化的依存关系分析如图4.2所示。其中“做工”、“屏幕”、“价格”均可作为话题词“三星”、“S6”的扩展词,并且情感词与“三星”、“S6”并不存在依存关系,而其与扩展词却存在,如SBV(做工，赞)、SBV(屏幕，不错)和SBV(价格，给力)，这些扩展词与情感词的依存关系直接影响了基于话题的微博情感倾向性分析。

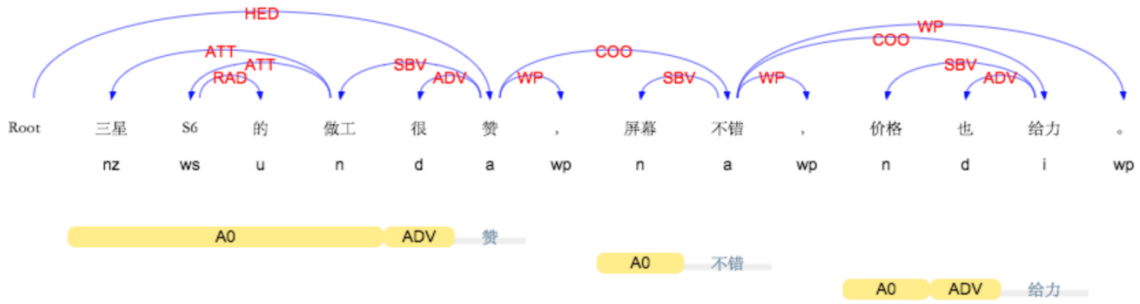

图4.2 依存分析结果示意图

获取word vector之后，我们分别计算每个话题下的名词与话题词的余弦相似度，取前N个词作为话题词的扩展词，加入ETW。至此，完成前期话题词的扩展的基础工作。

4.3.2 话题相关情感词提取

众所周知，人们在表达情感时往往是针对一个特定的话题或对象，而情感词与话题词之间，必定存在一定的依存关系。经过统计，发现话题词与情感词之间的依存关系类型主要有以下三种：

(1) 动宾关系。情感词作为动词,话题词是动词的宾语。例如“我喜欢三星。”，“喜欢”与“三星”存在动宾关系（VOB）。

(2) 主谓关系。情感词作为谓语,话题词作为情感词的主语。例如“三星很漂亮。”，“三星”和“漂亮”存在主谓关系（SBV）。

(3)定中关系。情感词作为定语，话题词作为定语中心语。例如“无与伦比的三星设计！”，“无与伦比”和“三星”存在定中关系（ATT）。

因此我们针对微博的依存分析结果，设计了一个基于依存关系的话题相关情感词提取算法，算法流程如下：

输入：依存分析结果 DP；扩展后的话题词 ETW。

输出：话题相关情感词 TSW。

Step 1：遍历依存分析结果 DP 中的每一个词，若词在扩展后的话题词 ETW 中，继续 Step 2；若词的父结点在扩展后的话题词 ETW 中，继续 Step 3；否则循环 Step 1。

Step 2：判断当前词的依存关系：若依存关系为‘SBV’、‘VOB’或‘ATT’，继续 Step 4；否则，继续循环 Step 1。

Step 3：判断当前词的依存关系：若依存关系为‘SBV’、‘VOB’或‘ATT’，继续 Step 5；否则，继续循环 Step 1。

Step 4：将当前词在 DP 中父结点对应的词语加入 TSW。

Step 5：将当前词加入 TSW。

Step 6：若所有分词都已被遍历，则停止迭代；否则，继续循环 Step 1。

4.4 多特征融合的情感倾向性分析过程

上文 4.2 和 4.3 节分别提出了一种基于局部和全局词图模型构建的 LTIGT 关键词特征提取算法，和一种基于 Word Embedding 和依存分析的话题相关情感词提取算法，本节主要介绍如何将以上两种算法与其他基础特征一起融入到向量空间模型中，使用不同的特征融合方式生成特征向量，完成话题相关的情感倾向性分析。

4.4.1 特征选择

由上文可知，通过 LTIGT 算法和话题相关情感词提取算法，可分别得到每个词的 LTIGT 得分与话题相关情感词集合 TSW。针对 LTIGT 特征，本章将每一个词对应的 LTIGT 得分作为每一维的值；而对 TSW 特征，本章采用一种特征聚类的方法，设置不同的聚类个数，采用 K-means[67]分别对其聚类，每一维度的值设置为当前句中属于该类别的话题相关情感词个数。此外，本章还结合前人工作，在张珊[68]和谢丽星[69]等人的基础上，充分考虑了前人所提出的特征。众所周知，副词，否定词和情感词在情感极性的判定中都占据着十分重要的作用。例如，“没有 MicroSD 卡插槽极之不方便，机身储存内容换机要拷贝转来转去。 [叉][叉][叉][叉][叉][叉][叉][炸弹][炸弹][炸弹][炸弹][炸弹][炸弹]”，这句话中“方便”是个正向情感词，但是它出现在否定词“不”的后面，并与情感词“方便”存在着“ADV”依存关系。因此本章融合以上所有特征，对话题相关的情感倾向性分析进行分类器构造，实验所用到的特征如表 4.1 所示。

表 4.1 话题相关情感倾向性分析任务的特征描述

特征	特征描述
表情符号/EM	正向与负向表情符号个数之差
否定和程度副词/NDA	每个否定和程度副词在句中出现，则设为 1，否则为 0
基础情感词典/BSL	正向情感词个数-负向情感词个数
三种依存关系/DP	与话题词存在 SBV、VOB、ATT 关系的正负向情感词个数之差
TFIDF	语料中每一个词的 TFIDF 词
LTIGT	语料中每一个词的 LTIGT 的值
TSW	话题相关情感词 K-means 聚类后每类中情感词个数

4. 4. 2 系统框架

综上所述，本章多特征融合的话题相关情感倾向性分析系统的流程如图 4.3 所示。

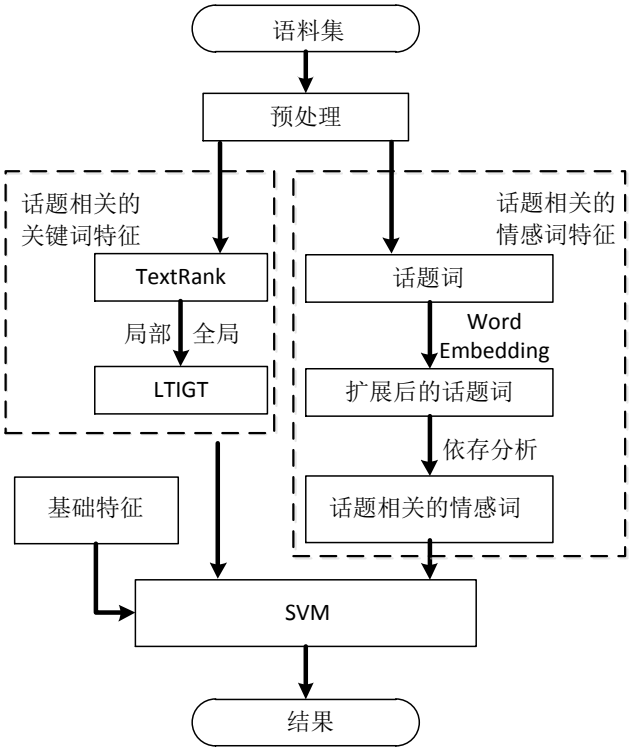

图 4.3 话题相关情感倾向性分析方法流程

1. 预处理：分词、词性标注、依存分析、句法分析、语义角色标注；
2. 按照 4.2 所述方法，构建词图模型获取 LTIGT 特征，作为话题相关的关键词特征；

3. 采用 4.3 节方法获得话题相关的情感词特征；
4. 将以上两种特征与基础的 TFIDF、否定词等特征一起，加入 SVM 进行分类，完成话题相关的情感倾向性分析。

