分类号	TP391.1	密级	公开
UDC	004	学位论文编号	D-10617-308-(2017)-02077

重庆邮电大学硕士学位论文

中文题目	基于产品评论的 细粒度情感分析研究
英文题目	**Research on Fine-grained Sentiment Analysis with Product Reviews**
学　　号	**S140201081**
姓　　名	王俊霞
学位类别	工学硕士
学科专业	计算机科学与技术
指导教师	张璞　副教授
完成日期	**2017 年 5 月 21 日**

独 创 性 声 明

本人声明所呈交的学位论文是本人在导师指导下进行的研究工作及取得的研究成果。尽我所知，除了文中特别加以标注和致谢的地方外，论文中不包含他人已经发表或撰写过的研究成果，也不包含为获得 重庆邮电大学 或其他单位的学位或证书而使用过的材料。与我一同工作的人员对本文研究做出的贡献均已在论文中作了明确的说明并致以谢意。

作者签名：刘俊霞　　日期：2017 年 5 月 21 日

学位论文版权使用授权书

本人完全了解 重庆邮电大学 有权保留、使用学位论文纸质版和电子版的规定，即学校有权向国家有关部门或机构送交论文，允许论文被查阅和借阅等。本人授权 重庆邮电大学 可以公布本学位论文的全部或部分内容，可编入有关数据库或信息系统进行检索、分析或评价，可以采用影印、缩印、扫描或拷贝等复制手段保存、汇编本学位论文。

（注：保密的学位论文在解密后适用本授权书。）

作者签名：刘俊霞　　导师签名：张璞

日期：2017 年 5 月 21 日　　日期：2017 年 5 月 21 日

摘要

随着网络信息化的发展，电商网站上出现大量产品评论，如何快速整理、归纳这些海量信息，成为当前迫切需要解决的问题。情感分析正是基于这一需求，通过自动分析、整理和归纳，挖掘出用户的情感倾向。细粒度情感分析作为情感分析的主要研究内容，不再对文本进行整体情感倾向性判断，旨在挖掘产品特征、情感词及对应情感倾向等要素。通过细粒度情感分析，可以发现用户对产品局部细节的满意程度，对改进产品、发掘潜在用户以及为用户提供购买依据起着极其重要的作用。

基于产品评论的细粒度情感分析研究中，产品特征抽取以及情感词典的构建是最主要的研究任务。本文针对产品特征抽取方法中领域移植性差、人工标注工作量大、无监督学习方法存在的准确率较低以及情感词典构建方法中情感词典的准确率不高、覆盖率低、依赖语义知识库等问题进行研究。具体而言，本文主要工作如下：

1. 基于产品特征在不同领域分布的差异，针对现有产品特征抽取方法中存在的领域移植性较差以及人工标注工作量大等问题，提出了一种无监督的基于领域相关性的产品特征抽取方法。该方法首先引入互信息确定名词短语，然后根据一定的句法规则抽取候选特征，再利用两种不同领域语料的差异性，得到候选特征的领域相关性值，进一步确定产品特征。实验结果表明，该方法可以有效提高产品特征抽取的准确率。

2. 针对传统方法中自动构建的情感词典中所存在的准确率不高、覆盖率低以及依赖语义知识库等问题，提出了基于标签传播的情感词典构建方法。该方法首先选取一定数量的情感种子词，然后利用 Word2Vec 抽取和种子词相似度高的词语；同时，通过依存句法分析抽取和种子词具有连词关系的词语；最后通过标签传播算法确定词语的极性，并得到构建的情感词典。实验结果表明，该方法可以得到准确率较高的情感词典。

3. 设计并实现了一个基于产品评论的细粒度情感分析原型系统。该系统可以从产品评论中自动抽取产品特征，并将用户的情感倾向性以图形化界面展示出来。

关键词：细粒度情感分析，产品特征抽取，情感词典构建，连词关系，标签传播

Abstract

With the development of Internet information, there comes out lots of product reviews in e-commerce. How to organize and summarize these massive information quickly become an urgent problem. Sentiment analysis is based on this demand, through analyzing, sorting and summarizing information automatically to find users' sentiment tendencies. Fine-grained sentiment analysis as the main content of sentiment analysis, the purpose is to extract opinion targets, sentiment words and the sentiment tendencies and so on. Through the fine-grained sentiment analysis, we can find users' satisfaction with products' details, which can play an important role to improve products' quality and find potential users, and also can provide purchasing basis for users.

In the study of fine-grained sentiment analysis with product reviews, the extraction of opinion targets and the expansion of sentiment lexicon are the most important tasks. In this thesis, we focus on the problems such as poor portability, a great deal of manual labeling effort, low accuracy in unsupervised learning approaches for the extraction of opinion targets, low accuracy, low coverage of sentiment words and rely on semantic knowledge base in the construction of sentiment lexicon. Specifically, the main work of this thesis is as follows:

1. Based on the difference of opinion target statistics across different corpora, aming at the problems of poor portability and a great deal of manual labeling effort in opinion target extraction, an unsupervised approach to opinion target extraction using domain relevance is proposed. This approach first introduces point mutual information to get noun phrases, and extracts candidate targets according to a pre-defined set of syntactic rules. Then use the difference of opinion target statistics across two corpora to get the value of domain relevance for each candidate target. Finally we find out the opinion targets. Experimental results show that the approach can effectively improve the accuracy of opinion target extraction.

2. Aming at the problems of low accuracy, low coverage of sentiment words and rely on semantic knowledge base in the traditional approaches, a method of constructing sentiment lexicon based on label propagation is proposed. This approach first selects a certain number of positive and negative sentiment seed words, then uses Word2Vec to train word embeddings to find out the words that have high similarity with seed words,

and then finds out the words that have conjunctive relations with the seed words through the analysis of universal dependencies. Finally we use the label propagation algorithm to determine the polarities of words and get sentiment lexicon. Experimental results show that the approach can construct a more accurate sentiment lexicon.

3. Design and implement a fine-grained sentiment analysis protype. The protype can automatically extract the opinion targets from the product reviews, and user's sentiment polarity is displayed in a graphical interface.

Keywords: fine-grained sentiment analysis, opinion target extraction, sentiment lexicon construction, conjunctive relations, label propagation

目录

图录

表录

第 1 章 绪论

1.1 研究背景及意义

随着互联网的迅猛发展，越来越多的用户利用网络渠道获取各种信息。根据中国互联网信息中心报告[1]调查显示，截至 2016 年 12 月，我国网民规模达 7.31 亿，全年共计新增 4299 万网民。互联网的普及率已经达到了 53.2%，相比 2015 年底，我国网民提升了 2.9 个百分点，超过了全球平均水平 3.1 个百分点。图 1.1 显示了我国网民规模近十年的变化情况。