4.5 实验和结果分析

本节对上文提出的多特征融合的话题相关情感倾向性分析方法进行实验和评估。实验过程是将实验数据集经过上述算法处理后得到每条微博针对每一个话题的情感倾向（-1,0,1），然后，按评测指标分别评估系统系能。

4.5.1 语料预处理

本文将 SIGHAN-8Task 2: Topic-Based Chinese Message Polarity Classification 提供的来自新浪、腾讯、网易等微博平台的语料集作为实验数据集，语料中包含 15 个话题:三星 S6、中国人疯抢日本马桶、央行降息、油价、雾霾等，每个话题下有 1000 条微博，其中包含 5 个话题作为训练集、10 个作为测试集，共计 15*1000 条微博。语料中的每一条微博都被标注了-1（负向情感）、0（中性情感）、1（正向情感）。其内容抽样如表 4.2 所示。

表 4.2 话题相关情感倾向性分析语料集

话题	微博	标注
三星 S6	S6edge 的推出再次证明三星是世界的三星，而小米仅仅是中国的小米	1
央行降息	未来的投资机会在于央行降息对估值的提升，超额收益越来越困难	-1
雾霾	以前不喜欢大风天气，现在觉得也不错，因为能吹散雾霾……	-1
12306 验证码	12306 图片验证码遭破解-手机新浪网 http://t.cn/RAJ55I5	0
就业季	高校就业季对我而言真是个苦不堪言的季节[晕][晕][晕][晕]	-1
……	……	……

实验中，分别作了封闭测试和开放测试。在封闭测试中，在 5 个话题的训练集合中，分别在每个话题下选择 800 条加入训练集，200 条加入测试集，最终 4000 条训练，1000 条测试语料；在开放测试中，使用 5 个话题、5*1000 条语料作为训练集，10 个话题、10*1000 条语料作为测试集，最终 5000 条训练，10000 条测试语料。预处理时，按照以下三条规则做简单处理：

规则 1：对句子进行“//”划分，并且使句子顺序翻转。（这样保证句子的转发关系，使后面的句子基于前面的句子进行分析）；

规则 2：对句子中用户名进行删除，即删除“@+用户名”结构；

规则 3：对连续出现多个标点符号，如“。。。。。。”，“!!!!”等，采用第一个标点符号进行替换；

评价指标分别采用微平均(Micro Average)和宏平均(Macro Average)下的准确率 P、召回率 R 和 F 值来表示。传统的准确率和召回率是从单个类别的角度出发，而没有考虑整体的分类精度，而微平均是从整体的角度出发不考虑单个类别的分类精度，其计算公式与公式（2.30）和（2.31）一样，只是其中的 A、B、C、D 指的是所有类别下的统计结果和；宏平均从单独一个类别的分类精度出发，首先在每一个类别上分别计算准确率、召回率和 F 值，然后分别对所有类别的准确率、召回率和 F 值取均值，即为宏平均下的准确率、召回率和 F 值。宏平均下的准确率MacroP计算公式分别如公式（4.4）所示。

$$\text{MacroP} = \frac{1}{n}\sum_{i=1}^{n} P_i \tag{4.4}$$

其中，n为分类类别总数，P_i为第i个类别下的准确率。宏平均下的召回率和 F 值计算方式与宏平均下的准确率计算方式类似，也是分别计算其在每一个类别下的值，最后取平均即可，其公式分别如（4.5）和（4.6）所示。

$$\text{MacroR} = \frac{1}{n}\sum_{i=1}^{n} R_i \tag{4.5}$$

$$\text{MacroF} = \frac{1}{n}\sum_{i=1}^{n} F_i \tag{4.6}$$

4.5.2 LTIGT 算法中不同词图模型构建方法的比较

在上文提出的LTIGT算法中，一个核心点就是话题相关的词图模型构建，如何构建出更准确体现词语间联系的词图模型，对话题相关情感倾向性分析任务是非常重要的。不同的图模型构建方法提取出的特征词是不同的，而特征词的选取对SVM分类，即话题相关的微博情感倾向性分析，有着非常重要的作用。因此，本节主要比较了五种不同的词图模型构建方法，分别采用不同的加权方法和随机跳转概率选择方法进行话题相关情感倾向性分析实验，实验结果如表4.3所示。其中，TR表示原始的TextRank方法，即权值为1，随机跳转概率等于节点个数的倒数；TR+W_co表示原始方法之上，

选用共现信息为词图模型设置权值；TR+W_PCFO表示在原始方法之上，选用本文提出的PCFO为图模型设置权值；TR+RJ_PMI表示在原始方法之上，选用本文提出的基于PMI的随机跳转概率方法；TR+W_CST+RJ_PMI表示全部采用本文提出的PCFO词图模型加权算法和基于PMI的随机跳转概率算法。

表4.3 LTIGT算法中不同词图模型构建方法的比较

Method	Macro Average			Micro Average		
	P(%)	R(%)	F(%)	P(%)	R(%)	F(%)
TR	62.35	57.98	60.09	82.69	82.69	82.69
TR+W_co	63.49	59.76	61.57	82.93	82.93	82.93
TR+W_PCFO	64.65	60.05	62.27	83.36	83.36	83.36
TR+RJ_PMI	64.98	61.15	63.01	83.41	83.41	83.41
TR+W_CST+RJ_PMI	**65.79**	**61.35**	**63.49**	**83.69**	**83.69**	**83.69**

从实验结果可以看出，综合考虑了话题、位置信息的PCFO加权方法，大大提高了实验效果。其主要原因就是因为其对话题题、位置和共现信息的融合。而使用本文提出的基于PMI的随机跳转概率，也在很大程度上提高了实验效果。这主要是因为原始的按照相同的概率跳转到其他节点的随机游走方式，没有考虑话题信息，且易产生局部最优。综上所述，基于局部和全局下词图模型构建的LTIGT算法在话题相关情感倾向性分析任务中表现出了良好的性能。

4.5.3 不同话题词扩展数目和 K-means 聚类个数的比较

上文4.2节主要进行了话题相关情感词的研究，首先通过对话题词扩展获取更多的相关话题词，进而根据依存关系提取更多的话题相关情感词，最后在特征生成部分，对话题相关情感词进行聚类，将每句话中属于每一类别的情感词个数作为向量每一维的特征值。根据以上过程，不难发现不同的话题词扩展数目与不同的K-means聚类个数对话题相关情感倾向性分析任务的实验效果有着巨大的影响。因此，本节使用不同的话题词扩展数目与不同的K-means聚类个数进行实验，图模型构建方法采用4.5.3节中的最好方法，实验效果如图4.4所示。其中，每一条折线表示话题词扩展个数，横轴表示K-means聚类的类别数目。

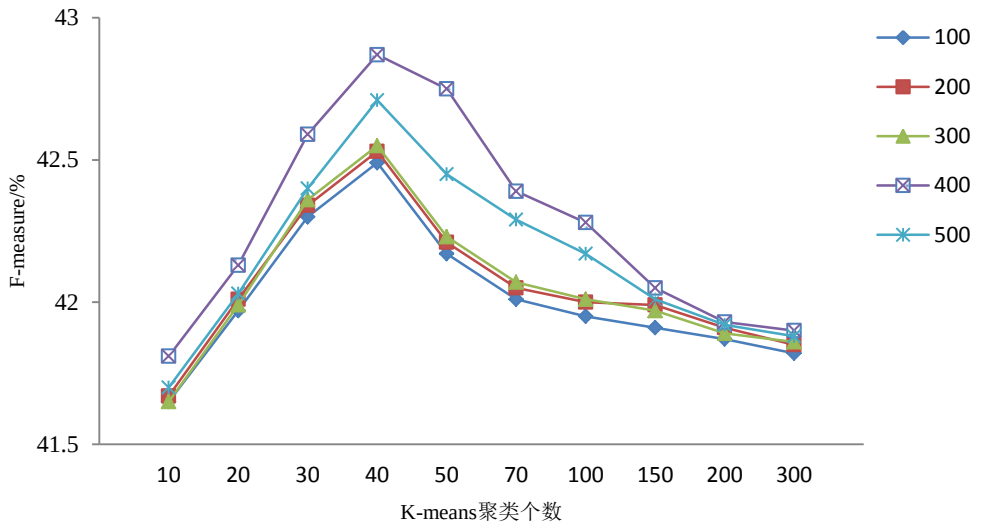

图4.3 不同话题词扩展数目和K-means聚类个数的比较

从实验结果中可以看到，当选择 40 个聚类个数和 400 个扩展话题词时，系统性能最好。观察曲线的走向，发现过大或过小的聚类个数，都不利于话题相关的情感倾向性分析，这主要是因为，过大的聚类个数引入了较多噪音，而较小的聚类个数，则不能充分展示各情感词的细微差别。此外，实验发现当使用 Word Embedding 对话题词扩展时，在扩展词为 400 时，系统达到最优性能，这主要是因为，在话题词扩展的过程中会不可避免的引入噪声，而只有在恰当的阈值内，在系统能够接受的范围内增加，才会提升系统性能。

4.5.4 不同 SVM 特征组合方法的比较

本节主要研究不同的特征组合对话题相关情感倾向性分析的影响，实验采用表 4.1 中 7 种不同的候选特征进行组合，分别在 4.5.1 节所述的封闭测试和开放测试两种规则下进行实验，实验结果如表 4.4 和表 4.5 所示。表格中的各特征描述如下：基础特征（BC）：“EM+NDA+BSL+SBV+VOB+ATT”；统计特征(TFIDF)：TFIDF；基础情感词特征(BSL)：Basic Sentiment lexicon；依存特征(DP)：SBV，VOB，ATT；基于局部和全局词图模型构建的关键词特征（LTIGT）：使用 4.2 节提出的 LTIGT 算法分别计算出语料中每个词的 LTIGT 值，并将其作为衡量词语重要性的标准；基于 Word Embedding 和依存关系的话题相关情感词特征（TSW）：使用 4.3 节的话题相关情感词提取方法提取话题相关情感词，并在特征生成时使用 K-means 聚类。

表 4.4 封闭测试下不同特征组合方法的实验结果

Method	Macro Average			Micro Average		
	P(%)	R(%)	F(%)	P(%)	R(%)	F(%)
BC[68]	55.37	46.91	50.79	81.93	81.93	81.93
TFIDF[70]	61.13	55.72	58.30	82.19	82.19	82.19
LTIGT	62.93	57.15	59.90	82.54	82.54	82.54
BC+ TFIDF	63.78	58.53	60.94	82.75	82.75	82.75
BC+ TFIDF+LTIGT	64.58	59.82	62.11	83.31	83.31	83.31
BC+ TFIDF+LTIGT+TSW	**65.79**	**61.35**	**63.49**	**83.69**	**83.69**	**83.69**

表 4.5 开放测试下不同特征组合方法的实验结果

Method	Macro Average			Micro Average		
	P(%)	R(%)	F(%)	P(%)	R(%)	F(%)
BC[68]	42.74	38.69	40.61	69.05	69.05	69.05
TFIDF[70]	47.92	45.65	46.76	69.97	69.97	69.97
LTIGT	49.03	46.15	47.55	70.34	70.34	70.34
BC+ TFIDF	50.71	47.28	48.93	70.65	70.65	70.65
BC+ TFIDF+LTIGT	51.49	48.32	49.85	71.32	71.32	71.32
BC+ TFIDF+LTIGT+TSW	**52.73**	**49.97**	**51.31**	**71.90**	**71.90**	**71.90**

通过使用不同的特征组合进行话题相关的情感倾向性分析，从实验结果可知，在加入依存、LTIGT和情感词特征之后，系统性能有了很大的提升。这主要是因为以上三个特征，在只考虑统计特征BC、TFIDF的基础上，充分挖掘了句法和语义信息。因此，此实验不仅证明了LTIGT模型的有效性，而且也充分揭示了话题词扩展对话题相关的情感倾向性分析任务的重要性，更进一步说明了多特征融合方法对话题相关情感倾向性分析任务的有效性。

4.6 本章小结

为解决话题相关情感倾向性分析中，一条微博包含多个候选话题和情感词都是基于特定话题的情况，本章提出两种全新的话题相关情感倾向性分析的特征提取算法。首先设计了一种融合局部和全局信息的词图模型构建的 LTIGT 算法，进行关键词特征

提取，其中包括一种全新的 PCFO 加权方法和基于点间互信息的跳转概率选择算法，并将其融入到话题相关的词图模型构建中；然后提出一种话题相关的情感词提取算法，首先基于 Word Embedding 进行话题词扩展，然后通过依存分析找出与扩展后的话题词存在一定依存关系的情感词，并在特征生成时将此部分的话题相关情感词进行 K-means 聚类；最后将此两类特征，融合传统的基础特征一起，选择不同的特征组合进行实验，实验结果充分证明了本章所提算法的有效性，也侧面反映了深度语义挖掘在情感分析任务中的重要性。

在未来工作中，将主要研究以下两个方面：

(1)本章主要使用依存分析挖掘句法信息，在未来工作中，我们将对短语结构分析进行研究，并比较其与依存分析的效果。

(2)本实验将 LTIGT、情感词等特征融入支持向量机 SVM 分类器中进行实验，未来可以考虑其他不同的分类模型，包括一些生成模型等。

第 5 章 基于句法和语义的话题细粒度情感分析原型系统设计

5.1 系统概述

基于句法和语义的话题细粒度情感分析系统共分为四个主要模块：语料预处理模块、情感关键句抽取模块、评价对象识别模块和话题相关的情感倾向性分析模块。

语料预处理模块包含了分词与词性标注、句法结构分析、依存关系分析、语义角色分析四个子模块，用于对语料集合做预处理，经过预处理的语料才能进行下一步的情感关键句抽取、评价对象识别和话题相关的情感倾向性分析。

情感关键句抽取模块，对经语料预处理模块后的语料进行情感关键句抽取；然后将抽取出的情感关键句集合送入评价对象识别模块，进行评价对象的识别；最后将评价对象识别结果与情感关键句抽取结果一并送入话题相关的情感倾向性分析模块，分别以各评价对象为话题，分析每个话题对应的情感倾向，进而完成基于句法和语义的话题细粒度情感分析系统的抽取过程。