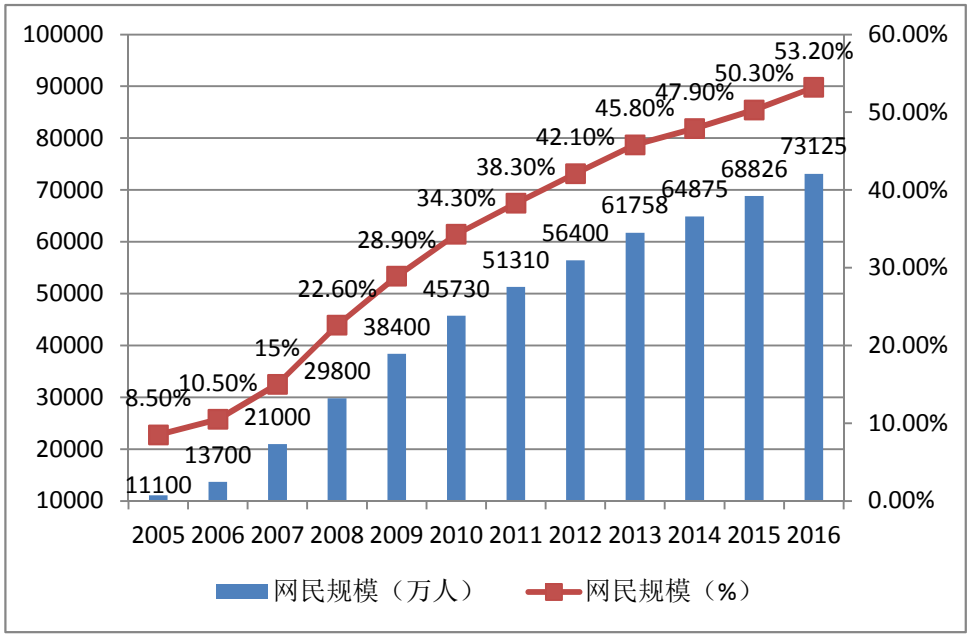

图 1.1 我国网民规模以及互联网普及率

网络已经融入到了人们生活的方方面面，不断改变着人们的传统生活以及消费方式，成为各行业发展以及人们生活息息相关、密不可分的一部分。网络的普及促使电子商务快速发展，使 B2C 交易规模所占比例不断提升，线下线上的融合得到了进一步加深，越来越多的用户倾向于在网站上购买自己所需要的产品。数据显示，随着网民增长速度逐渐变缓，网上购物的用户依然呈快速增长的趋势[1]。截至 2016 年 12 月，网上购物的用户达到了 4.67 亿，占网民比例的 63.8%。图 1.2 显示了 2015 年到 2016 年网络购物用户的规模。

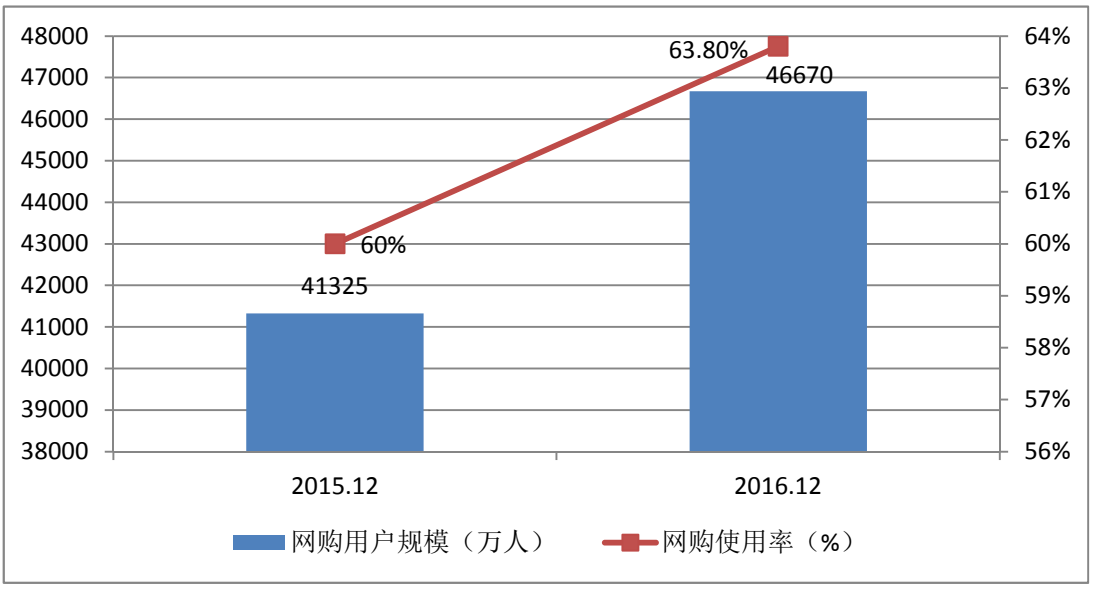

图 1.2 2015.12-2016.12 我国网购用户规模及使用率

由于互联网的开放性，已购买产品的用户可以自由发布言论和观点，导致各电商网站上产生了大量带有情感色彩的评论信息。而面对电子商务网站大量产品信息，用户在购买某产品前，除了查看商家对产品的介绍，更需要查看该产品的评论信息，进一步了解用户对产品品质的真实反映。但是大量且分散的产品评论使用户不能快速定位自己所需要的信息，往往为了解一个产品要去浏览很多网站进行比较分析，同时还要从众多评价中耗费很多精力来获取自己需要的有用信息。因此，查看产品评论成为用户网上购物最为头疼的一个问题。对商家而言，在大量多样化及分散的评论下，商家很难了解自己产品特征的优缺点，同时也很难发现其他商家产品特征的优点和不足，不利于商家改善自己产品的不足之处。

为了向用户清晰地展示产品评论中产品相关信息，同时为商家提供产品设计依据和其他商家的竞争情报，对用户的需求和产品的改进方向做出有效反应，就需要对海量评论信息进行分析、处理，从中挖掘、归纳出具有价值的观点信息[2]。因此，文本情感分析技术[3,4]应运而生。文本情感分析，也称意见挖掘，其目的是分析人们对实体(如产品、服务、属性等)的意见、情绪和态度[5]。早期的文本情感分析主要面向篇章级文本或句子级文本来判定相应的情感极性。

文本情感分析具有很大的研究价值，受到了国内外众多研究者的重视。近年来，随着应用的深入，用户不再满足于只了解评论文本的整体观点倾向，而是提出了更高的需求，想要对评论信息有更细粒度地了解[6]。细粒度情感分析也称为情感倾向

信息挖掘，其任务是通过分析文章、段落或者句子，从而得到文本中的产品特征、评价词语以及情感极性等评价的关键要素[7]。产品特征也称作评价对象，是指某段评论中所讨论的主题，主要是指产品的部件、属性以及性能等对象，它能反映产品对顾客的吸引力；评价词也称作极性词、情感词，是指对产品特征有情感倾向的描述词语。如在评论“这个手机待机时间长，外壳很漂亮，就是屏幕有点小，不过整体还是不错的”中，“待机时间”、“外壳”以及“屏幕”等是产品特征，“长”、“漂亮”、“小”以及“不错”等是评价词。如果想要了解该评论中对“待机时间”、“外壳”以及“屏幕”等产品特征的评价极性，就需要对评论进行细粒度的情感分析。