系统结构图如图 5.1 所示。以下小结分别介绍每个模块的功能与实现。

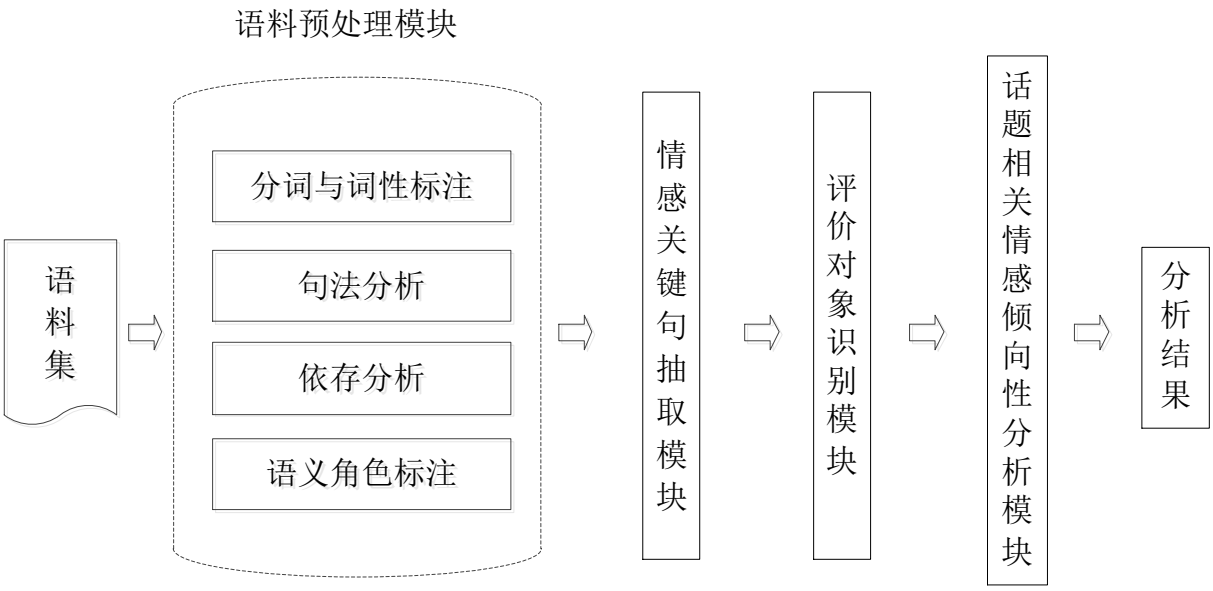

图 5.1 话题细粒度情感分析系统结构图

5.2 语料预处理模块

语料预处理模块主要包含分词与词性标注、句法结构分析、依存关系分析和语义角色标注四个子模块。

5.2.1 分词与词性标注

词语是自然语言处理任务中粒度最小的研究对象，分词往往是进行自然语言处理任务的第一步，其准确率在很大程度上严重影响整个系统的性能。词性标注与分词是相伴产生的，通过对每一个分词结果进行词性分析，确定每一个分词的词性。在本文的细粒度情感倾向性分析系统中，情感关键句抽取、评价对象识别和话题相关的情感倾向性分析三个模块都用到了词和词性的特征，因此，该部分的效果直接影响整个细粒度情感分析模块的性能。因此，本文采用哈工大社会计算与信息检索研究中心研发的“语言技术平台（LTP）”，该系统曾在多次国内外的技术评测中获得优异成绩，例如获得 CoNLL 2009 国际句法和语义分析联合评测的第一名等。

5.2.2 句法结构分析

句法分析是一项分析语言的句法结构、剖析句中各词语的语法功能的自然语言处理任务，可以为上层任务提供支持。在本文的细粒度情感倾向性分析系统中，评价对象识别模块在构建领域词典时，用到了句法结构分析结果，因此，句法结构分析的准确性，直接影响领域词典的构建效果与评价对象识别模块的准确率。因此，在本文的细粒度情感分析系统中，使用的斯坦福大学的句法分析工具。

5.2.3 依存关系分析

依存关系分析方法最早是由 McDonald 提出的，他将依存关系分析问题归结为在一个有向图中寻找最大生成树的问题。依存关系指的是词语的二元关系，其表示的是核心词与依存词之间的依赖关系。依存关系分析通过分析句中的“主谓宾”、“定状补”成分，来挖掘各词语之间的联系。在本文的细粒度情感倾向性分析中，情感关键句抽取、评价对象识别和话题相关的情感倾向性分析三个模块都用到了依存关系分析，因此，依存关系分析的准确率对本系统的影响巨大。所以，经过比对，本文选择使用哈尔滨工业大学信息检索研究室提供的语言技术平台（LTP）中的汉语句法依存分析器，来进行依存关系分析。

5.2.4 语义角色标注

语义角色标注是一种浅层的语义分析技术，主要进行谓词和论元的识别和分类，为句子中的给定谓词，标注出该谓词所对应的论元，即语义角色，如常说的施事、受事、时间、地点、方式等。本文在评价对象识别模块中，主要分析了语义角色标注中的施事和受事者，通过语义角色与词性结合进行评价对象提取，构建评价对象领域词典。由此可见，语义角色标注对本文细粒度的情感倾向性分析系统具有着十分重要的作用。因此，通过比较，本文继续采用哈尔滨工业大学的语言技术平台（LTP）进行语义角色标注。

5.3 情感关键句抽取模块

此模块通过对经分词、词性标注、句法结构和依存分析预处理过的篇章集合进行情感关键句识别，抽取出既能体现篇章情感、又涵盖篇章主题的句子。首先，提出一种基于点间互信息的领域相关情感词典扩展方法，并通过实验验证了其高置信度；然后，提出了一种基于主题模型与词图模型的关键词词典构建方法，并将其作为特征加入情感关键句抽取实验中，验证了其高准确率和高召回率；紧接着，针对依存分析结果，提出了一种依存模板提取算法，并用实验验证了模板的有效性；最后，将以上三种特征与位置特征按一定规则组合成向量，加入 SVM 进行分类，将分类结果的正类，即情感关键句集合（中间结果 1）送往下一个模块（评价对象识别模块）中做进一步处理，其具体结构图如图 5.2 所示。

针对语料预处理结果分四个步骤提取特征：

1. 在 Hownet 和 NTUSD 组成的基础情感词典上运行 PMI 算法，获取领域相关的情感词典，并对词典中的每一个词进行概率统计，获得情感词与其概率得分值；
2. 对篇章进行 LDA 建模，获取篇章中每一个词属于每一个主题的概率，并将其作为跳转概率值保存；同时运行 PCFO 算法，获取边的权值；然后在此基础上构建词图模型，并采用一定的打分函数计算图模型中每个节点的得分，最红获得每一个关键词及其 Rank 得分值；
3. 针对依存分析的结果，首先提取核心词，然后根据依存关系提取出核心词与依存词之间的依存关系，并组合成依存模板，同时进行概率统计，获得每一个依存模板及其概率得分；

4. 以篇章中位句子的得分为 0，计算出打分函数中的参数值，然后使用抛物线函数，分别计算篇章中每个句子的位置得分；
5. 最后将以上四部分特征进行融合，分别选择各对应特征的得分值组合成 SVM 向量，并进行分类，获得最终的情感关键句集合。

图 5.2 情感关键句抽取模块结构图

5.4 评价对象识别模块

此模块通过对经情感关键句抽取模块（中间结果 1）进行评价对象识别，抽取出每个句子的评价对象。首先，通过挖掘词性序列关系、句法、依存、语义角色标注关系，构建候选评价对象词典；然后，使用 Word Embedding 的方法，对以上候选评价对象词典进行扩展，选取排名较高的扩展词作为评价对象扩展词加入评价对象词典；最后，将此扩展后的候选评价对象词典与一些其他词、词性、句法、依存特征一起，使用 CRF 进行预测，标注出每个句子对应的评价对象（中间结果 2），并将其与情感关键句（中间结果 1）一起送往下一个模块（话题相关的情感倾向性分析）中做进一步处理。

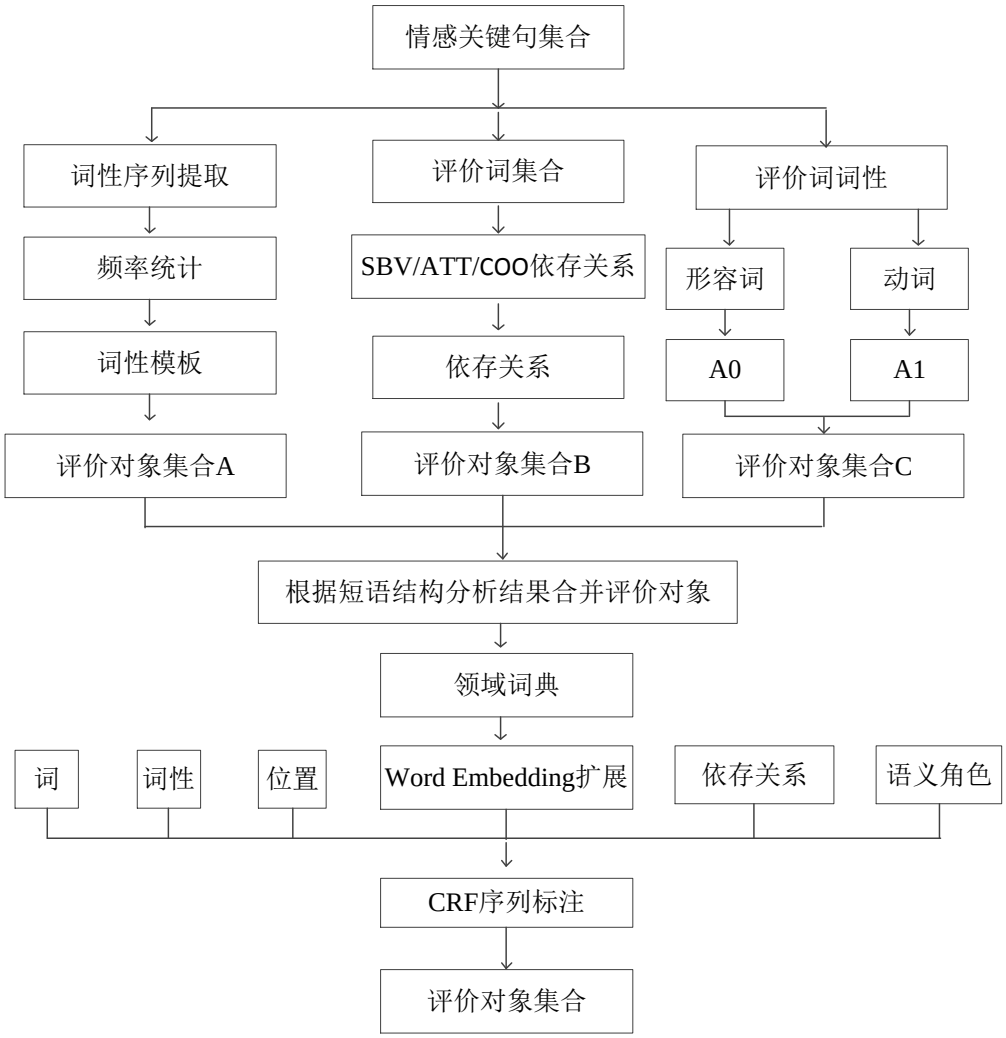

图 5.3 评价对象识别模块结构图

针对情感关键句抽取模块抽取出的情感关键句集合，即中间结果 1，进行评价对象领域词典构建。首先采用 3-gram 的方法对词性序列进行提取，并进行频率统计后，选择出现频率较高的词性模板加入到评价对象集合 A 中去；然后根据 Hownet 与 NTUSD 组合成的评价词集合，分析依存结果中的 SBV、ATT 和 COO 关系，通过其依存关系提取出评级对象集合 B；接着通过分析评价词的词性，分别提取形容词性评价词的施事者（A0）和动词性评价词的受事者（A1）作为评价对象集合 C；最后根据短语结构分析结果对评价对象集合 A、B、C 进行合并，得到评价对象领域词典，然后对其进行 Word Embedding 扩展，与词、词性、位置、依存关系、语义角色一起加入 CRF 进行序列标注，得出最终的评价对象集合，即中间结果 2。

5.5 话题相关的情感倾向性分析模块

此模块以评价对象识别结果（中间结果 2）为话题，分析每个话题对应的情感倾向。首先，通过一种基于局部和全局词图模型的 LTIGT 算法提取关键词特征，并提出一种基于 Word Embedding 的情感词扩展方法提取情感词特征、与 TF_IDF 等基本词特征一起作为特征；然后，通过选择不同的特征维数，寻求最优特征组合形式；最后，按照已选特征将句子分别构造成特征向量的形式，用训练语料训练出的 SVM 分类模型对句子进行分类（-1，0，1），判定结果即为不同话题下的情感倾向（负向，中性，正向）。