对产品评论进行细粒度的情感分析研究，不仅能够得到产品评论的倾向，还可以从更多细粒度角度得到分析结果，不仅具有重要的学术研究价值，还具有极大的社会价值和商业价值。对消费者来说，可以更快速地了解其他用户对产品的评论，在节省大量时间和精力的前提下，可以得到商品属性粒度级别的购买依据。对商家来说，产品评论通常蕴含着大量消费者关心的问题以及对产品的情感倾向，通过了解这些信息，可以更加明确产品的优势和不足，同时也可以更好地了解消费者的购买需求，这样不仅可以为商家提供产品设计依据和其它企业的竞争情报，还能对用户需求和产品改进方向做出有效反应，提高企业竞争力。

综上所述，对产品评论进行细粒度情感分析研究具有重要的研究意义和应用价值。

1.2 国内外研究现状

在产品评论的细粒度情感分析研究中，产品特征抽取以及情感词典的构建是最基础也是最重要的任务。本节对产品特征抽取方法以及情感词典构建方法的国内外研究现状进行阐述，并总结了存在的不足之处。

1.2.1 产品特征抽取

产品特征是指产品的部件、属性以及性能等对象。大部分产品特征是显式的，但有部分是隐式的。如在产品评论“这个手机很贵”中，评价词“贵”形容的是产

品特征“价格”，而“价格”并没有出现在评论中，所以它是隐式的。本文研究的是显式产品特征的抽取。

已有研究工作中，对产品特征抽取的方法主要可以分为基于名词和名词短语的抽取方法、基于评价词和产品特征关系的抽取方法、基于主题模型的抽取方法以及基于有监督学习[8]的抽取方法。本文对这四类方法进行介绍：

1. 基于名词和名词短语的产品特征抽取方法

该方法主要通过从大量的语料中找出名词和名词短语作为产品特征。Hu 等人[9]通过词性标注识别名词和名词短语，同时统计它们出现的频率，最后将出现频率大于某阈值的名词和名词短语作为产品特征。尽管该方法较为简单，但是比较有效，可以抽取出用户经常评论的产品特征。A. M. Popescu 等人[10]考虑到抽取出的一些名词短语并不是产品特征，因此在 Hu 的基础上做了改进，利用点对互信息(Pointwise Mutual Information，简称 PMI)将该值太低的名词短语过滤掉。S. Blair-Goldensohn 等人[11]主要考虑那些具有情感倾向句子中的名词和名词短语，使用了一些过滤器来过滤掉非特征词语。Long 等人[12]基于频率和词语之间的距离抽取产品特征，他们首先基于词频找到核心的产品特征，然后利用词语之间的距离找出和它关联性较大的其它产品特征。邱培超等人[13]首先提取名词和名词短语作为候选特征，同时通过观察语料中的词性组合提取出几种模式，再根据这些模式提取候选特征，进一步将出现频率大于某阈值的候选特征作为产品特征。

2. 基于评价词和产品特征关系的产品特征抽取方法

该方法的思想是：通常，对于每条产品评论，评价词都有相应评价的产品特征。所以，依据评价词和产品特征的这种关系可以抽取产品特征。Hu 等人[9]基于这种思想抽取不经常出现的产品特征。其方法是：一个评价词可以修饰多个产品特征，如果在一条评论中有评价词，但是没有找到出现频率较多的产品特征，则将距离该评价词最近的名词或名词短语作为产品特征，最终证实了这种方法的有效性。S. Somasundaran 等人[14]首先根据依存句法分析器来识别产品特征的依赖关系，进一步抽取产品特征。随后，Qiu 等人[15]将依存分析利用在双向传播方法中，进一步抽取产品特征和情感词。唐晓波等人[16]根据句法依存分析结果来确定特征词之间的语义关联。作者首先利用特征词的依存方向来确定该词的关联方向，然后利用改进的 PageRank 算法来计算节点的重要程度，继而进行特征词的提取。赵妍妍等人[17]提出

了根据句法路径来抽取产品特征和评价词的方法。通过自动获得句法路径来得到产品特征和评价词语的修饰关系，进一步根据句法路径距离的计算，对情感评价单元的抽取系统进行性能提升。最后证明了该实验算法的有效性。L. Ferreira 等人[18]基于亚马逊的数据比较了两种产品特征抽取的方法。第一种方法通过应用一组词性标注模式来识别产品特征，进而再利用似然比检验筛选掉与主题无关的产品特征。第二种方法应用关联规则抽取产品特征，进而基于情感词存在的启发式教育法来筛选产品特征。最终实验结果表明两种方法各有优势和不足，也分析了各自适用的情况。K. Khan 等人[19]在 Liliana Ferreira 等人研究的基础上，对利用似然比技术进行产品特征抽取做了进一步的优化。通过 WordNet 将被过滤掉的候选特征进行重新筛选，该方法取得了较好的的效果。N. D. Boer 等人[20]基于统计和句法关系来提取特征，最后和三种已有的方法进行比较，证明了方法的有效性。郁亚辉等人[21]利用情感词和产品特征之间的语法依赖关系进行双向传播，抽取情感词和产品特征。该方法不需要大量数据的标注就可以实现产品特征的抽取。

3. 基于主题模型的产品特征抽取方法

目前，已有很多方法运用主题模型来从大量的文本中发现主题。基于主题的模型主要基于两个基本的模型：pLSA[22]和 LDA[23]。

Mei 等人[24]提出了一个联合模型。作者首先基于模型 pLSA 和 LDA 构建了主题模型、积极情感词模型和消极情感词模型，然后在此基础上建立了一个产品特征-情感词的混合模型来抽取产品特征。Zhao 等人[25]提出了一个最大熵和 LDA 结合的混合模型，该模型可以将产品特征和情感词同时抽取出来，然后利用句法结构将产品特征和情感词分离开。其中，最大熵用标记的数据来学习变量的参数。C. Sauper 等人[26]提出了一个用于短评论中抽取产品特征的联合模型，该模型能够将隐马尔可夫模型和主题模型结合。刘羽等人[27]通过分析产品的相关评论信息，构建产品信息-整体评论-细节评论三层挖掘模型，这个模型根据不同层特点的不同，选择不同的方法抽取产品特征。最后，将三层模型得到的实验结果与只用一层模型得到的实验结果进行对比，验证了三层模型方法的有效性。

4. 基于有监督学习的产品特征抽取方法

有监督学习方法需要数据集，数据集一般由人工进行标注。一般的做法是通过机器学习的方法来训练和构建模型，即利用学习函数对这些训练数据中的实体和其

中的关系进行学习，进一步确定模型中的未知因素。然后通过得到的标注信息确定抽取规则，再分析新数据。Zhang 等人[28]选择词级特征，将领域词典的知识加入模型训练中，采用条件随机场(CRF)模型抽取产品特征和属性特征。王山雨等人[29]采用条件随机场模型和最大熵模型在产品特征抽取任务中进行比较分析，同时分析了加入词性、浅层句法等相关特征对产品特征抽取的影响。刘丽等人[30]结合 CRF 和语法树剪枝的方法识别情感单元，和传统粗粒度方法相比取得了更好的性能。王荣洋等人[31]利用 CRF 模型，研究了产品特征抽取中多类特征的区别，将特征总结为四大类别，分别是词法、依存关系、相对位置以及语义，并且，作者重点利用语义角色对新的特征进行标注。实验结果证实了对特征标注语义角色能够较好地抽取产品特征。接着，戴敏、王荣洋等人[32]又利用基于 CRF 模型的监督学习方法实现了对英文产品特征的抽取。文中为了发现情感词和产品特征之间的关系，引入了句法分析，提高了产品特征抽取的性能。徐冰等人[33]也使用类似的方法，在过程中利用情感词表引入启发式位置信息，再通过浅层句法分析引入浅层句法的信息，提高了识别特征的准确率。张玥等人[34]将依存句法树和 CRF 结合起来抽取产品特征和评价词语，使 CRF 在标注细粒度情感要素时对语义依赖较长距离的情感要素同样适应。丁晟春等人[35]将 CRFs 和本体特征结合起来实现产品特征的抽取，提高了算法的准确率。但是该方法也有一定的不足，如本体构建不够完善等。

总体来说，针对产品特征抽取方法的研究取得了较多成果；但是也存在一些不足，如有监督学习方法不仅需要大量的人工标注，对于不同语料还需要重新训练模型，使得领域移植性较差；而无监督方法存在准确率较低的问题等。因此，在现有工作基础上，为了减少人工标注量、提高领域移植性以及进一步提高无监督方法中产品特征抽取的准确率，本文提出了基于领域相关性的产品特征抽取方法。

1.2.2 情感词典的构建

情感词也称为评价词或者极性词，一般是指带有情感倾向性的词语，情感词典的全面性以及准确性会严重影响情感分析的准确率。情感词典不仅可以准确判断词语的极性，还可以有效辅助句子级文本以及篇章级文本的情感分析。因此，构建一个准确率较高的情感词典，具有极其重要的意义。

目前，大部分通用情感词典是通过人工方法进行构建的。主要通过阅读大量的相关语料，或者利用现有的词典，人工总结出具有情感倾向的词并标注其极性构成情感词典。虽然人工构建的情感词典准确率较高，但在实际应用中，人工情感词典的构建不仅需要花费大量人力和物力，而且领域适应性差。因此，更多的学者们聚焦于自动构建情感词典的工作。现有情感词典的自动构建方法主要分为三类：基于知识库的方法、基于语料库的方法以及知识库和语料库相结合的方法[36]。

1. 基于知识库的方法

A. Neviarouskaya 等人[37]基于 WordNet 的同义词关系进行情感词典的扩充。首先人工构建少量积极和消极种子词，然后通过 WordNet 中词语的同义词关系来扩充情感词典。但是由于词语与词语之间复杂的关系，有些词语经过迭代后，可能会得到该词语的反义词，所以情感词典有一定的误差。A. Hassan 等人[38]认为，两个词语的意思越相近，它们之间就需要越少的次数进行同义词迭代。因此，作者首先选取部分词语作为积极种子词集合和消极种子词集合；然后利用 WordNet 构建了一幅词语间的关系图，再分别计算词语移动到积极种子词集合和消极种子词集合的平均移动次数，词语距离哪个种子词集合的平均移动次数少，该词语就和该种子词集合的极性相同。

2. 基于语料库的方法

语料和语义知识库相比，不仅数量充裕而且更容易得到。因此，利用基于语料库的方法来构建情感词典不仅可以节省大量的人力、物力，而且在不同领域的语料上可以得到领域特定的情感词典，更加具有实用意义。H. Kanayama 等人[39]利用一定的规则模式来抽取情感词，作者认为连续出现的单词具有相同的极性，只有当出现转折关系的连词时，情感极性会发生改变，由此来确定情感词的极性。但是由于规则模式的多样性，该方法可扩展性差、覆盖率低、人工编写工作量大。Huang 等人[40]通过词语间的连词关系来判断词语的极性，作者首先结合词语的否定形式构建了一个情感极性的约束矩阵，然后通过点对互信息值来判断词语的情感极性。这类方法对于主观性较强而且语句之间有情感变化的评论有较强的适用性，对主观性较弱的评论适用性较差。郁亚辉等人[41]利用词语之间的共现信息以及标签传播算法构造情感词典，最终取得了较理想的效果。

3. 知识库与语料库相结合的方法

基于语料库的情感词典扩充方法可以通过对大量的语料进行无监督学习，进而获得相应的情感词典。相对基于知识库的方法来说，还存在准确率较低的问题。所以，很多学者把语料库和知识库方法相结合扩充情感词典。由于知识库主要包含了词语之间的语义关系，而语料库侧重于语料中词语之间的关系，如共同出现的信息、并列关系和转折关系等。因此，通过知识库来提供种子词集，再利用语料库中并列、转折以及其它信息等，确定扩充得到的情感词极性，可以得到一个比较准确的情感词典。如 Peng 等人[42]提出了约束对称非负矩阵分解算法，该方法首先选取了一定的种子词，然后根据 WordNet 中词语之间的关系来对种子词进行扩充；之后作者利用语料中的并列关系和转折关系将词语提取出来，并构建了一张关系图和一个限制矩阵；最后使用限制的非负矩阵分解算法来判断词语的极性，构成最终的情感词典。实验表明了利用 WordNet 和语料库相结合的方法比只使用一种方法扩充得到的情感词典准确率高。D. Rao 等人[43]利用基于图的半监督的标签传播算法对扩充得到的情感词进行情感极性判断，图中每个节点表示要确定极性的词语。作者选取三种不同语言进行实验，研究表明该方法相比基线方法准确率有显著的提高。杨阳等人[44]使用谷歌开源工具 Word2Vec 训练词向量，选取的种子词为大连理工大学信息检索研究室提供的情感词汇本体，然后作者将语料中和种子词相似度大于某一阈值的情感词作为候选情感词，再计算候选情感词和种子词的余弦相似度，进而得到候选情感词分别属于褒贬词典的概率，最后判断词语的情感倾向。

综上所述，已有情感词典构建方法借助各种语义或语料信息来构建情感词典。但是这些方法仍存在一定的问题，如情感词极性不够准确，情感词典构建依赖于语义知识库、覆盖率较低以及领域移植性差等。因此，在业界现有工作基础上，进一步研究更准确且更完善的情感词典构建方法，具有重要的研究意义。

1.3 论文主要工作

本文的研究内容主要包括以下几个方面：

1. 产品特征的抽取：通过研究已有学者所提出的产品特征抽取方法，分析产品特征抽取中各方法引入噪音的原因，本文提出了一种新颖的产品特征抽取方法。该方法首先基于一定的句法规则抽取候选特征，再利用领域相关性和领域无关性两种

语料来计算候选特征的领域相关性值，然后利用领域相关性值剔除无关候选特征，最后通过设计相应的实验，验证了该方法的有效性。

2. 情感词典的构建：在分析已有情感词典的构建过程后，本文提出利用 Word2Vec 训练词向量，得到词语之间的相似度，同时利用依存句法分析语料中词语的关系，对种子词进行扩充。然后利用标签传播算法判断扩充得到的词语极性，并得到情感词典。最后通过一系列实验对比，证明了本文方法能有效提高情感词典的准确率。