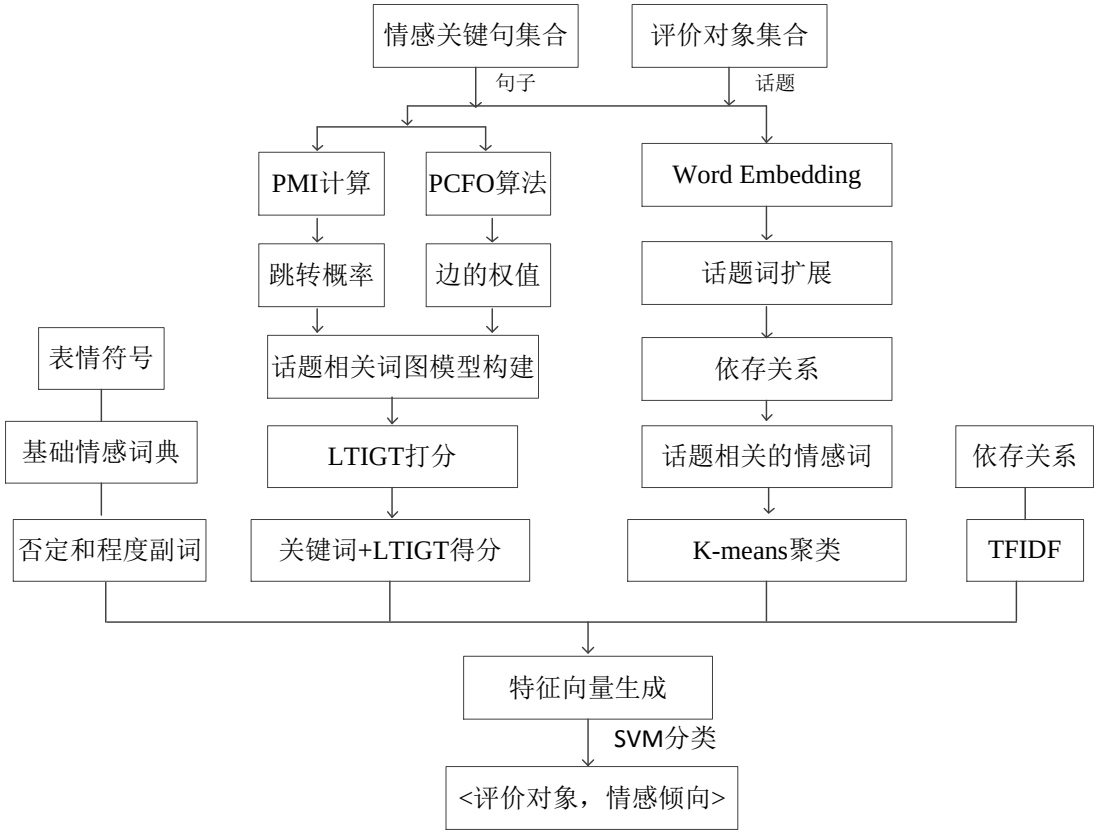

图 5.4 话题相关的情感倾向性分析模块结构图

针对情感关键句抽取和评价词抽取模块的抽取结果，话题相关的情感倾向性分析模块首先将情感关键句集合和评价对象集合分别作为句子和话题输入，然后分别构建关键词特征和话题相关的情感词特征；针对关键词特征，首先通过 PMI 值计算出跳转概率，然后运行 PCFO 算法获得边的权值，最后分别以句子和话题为单位进行话题相关的词图模型构建，并利用 LTIGT 算法对每一个词进行打分，获取最终的关键词与 LTIGT 得分；针对话题相关的情感词特征，首先采用 Word Embedding 进行话题词扩

展，然后基于依存关系抽取出与话题词相依存的情感词，并使用 K-means 对其进行聚类，将属于每一类别的情感词数目作为分类特征；最后将以上两种特征与表情符号、基础情感词典、否定和程度副词、依存关系、TFIDF 特征一起，生成 SVM 特征向量，识别出<评价对象，情感倾向>二元组，完成细粒度的情感分析。

5.6 本章小结

本章介绍了细粒度的情感分析原型系统。系统共分为四个主要模块：语料预处理模块、情感关键句抽取模块、评价对象识别模块和话题相关的情感倾向性分析模块。语料预处理模块是对语料做分词与词性标注、句法分析、依存关系分析和语义角色标注这些预处理；情感关键句抽取模块基于情感词典扩充、关键词典构建、依存模板抽取，识别情感关键句；评价对象识别模块基于句法和语义信息构建候选评价对象词典，使用序列标注模型 CRF 抽取出每个句子对应的评价对象；话题相关的情感倾向性分析模块通过提出一种基于局部和全局信息的 LTIGT 算法，与基于 Word Embedding 的情感词算法，利用 SVM 分类器进行话题相关的情感倾向性分析，最终得到每个话题下的句子的情感倾向。

总结

本文针对细粒度的情感分析技术进行了探索，分别从情感关键句抽取、评价对象识别和话题相关的情感倾向性分析三个层面进行分析，使用 PMI、Word Embedding、词图模型构建的方法深度挖掘隐含句法和语义信息。

1、针对情感关键句提取，本文按照二分类思想来处理情感关键句的抽取问题，通过词汇语义和句法依存两个方面挖掘情感关键句的隐含特征，为情感关键句的抽取奠定基础。本文通过基于点间互信息的情感词典扩充与基于 LDA 和 TextRank 的关键词词典创建，来获取词汇语义信；然后，基于句法依存分析的结果进行依存模板提取，获取依存知识库；并对所有文章中的句子进行规则过滤，然后将情感词、关键词、依存模板和位置特征一起加入支持向量机分类器中，选取不同的分类特征进行分类。实验结果表明，该方法能表现出比传统方法更好的性能。

2、针对评价对象提取，本文提出了一种基于句法和语义的评价对象识别方法，采用序列标注的方法对评价对象进行抽取。在特征提取方面，提出一种融合词性模板、依存结构分析、语义角色标注和短语结构分析的领域词典构建方法，并将此方法构建出的领域词典，与基本的词法特征、相对位置特征、依存关系特征、语义特征一起，选择不同的特征组合，并使用序列标注模型条件随机场 CRF 进行实验，并将机器学习识别结果与领域词典相结合，完成最终的评价对象识别工作。实验结果表明，该评价对象提取方法达到了比较满意的准确率和召回率。

3、针对话题相关的情感倾向性分析，本文提出了一种新的特征并进行验证。在前人的研究基础上，首先提出了一种融合局部和全局信息的 LTIGT 算法，分别在句子级别和话题级别下构建词图模型，并进行迭代排序，得出每个词的得分值；然后提出了一种基于 Word Embedding 的话题词扩展方法和基于依存关系的情感词提取算法；最后利用 LTIGT、扩展后的情感词典、以及前人提出的特征，采用一定的特征构造形式，构造 SVM 分类器，进行情感倾向性分析。实验结果显示，该方法的情感分析效果较以往方法有明显提高。

4、本文构建了一个基于句法和语义的话题细粒度情感分析系统，该系统共分为两个个主要模块：语料预处理模块、细粒度情感分析模块。基于分词、词性标注、句法分析、依存分析、语义角色分析对待处理集经语料进行预处理，之后将其送往细粒度情感分析模块。经三个子模板，首先由情感关键句抽取模块抽取出情感关键句，进

而送往评价对象识别模块抽取评价对象，最后，使用话题相关的情感倾向性分析模块，对评价对象进行情感倾向判定，完成细粒度的情感分析。

本论文主要是基于本人在研究生期间，在海量语言信息和云计算机实验室的科研学术工作，现在虽然论文已经告一段落，但在实际工作中仍然有一些遗憾，需要进一步改进。针对中文观点句抽取的研究，未来可以考虑从以下几个大方向进一步研究：

(1) 可以对现有的依存模板进行同义词扩展，或改进依存关系抽取算法，尝试提出更加具有普遍性的依存关系模板。

(2)实体识别中的指代消解等问题考虑的还不够完善，有待进一步研究和实验，提高系统的准确性。

(3)可以考虑做进一步尝试使用不同分类算法、或选用不同的分类模型对结果进行测试。

参考文献

[1] 杨立公，朱俭，汤世平．文本情感分析综述[J]．计算机应用, 2013, 33(6): 1574-1578.

[2] Liu B. Sentiment analysis and subjectivity [J]. Handbook of natural language processing, 2010, 2: 568.

[3] O'Connor B, Balasubramanyan R, Routledge B R, et al. From Tweets to Polls: Linking Text Sentiment to Public Opinion Time Series[J]. ICWSM, 2010, 11: 122-129.

[4] Asur S, Huberman B A. Predicting the future with social media[C]//Web Intelligence and Intelligent Agent Technology (WI-IAT), 2010 IEEE/WIC/ACM International Conference on. IEEE, 2010, 1: 492-499.

[5] Pang B, Lee L, Vaithyanathan S. Thumbs up? Sentiment Classification using Machine Learning Techniques[J]. Proceedings of Emnlp, 2002:79-86.

[6] Yu H, Hatzivassiloglou V. Towards answering opinion questions: separating facts from opinions and identifying the polarity of opinion sentences[C]//In Proceedings of EMNLP-03. 2003:129-136.

[7] 孙宏纲，陆余良．中文博客主题情感句自动抽取研究[J]．计算机工程与应用, 2006，44(20):165-168.

[8] 杨江，侯敏，王宁．基于主题情感句的汉语评论文倾向性分析[C] .第五届全国青年计算语言学研讨会，2010, 28(2):569-572.