3. 系统展示：在细粒度情感分析研究工作的基础上，本文设计了一个原型系统。该系统可以展示各产品的产品特征及其相应的情感极性。

1.4 论文组织结构

本文针对产品评论进行细粒度的情感分析，研究了已有产品特征抽取方法以及情感词典构建方法，提出了基于领域相关性的产品特征抽取方法和基于标签传播的情感词典构建方法。本文共分为六章，具体的组织结构如图 1.3 所示。

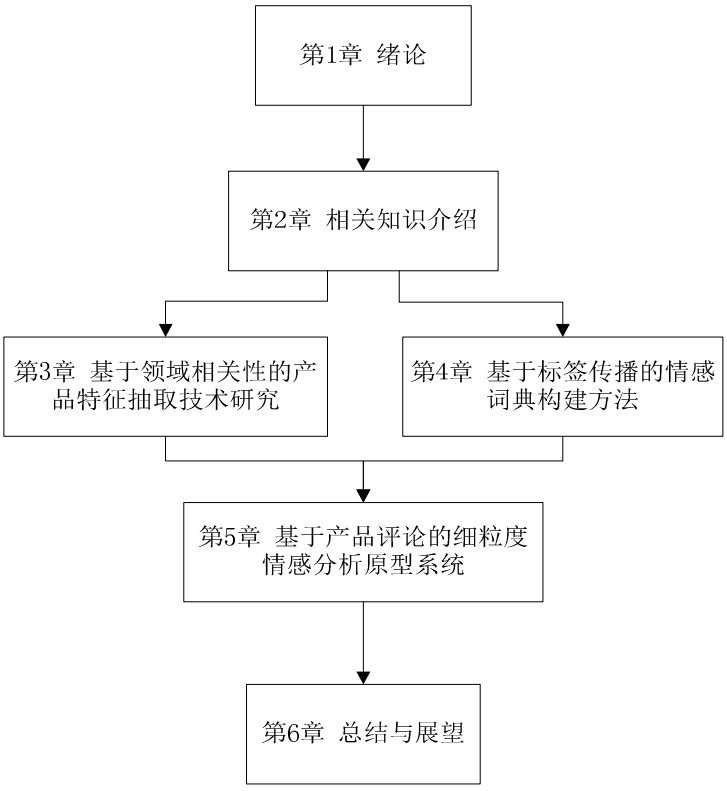

图 1.3 本文组织结构图

第 1 章：绪论。简述了基于产品评论的细粒度情感分析的研究背景、目的及意义，并阐述了产品特征抽取以及情感词典构建的研究现状和本文的主要工作。

第 2 章：相关知识介绍。主要介绍了本文进行细粒度情感分析研究所涉及的相关基础知识，包括分词、词性标注、句法分析、语言技术平台 LTP、Word2Vec 以及本文实验的评价标准等。

第 3 章：基于领域相关性的产品特征抽取技术研究。详细阐述了本文所提出的产品特征抽取方法，该方法首先进行候选特征的抽取，然后对候选特征进行筛选，最后给出了该方法进行的实验并对结果进行了分析。

第 4 章：基于标签传播的情感词典构建方法。详细阐述了情感词典构建以及极性判断的方法。该方法首先利用 Word2Vec 和连词关系对种子词进行候选情感词的扩充，然后利用标签传播算法对候选情感词进行多次迭代确定其极性并得到情感词典。最后本文对该方法进行了实验，并对实验结果做了进一步地分析。

第 5 章：基于产品评论的细粒度情感分析原型系统。该系统主要将产品特征的优势和不足以图形化界面清晰地展现出来。

第 6 章：总结与展望。主要对本文的工作进行总结和对未来工作进行展望。

第 2 章 相关知识介绍

2.1 引言

基于产品评论的细粒度情感分析研究的相关知识包括分词、词性标注、句法分析、哈工大语言技术平台 LTP、Word2Vec、WordNet 以及实验评价标准等。本章将对以上这几个相关知识进行阐述分析。

2.2 语料预处理

在文本情感分析之前，需要对语料进行预处理。下面主要介绍中文分词技术、词性标注以及句法分析。

2.2.1 中文分词技术

尽管中文与英文都是人类的自然语言，但是在文本表示方式上两者还是有很多区别。英文使用空格分隔每个单词，但是中文是以字为单位的，词与词之间无空格进行分隔，因此对中文进行情感分析，首先就需要进行分词处理。分词(Word Segmentation，简称 WS)，有时也称为切词，是指将连续的汉字序列根据某一规范切分成新的词序列。由于在汉语中，词是承载语义的最基本单元，所以分词是文本情感分析中最基础的任务。目前，中文分词的方法大体可分成两类：基于字符串匹配的分词方法以及基于统计的分词方法。

1. 基于字符串匹配的分词方法

该方法的基本思想是不使用规则知识和统计信息，只根据词典信息，利用某种规则将中文字符串和词典中收录的词语进行逐一匹配。如果中文字符串与词典中内容匹配成功，则将这个字符串切割开，作为一个词语；再从切割位置开始，进行重新匹配，直到将待分词的中文字符串切分为一个个词语序列为止。这种分词方法主要由三个要素组成，分别为字符串匹配的起始位置、分词的词典以及匹配的原则。根据字符串匹配的起始位置不同，分词方法可以分为正向匹配的分词方法和逆向匹配的分词方法；根据匹配原则的不同，分词方法可以分为最大匹配的分词方法和最

小匹配的分词方法。基于字符串匹配的分词方法操作简单且实现起来比较轻松，但是由于字典的限制以及歧义切割的问题，该算法效率还不是很高。

2. 基于统计的分词方法

由于词是由字组合成的，且较为稳定。所以，在文本中，若相邻几个字共同出现的次数越多，则越有可能是一个词。基于统计的分词方法就是根据这个思路，统计语料中相邻且共现的字组合的频度，当频度大于某个阈值时，则这几个字可以组合成一个词。这种方法不需要分词的词典。但是，由于有些相邻的字出现的频度高，但并不是词，如“我的”。因此，该方法也具有一定的误差。在实际应用中，经常将基于字符串的分词方法以及基于统计的分词方法结合起来使用，使分词速度快且效率高。

2.2.2 词性标注

词性是指根据词的特点对词进行分类。如对一个概念进行描述的词语叫做名词，在文中引用这个词的词叫做代词。词性标注(Part-of-Speech tagging 或 POS tagging)，也称为词类标注，是指为分词结果中的每个词语标注一个正确词性，即确定每个词是名词、形容词、动词还是其它词性。在文本情感分析中，很多词语的抽取以及极性的判断都是基于词性的。如绝大多数产品特征的词性都是名词，很多情感词的词性是形容词或者副词。所以，词性标注是文本情感分析中必不可少的一部分。常用的词性如表 2.1 所示。

表 2.1 常见的词性

词性	符号表示	词性	符号表示
形容词	a	连词	c
副词	d	数词	m
名词	n	人名	nr
地名	ns	方向名词	nd
机构名称	ni	位置名词	nl
时间名词	nt	其它名词	nz
代词	r	动词	v
量词	q	介词	p
感叹词	e	标点	wp

2.2.3 句法分析

依存句法(Dependency Parsing，简称 DP)，通过分析语句单位内成分之间的依存关系表明语句的句法结构。即通过识别语句中的“主谓宾”、“定状补”等这些语法成分来分析各词语之间的关系。依存句法分析的标注关系如表 2.2 所示。

表 2.2 依存句法的标注关系

关系类型	符号表示	关系类型	符号表示
主谓关系	SBV	动宾关系	VOB
间宾关系	IOB	前置宾语	FOB
定中关系	ATT	状中结构	ADV
并列关系	COO	介宾关系	POB
动补结构	CMP	兼语	DBL
左附加关系	LAD	右附加关系	RAD
核心关系	HED	标点	WP

2.3 语言技术平台

语言技术平台(Language Technology Platform，简称 LTP)是由哈工大社会计算与信息检索研究中心研发的一套中文自然语言处理系统，它可以为用户提供较为准确的中文自然语言处理云服务。LTP 提供了多个较为准确的中文自然语言处理模块，包括词法分析、句法分析以及语义分析等。

如使用 LTP 对产品评论“太失望了，声音太小了。诺基亚手机都能造出来，一个小小的声音器件不能选好的？是没有能力还是不注重整体品质？”进行分词及词性标注，结果如下：

太/d 失望/a 了/u，/wp 声音/n 太/d 小/a 了/u。/wp 诺基亚/nz 手机/n 都/d 能/v 造/v 出来/v，/wp 一个/m 小小的/z 声音/n 器件/n 不/d 能/v 选/v 好/a 的/u？/wp 是/v 没有/v 能力/n 还是/d 不/d 注重/v 整体/n 品质/n？/wp

对于句子中“太失望了，声音太小了。”的句法分析结果如图 2.1 所示。

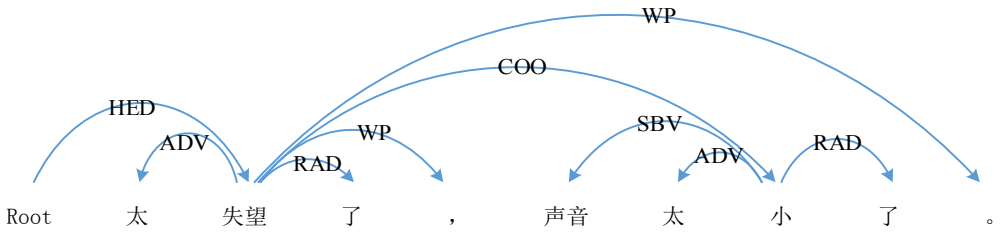

图 2.1 例句的句法分析结果

2.4 Word2Vec

2.4.1 词向量

由于计算机无法直接理解人类的语言，因此，就需要将语言转化成数学，转化成计算机能够识别的格式来处理情感分析的问题。词向量就是一种有效的方式。

One-hot Representation 是词向量的一种表示方式，在该种表示方式中，词语和词语之间是相互独立的，向量维度和词表大小相同。词语表示为对应的维度 1，其余全是 0，即 1 的位置表示该词语在词表中所在的位置。这种方式表示简单，但是有可能会导致维数灾难[44]。

Hinton 在 1986 年提出了使用 Distributed representation[45]来表示词向量。该表示方式的基本思想是：通过对语料进行训练，将语料中每个词语都映射为具有相同长度的多维实数向量，进而通过计算实数向量间的距离来衡量两个词语的相似度，距离越近，则相似度越高。这种低维的词向量表示方式优于 One-hot Representation，可以有效地避免维数灾难。谷歌开源的学习工具 Word2Vec 就是使用了 Distributed representation 的词向量表示方式，具有较出众的效果。

2.4.2 Word2Vec 介绍

Word2Vec 使用的是 Distributed representation 的词向量表示方式。其基本思想是利用深度学习，通过训练将每个词映射成多维的实数向量，通过词之间的距离(比如余弦相似度等)来判断它们之间的语义相似度。Word2Vec 工具中有两种模型，分别

为 CBOW(Continuous Bag-Of-Words，连续词袋模型)模型和 skip-gram 模型。其中，CBOW 模型如图 2.2 所示。

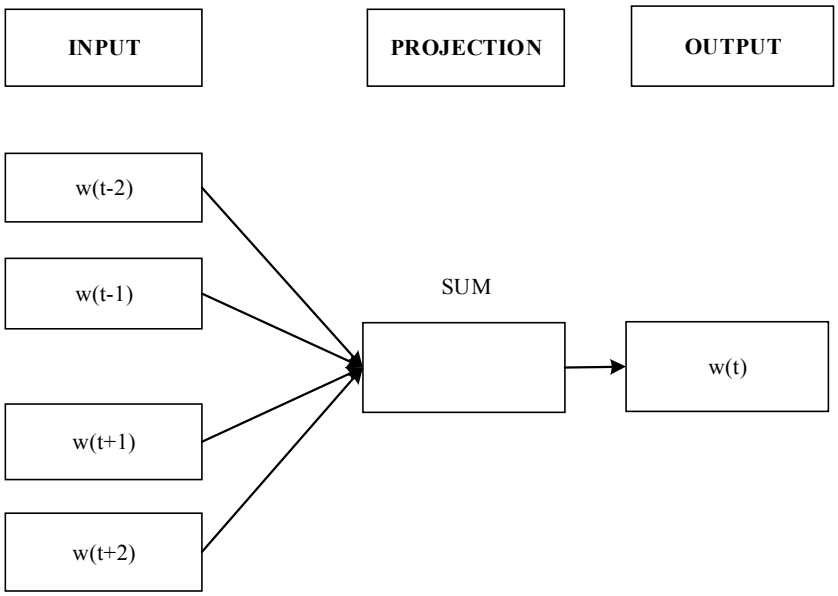

图 2.2 CBOW 模型

Skip-gram 模型如图 2.3 所示。

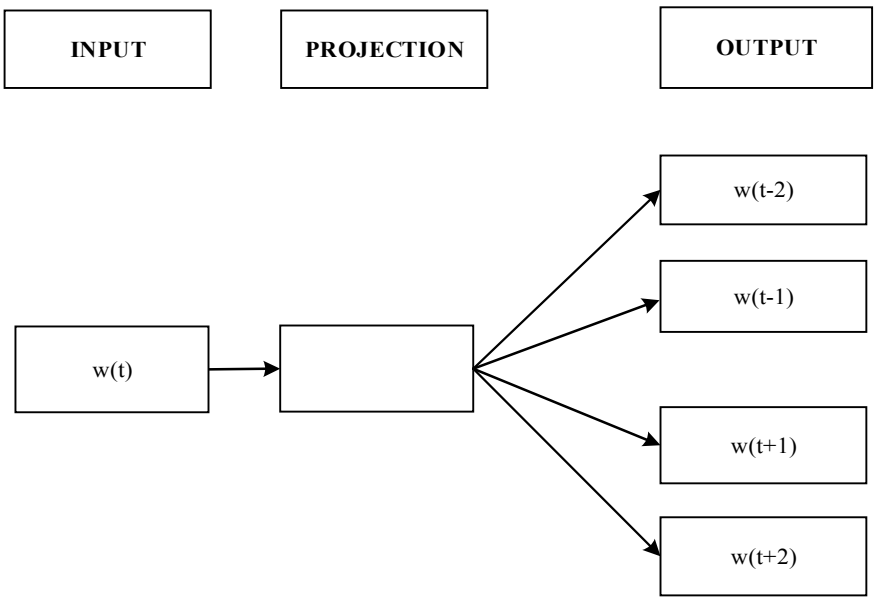

图 2.3 skip-gram 模型

由图 2.2 和 2.3 可知，CBOW 模型和 skip-gram 模型都包含三层，即输入层、投影层和输出层。CBOW 模型通过上下文来预测当前词。已知上下文 $w(t-2), w(t-1), w(t+1), w(t+2)$，预测未知词 $w(t)$。Skip-gram 模型则是通过当前词来预测上下文。已知词 $w(t)$，预测其上下文 $w(t-2), w(t-1), w(t+1), w(t+2)$。