[9] 林政，谭松波，程学旗．基于情感关键句抽取的情感分类研究[J]．计算机研究与发展, 2012, 49(11):2376-2382.

[10] Hu M, Liu B. Mining and summarizing customer reviews[C]//Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining. ACM, 2004:168-177.

[11] Li B, Zhou L, Feng S, et al. A Unified Graph Model for Sentence-Based Opinion Retrieval.[J]. Proceedings of Annual Meeting of the Association for Computational Linguistics, 2010:1367-1375.

[12] Popescu A M, Etzioni O. OPINE: extracting product features and opinions from reviews[C]//Proceedings of HLT/EMNLP on Interactive Demonstrations. Association for Computational Linguistics, 2005:32-33.

[13] 刘鸿宇，赵妍妍，秦兵等．评价对象抽取及其倾向性分析[J]．中文信息学报, 2010, 24(1):84-88.

[14] Christopher Scaffidi, Kevin Bierhoff, Eric Chang, Mikhael Felker, Herman Ng and Chun Jin. Red opal: Product-feature scoring from reviews[C]//Proceedings of the 8th ACM conference on Electronic commerce, 2007: 182-191.

[15] Blei D M, Lafferty J D. Correlated Topic Models.[C]//International Conference on Machine Learning. 2006:113-120.

[16] Ivan Titov and Ryan McDonald. Modeling online reviews with multi-grain topic models[C]//Proceedings of the 17th international conference on World Wide Web, 2008: 111–120.

[17] Qiaozhu Mei, Xu Ling, Matthew Wondra, Hang Su and ChengXiang Zhai. Topic sentiment mixture: Modeling facets and opinions in weblogs[C]//Proceedings of the 16th international conference on World Wide Web, 2007: 171-180.

[18] Li Zhuang, Feng Jing and Xiao-Yan Zhu. Movie review mining and summarization[C]//Proceedings of the 15th ACM international conference on Information and knowledge management, 2006: 43-50.

[19] Jason S Kessler and Nicolas Nicolov. Targeting sentiment expressions through supervised ranking of linguistic configurations[C]//ICWSM, 2009.

[20] Niklas Jakob and Iryna Gurevych. Extracting opinion targets in a single-and cross-domain setting with conditional random fields[C]//Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing, 2010: 1035-1045.

[21] Shoushan Li, Rongyang Wang and Guodong Zhou. Opinion target extraction using a shallow semantic parsing framework[C]//Twenty-Sixth AAAI Conference on Artificial Intelligence, 2012.

[22] 徐冰，赵铁军，王山雨等. 基于浅层句法特征的评价对象抽取研究[J]. 自动化学报, 2011, 37(10):1241-1247.

[23] 王荣洋，鞠久朋，李寿山等. 基于 CRFs 的评价对象抽取特征研究[J]. 中文信息学报, 2012, 26(2):56-61.

[24] 郑敏洁，雷志城，廖祥文等. 中文句子评价对象抽取的特征分析研究[J]. 福州大学学报：自然科学版, 2012, (5):584-590.

[25] Ku L W, Wu T H, Lee L Y, et al. Construction of an evaluation corpus for opinion extraction[J]. Ntcir, 2005:513-520.

[26] Chen J M, Tang Y, Li J G, et al. Community-Based Scholar Recommendation Modeling in Academic Social Network Sites[J]. Lecture Notes in Computer Science, 2014, 8182:325-334.

[27] Hatzivassiloglou V, Wiebe J M. Effects of adjective orientation and gradability on sentence subjectivity[C]//International Conference on Computational Linguistics. 2000:299-305.

[28] Turney P D. et al. Thumbs up or thumbs down? Semantic orientation applied to unsupervised classification of reviews[J]. Proceedings of Annual Meeting of the Association for Computational Linguistics, 2002:417-424.

[29] Zha Z J, Yu J, Tang J, et al. Product Aspect Ranking and Its Applications[J]. IEEE Transactions on Knowledge & Data Engineering, 2014, 26(5):1211-1224.

[30] X. L. M. Z. M. Z. Xiaolong Wang, Furu Wei, Topic sentiment analysis in twitter: a graph-based hashtag sentiment classification approach[C]//CIKM'11, 2011:1031–1040.

[31] Pang B, Lee L, Vaithyanathan S. Thumbs up? Sentiment Classification using Machine Learning Techniques[J]. Proceedings of Emnlp, 2002:79-86.

[32] Dasgupta S, Ng V. Mine the Easy, Classify the Hard: A Semi-Supervised Approach to Automatic Sentiment Classification[C]// Meeting of the Association for Computational Linguistics. 2009:701-709.

[33] Li F, Liu N, Jin H, et al. Incorporating Reviewer and Product Information for Review Rating Prediction.[C]// Twenty-second International Joint Conference on Artificial Intelligence. 2011:1820-1825.

[34] Yi J, Nasukawa T, Bunescu R, et al. Sentiment analyzer: Extracting sentiments about a given topic using natural language processing techniques[C]// In IEEE Intl. Conf. on Data Mining (ICDM. 2003:427-434.

[35] 洪堡特. 洪堡特语言哲学文集[M]// 商务印书馆, 2011.

[36] 马庆株. 结构、语义、表达研究琐议——从相对义、绝对义谈起[J]. 中国语文, 1998, (3):173-180.

[37] Riloff E M, Shepherd J. A Corpus-Based Approach for Building Semantic Lexicons[C]// Eprint Arxiv: Cmp-lg. 1997:117--124.

[38] Hatzivassiloglou V, Mckeown K R. Predicting the Semantic Orientation of Adjectives[J]. Proceedings of the Acl, 2002:174--181.

[39] Turney P D, Littman M L. Measuring Praise and Criticism: Inference of Semantic Orientation from Association[J]. Acm Transactions on Information Systems, 2003, 21(4):315--346.

[40] 朱嫣岚, 闵锦, 周雅倩等. 基于 HowNet 的词汇语义倾向计算[J]. 中文信息学报, 2006, 20(1), 14-20.

[41] 田久乐, 赵蔚. 基于同义词词林的词语相似度计算方法[J]. 吉林大学学报（信息科学版）, 2010,28(6): 602-608.

[42] 王素格, 李德玉, 魏英杰, 宋晓雷. 基于同义词的词汇情感倾向性判别方法[J]. 中文信息学报, 2009,23(5): 68-74.

[43] 徐峻岭, 周毓明, 陈林,等. 基于互信息的无监督特征选择[J]. 计算机研究与发展, 2012, 02 期 (02):372-382.

[44] Kaji N，Kitsuregawa M.Building lexicon for sentiment analysis from massive collection of HTML documents[C]// Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning，2007.

[45] Frank E, Paynter G W, Witten I H, et al. Domain-Specific Keyphrase Extraction[C]// Proceedings of the Sixteenth International Joint Conference on Artificial Intelligence. Morgan Kaufmann Publishers Inc., 1999:668--673.

[46] Turney P D. Learning Algorithms for Keyphrase Extraction[J]. Information Retrieval, 2002, 2(4):303-336.

[47] Blei D M, Ng A Y, Jordan M I. Latent dirichlet allocation[J]. Journal of Machine Learning Research, 2003, 3:2003.

[48] Pasquier, C. Task 5: Single document keyphrase extraction using sentence clustering and Latent Dirichlet Allocation[C]// Proceedings of the 5th international workshop on semantic evaluation, 2010: 154-157.