2.5 WordNet

WordNet 是由 Princeton 大学的心理学家、语言学家以及计算机工程师研制的一种基于认知语言学的英语词典[46]。它和传统字典不同，不是按照字母组织词条，而是基于同义词集对词语进行组织。

WordNet 由名词、动词、形容词和副词组成。包含了 155287 个词语，117659 个同义词集，206941 个词义对。WordNet 词语、同义词集及词义对的数量如表 2.3 所示。

表 2.3 词语、同义词集及词义对的数量

词性	词语数	同义词集数	词义对数
名词	117798	82115	146312
动词	11529	13767	25047
形容词	21479	18156	30002
副词	4481	3621	5580
总数	155287	117659	206941

WordNet 中定义了多种关系，主要包括以下几种语义关系：

1. 同义关系(Synonymy)：同义关系是 WordNet 中最基本的语义关系，同义词集就是通过同义关系组成的。同义关系的词语在词性上是相同的。

2. 反义关系(Antonymy)：反义关系也是词性相同的语义关系。

3. 部分整体关系(Meronymy)：部分整体关系包括三种关系。如 A 是整体，B 是部分。第一种关系是 B 是 A 的组成部分，第二种关系是 B 是构成 A 的材料，第三种关系是 B 是 A 的成员。

4. 从属关系(Hyponymy)：从属关系是父类和子类的关系。通常只有一个父类，按照这种关系可以将名词的含义组织成一个层次结构。

WordNet 中的部分语义关系如表 2.4 所示。

表 2.4 部分语义关系

语义关系	语法类别	例子
同义关系	名词、动词、形容词、副词	“sad”, “unhappy”
反义关系	形容词、副词	“wet”,“dry”
部分整体关系	名词	“brim”,“ hat”
从属关系	名词	“tree”,“plant”

2.6 实验性能评价标准

本文使用信息检索领域标准的评价准则：准确率 P 、召回率 R 以及调和评价值 $F1$ 值来衡量产品特征抽取算法的性能。具体定义如公式 2.1、2.2、2.3 所示：

$$P = \frac{num}{all\_num} \tag{2.1}$$

$$R = \frac{num}{man\_num} \tag{2.2}$$

$$F1 = 2 * \frac{P * R}{P + R} \tag{2.3}$$

其中， num 表示抽取出正确的产品特征个数， all_num 表示抽取出的所有产品特征个数， man_num 表示人工标注的产品特征个数。

本文使用人工评判的方法对构建的情感词典进行评估。具体做法是随机选取情感词典中一定比例(50 或 100 个)的词，人工判断该情感词极性是否正确，以这些情感词的正确率来衡量构建的情感词典性能[36]。情感词典正确率的计算如公式 2.4 所示：

$$A = \frac{c}{a} \tag{2.4}$$

其中， A 表示情感词典的正确率， a 表示人工评判情感词典性能时所选取的情感词个数， c 表示选取的情感词中情感极性正确的个数。

2.7 本章小结

本章主要介绍了本文中所涉及的相关知识，包括分词、词性标注、句法分析，哈工大的语言技术平台、Word2Vec、WordNet 以及实验评价标准等。其中，分词和

词性标注是进行情感分析最基础的任务。通过对语句进行句法分析，可以获得词语之间的关系。Word2Vec 可以将词语表示为空间向量，进而通过词语之间的距离计算出词语之间的相似性。利用这一点，在情感词典的扩充中可以有效找出语料中具有相似性的词语。WordNet 是一个基于同义词集的语义知识库，可以通过 WordNet 对种子词进行扩充，得到构建的情感词典。最后简单介绍了本文实验的评价标准。

第 3 章 基于领域相关性的产品特征抽取技术研究

3.1 引言

在大量的产品评论中，对用户和商家来说，不仅需要了解用户对产品的整体看法，还需要了解产品的细节评价，如：

评论 1：“华为的手机整体不错，电池很耐用，价格也便宜。”

用户除了要了解华为手机的整体评价，更想了解自己看重的某些产品特征如“电池”、“价格”等是否满意。这些产品特征及对应评价也正是用户决定是否购买该产品的主要因素。为了向用户和商家提供他们想要了解的有关产品的细粒度信息，就需要对产品特征抽取技术进行研究。

在已有产品特征抽取的方法中，有监督学习的方法可以在一个给定领域语料中发挥较好的效果，但是对于不同领域的语料必须重新训练模型，存在语料标注工作量大、领域可移植性差等问题。无监督学习的方法通过定义一些模板规则来识别产品特征，但是由于产品评论比较口语化，缺乏一定的结构，因此，模板规则并不完全适用。并且研究者们抽取产品特征时，通常只基于一个领域识别产品特征，而忽略了产品特征分布的差异性。例如，手机评论中经常出现产品特征“摄像头”，但是“摄像头”在领域无关性语料(美妆评论)中却很少出现。本文充分利用产品特征在不同领域语料中分布信息的差异性，提出了基于领域相关性的产品特征抽取方法，该方法不仅领域独立性好，而且也便于在不同领域的语料下进行产品特征的抽取。

本章提出了基于领域相关性的产品特征抽取方法(A Statistical Approach to Opinion Target Extraction Using Domain Relevance，简称 SDR)。SDR 方法的总体流程如图 3.1 所示。方法总体思路如下：首先，利用点对互信息合并该值大于某一阈值的相邻名词，然后根据制定的句法规则抽取候选特征。对于每个候选特征，利用它在两个不同领域语料中分布的差异性来计算领域相关性值。最后，将领域相关性值大于某一阈值的候选特征作为最终产品特征。