[49] Liu Z, Huang W, Zheng Y, et al. Automatic Keyphrase Extraction via Topic Decomposition.[C]// Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing. Association for Computational Linguistics, 2010:366-376.

[50] Page L, Brin S, Motwani R, et al. The PageRank Citation Ranking: Bringing Order to the Web[C]// Stanford InfoLab. 1999:1-14.

[51] Mihalcea R, Tarau P, Mihalcea R. TextRank: Bringing Order into Text.[J]. Unt Scholarly Works, 2004:404-411.

[52] 陆俭明. 词的具体意义对句子意思理解的影响[J]. 汉语学习, 2006, (02):1-5.

[53] Hermjakob U. Parsing and question classification for question answering[C]//Proceedings of the workshop on Open-domain question answering-Volume 12. Association for Computational Linguistics, 2001: 1-6.

[54] Corinna Cortes, Vladimir Vapnik. Support-Vector Networks[J]. Machine Learning, 1995, 20：273-297.

[55] Liu, Q., Feng, C., Huang, H.: Emotional Tendency Identification for Micro-blog Topics Based on Multiple Characteristics[C]//Proceedings of the 26th Pacific Asia Conference on Language, Information and Computation, 2012: 280–288.

[56] 韩忠明, 张玉沙, 张慧, 等. 有效的中文微博短文本倾向性分类算法[J]. 计算机应用与软件, 2012, 29(10): 89-93.

[57] 张辰, 冯冲, 刘全超, 师超，黄河燕等. 基于多特征融合的中文比较句识别算法[J]. 中文信息学报, 2013, 27(6): 110-116.

[58] Che W, Li Z, Liu T. LTP: A Chinese Language Technology Platform.[C]// International Conference on Computational Linguistics. 2010:13-16.

[59] 安强强，张蕾. 基于依存树的中文语义角色标注[J]. 计算机工程, 2010: 36(4), 161-163.

[60] Socher R, Bauer J, Manning C D, et al. Parsing with Compositional Vector Grammars[C]// In Proceedings of the ACL conference. 2013:455-465.

[61] Tomas Mikolov, Kai Chen, Greg Corrado& Jeffrey Dean. Efficient Estimation of Word Representations in Vector Space[C]//Proceedings of Workshop at ICLR, 2013:1-12.

[62] D. Tang, F. Wei, N. Yang, M. Zhou, B. Qin, and T. Liu. Learning sentiment-specific word embedding for twitter sentiment classification[C]//Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL), 2014:1555–1565.

[63] Lafferty J, McCallum A, Pereira F C N. Conditional random fields: Probabilistic models for segmenting and labeling sequence data[C] //Proceedings of the 18th International Conference on Machine Learning. 2001:282-289.

[64] 朱艳辉, 徐叶强, 王文华, 等. 中文评论文本观点抽取方法研究[C]. 第三节中文倾向性分析论文集, 2011: 126-135.

[65] Sun H. Shallow parsing: an overview[J]. Contemporary Linguistics, 2000:74-83.

[66] Salton G, Yu C T. On the construction of effective vocabularies for information retrieval[J]. Acm Sigplan Notices, 1975, 10:48-60.

[67] Macqueen J. Some Methods for Classification and Analysis of Multivariate Observations[C]// In 5th Berkeley Symp. Math. Statist. Prob. 1967:281-297.

[68] 张珊，于留宝，胡长军．基于表情图片与情感词的中文微博情感分析[J]．计算机科学，2012, 39(S3):146-148.

[69] 谢丽星，周明，孙茂松．基于层次结构的多策略中文微博情感分析和特征抽取[J]．中文信息学报, 2012, 26(1):73-83.

[70] Martineau J, Finin T. Delta TFIDF: An Improved Feature Space for Sentiment Analysis.[J]. Proc.of Int'l Aaai Conf.on Weblogs & Social Media, 2009:1-4.

攻读硕士学位期间发表的论文

[1] Feng C, Liao C, LiuZ, et al.A Hybrid Method of Sentiment Key Sentence Identification Using Lexical Semantics and Syntactic Dependencies[C]// APWeb 2014 Workshops, SNA, NIS, and IoTS.Web Technologies and Applications, 2014:11-22.

[2] Liao C, Feng C, Yang S, et al. A Hybrid Method of Domain Lexicon Construction for Opinion Targets Extraction Using Syntax and Semantics[C]// Social Media Processing. Springer Berlin Heidelberg, 2014:108-116.

[3] Liao C, Feng C, Yang S, et al.Topic-Based Chinese Message Polarity ClassificationSystem at SIGHAN8-Task2[C]//Proceedings of the Eighth SIGHAN Workshop on Chinese Language Processing (ACL-SIGHAN-8), 2015:158-163.

[4] 冯冲, 廖纯, 刘至润, 黄河燕. "基于词汇语义和句法依存的情感关键句识别", 电子学报. 录用.

致 谢

时光荏苒，转眼又到了离别的季节。回首这段弥足珍贵的研究生时光，除了学习到丰富的知识，提高了自己的能力之外，更重要的是感恩，是有幸结识了这么多知识渊博的师长和勤奋好学的同学们，他们对我的影响足以令我受益一生。

首先，我要感谢我的导师冯冲老师，感谢他一直鞭策着我，才能让我的研究生生活如此充实，并获得了一定的成就和种种荣誉。科研工作中，他严谨认真，用他渊博的学术知识在学术上给我以很大帮助，并鼓励我参加多种科研项目，很大程度上提升了我的理论和实践能力，让我在以后的学习工作中受益颇丰；生活中，他为人平和，百忙之中不忘关心和鼓励着我，让我倍感温暖。每当遇到学习或者生活上的困难时，冯老师总是不遗余力地给与帮助。与此同时，他还以他丰富的人生阅历给了我许多指导性意见，帮助我解决了许多难题，让我在人生道路上不再彷徨无措。

其次，我要感谢黄河燕老师、史树敏老师、毛先领老师、辛欣老师、鉴萍老师和郭宇航老师。感谢黄老师，每当遇到困难时，黄老师总是不厌其烦地指导我，帮助我找到解决问题的方案，并着力培养我解决问题的能力。感谢史老师的细心培养，史老师就像姐姐一样关心我们的学习生活，让我们在紧张的工作学习中感受到阵阵温暖。感谢毛先领老师、辛欣老师、鉴萍老师和郭宇航老师，四位老师年轻有活力，对待学术工作认真负责，在学术探讨上给了我非常大的帮助，是我学习的榜样。

我还要感谢亲爱的同学们。感谢杨森，陪我度过了快乐的两年时光，不管在学术还是生活上都给了我很大的帮助。感谢刘茜，一直陪伴在我左右，在我生活迷茫的时候总能给我力量。感谢刘全超师兄和魏骁驰师兄在毕业设计开展过程中给予的宝贵意见。感谢王天航、刘至润、刘敏，感谢实验室所有同学们，我们一起探讨学术、积极向上的日子让我永生难忘。

感谢参考文献中的所有作者们，在他们的前期研究基础之上，我的毕业设计才能够顺利完成。

再次，感谢我的家人，感谢他们长期以来对我的培育、鼓励和支持，让我不畏艰难，用于接收挑战。家人是我我坚实的后盾，无论身在何方，家人总能给予我无限力量，让我的前行之路不再孤单。

最后，感谢所有支持、鼓励、帮助过我的老师、同学、朋友和亲人，他们无私的帮助让我受益颇丰，祝福大家在以后的生活中走得更远。