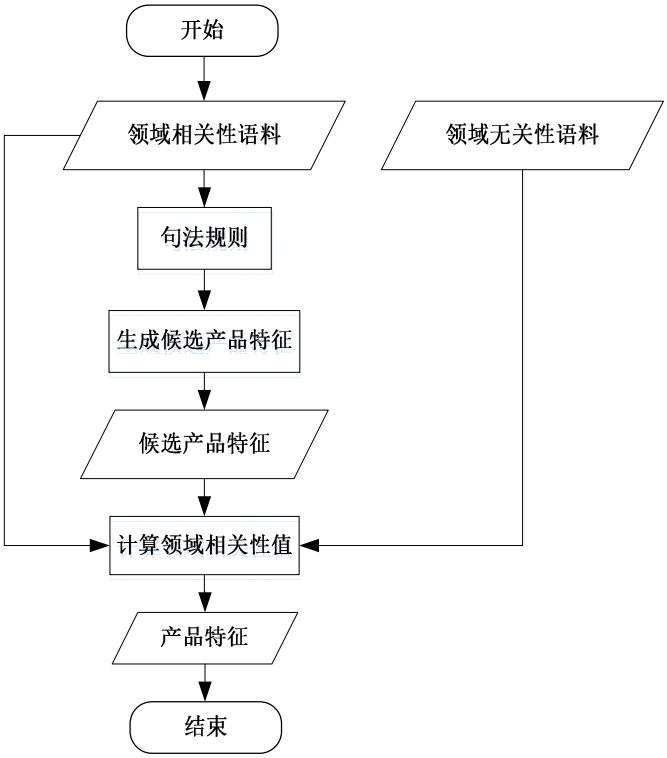

图 3.1 SDR 方法的总体流程图

3.2 点对互信息

点对互信息(PMI)是机器学习领域经常使用的一种特征相关性判别准则，它是对一个随机变量与另一个随机变量有关信息量的度量，即对两个随机变量共同具有的信息量的度量。最早是由 K.W.Church 和 PHanks 在 1989 年提出[47]，本文利用点对互信息通过统计词语共同出现的概率来计算两个词语间的关系，其定义如公式 3.1 所示：

$$PMI(w_1,w_2)=\log_2\left(\frac{P(w_1,w_2)}{P(w_1)P(w_2)}\right) \tag{3.1}$$

其中，$P(w_1,w_2)$表示词语w_1和w_2共同出现的概率，$P(w_1)$表示词语w_1出现的概率，$P(w_2)$表示词语w_2出现的概率，$\frac{P(w_1,w_2)}{P(w_1)P(w_2)}$是对$w_1$和$w_2$统计独立性的度量，值越大，则两者的统计相关性也越大。因此，两个词语的点对互信息值越大，表示这两个词语的相关性也越大。

3.3 似然比技术

由于似然比检验(Likelihood ratio test，简称 LRT)[48,49]在产品特征抽取中有较好的应用。因此，本文结合似然比技术进行产品特征的抽取。似然比计算方法如公式 3.2 所示：

$$lr=\begin{cases}-2*l, r_2<r_1\\ 0, r_2\geq r_1\end{cases} \tag{3.2}$$

其中，r_1、r_2和l的计算方法分别如公式 3.3、3.4 和 3.6 所示：

$$r_1=\frac{n_1}{n_1+n_2} \tag{3.3}$$

$$r_2=\frac{n_3}{n_3+n_4} \tag{3.4}$$

$$r=\frac{n_1+n_3}{n_1+n_2+n_3+n_4} \tag{3.5}$$

$$\begin{aligned}l=&\left(n_1+n_3\right)\log r+\left(n_2+n_4\right)\log\left(1-r\right)\\&-n_1\log r_1-n_2\log(1-r_1)-n_3\log r_2-n_4\log(1-r_2)\end{aligned} \tag{3.6}$$

n_1、n_2、n_3、n_4的计算如表 3.1 所示。

表 3.1 候选特征的统计量

—	D	I
BNP	n_1	n_2
$\overline{BNP}$	n_3	n_4

其中，D是领域相关性语料，I是领域无关性语料。BNP是出现在D中的一个候选特征，$\overline{BNP}$是出现在D中除BNP以外的其它候选特征。n_1和n_2分别是候选特征BNP在领域相关性语料D和领域无关性语料I中出现的频率，n_3和n_4分别是$\overline{BNP}$在领域相关性语料D和领域无关性语料I中出现的频率。似然比lr的值越大，表示该候选特征BNP和领域相关性语料的关联程度越大。

3.4 候选特征识别

本节通过名词短语的识别以及制定的句法规则来识别候选产品特征。

3.4.1 名词短语的识别

本文使用哈尔滨工业大学开发的语言技术平台(LTP)对中文语料进行分词、词性标注以及依存句法分析。

由于在实际评论中，某些产品特征是由多个相连的名词组成的。例如，在产品评论语句“手机壳粗糙”中，产品特征为“手机壳”，但是利用分词软件进行分词时，“手机壳”会被分为“手/n 机壳/n”，从而使得本应完整的产品特征被分词为两个不相关的部分。

针对这类问题，本章利用点对互信息确定是否合并相邻的名词，当相邻名词的点对互信息值达到某一阈值时，将其合并作为新的名词。

3.4.2 候选特征抽取

在本文的方法中，首先根据一定的句法规则从语料中抽取候选产品特征。抽取候选产品特征主要有以下两步：首先对产品评论进行依存分析，得到词语之间的关系，然后按照预先定义的句法规则来抽取候选产品特征。

在大部分产品特征抽取的研究中，考虑将名词和名词短语作为候选产品特征。基于句法规则，本文按照如下几种规则进行候选产品特征的抽取：N+SBV，N+VOB 和 N+POB。其中，N 代表名词或名词短语，SBV、VOB 和 POB 分别代表主谓结构、动宾结构和介宾结构。“N+SBV”表示如果一个词语是名词或名词短语，并且具有主谓结构，则将该词语作为候选特征；“N+VOB”表示如果一个词语是名词或名词短语，并且具有动宾结构，则将该词语作为候选特征；“N+POB”表示如果一个词语是名词或名词短语，并且具有介宾结构，则将该词语作为候选特征。如：“手机的做工很差”，其中“做工”是名词，且和“差”具有主谓关系；在评论“我很喜欢这个款式”中，“款式”是名词，且与“喜欢”具有动宾关系；在评论“我对手机电池很失望”中，“电池”是名词，且与“对”具有介宾关系，则将“做工”、“款式”和“电池”作为候选产品特征。

3.5 产品特征筛选

基于领域的相关性算法利用两种领域的差异性来计算一个词和一个特定领域的关联程度。它避免了在不同评论语料中需要找不同规则来进行产品特征筛选，同时针对不同语料也不需要重新标注，从而使该方法可以方便地应用于不同领域中，进而获得良好的可移植性。

基于领域相关性的产品特征抽取方法本文参考了文献[50]的方法，并在其上进行了改进。与文献[50]中方法不同的是，本文没有针对候选产品特征分别计算它在两个领域的内、外领域相关性值，而是把候选产品特征在两个不同领域语料中出现频率的差异性与该特征在领域相关性语料中的分布信息结合起来，从而计算候选产品特征的领域相关性值。该方法不仅减小了计算量，而且实验结果也表明，该方法提高了算法的准确率。本文提出的方法如下所述：

1. 首先计算每个候选产品特征 C_i 在领域相关性语料中的离差 $disp_i$。

(1) 利用 TF-IDF 方法来计算候选特征在语料中的权重值。对于每个候选产品特征 C_i，在每条评论 D_j 中都有一个词频 TF_{ij}，在整个语料中有一个频率 DF_i。则候选产品特征 C_i 在评论 D_j 中的权重 ω_{ij} 的计算如公式 3.7 所示：

$$\omega_{ij}=\begin{cases}\left(1+\log TF_{ij}\right)\times\log\dfrac{N}{DF_i}, TF_{ij}>0\\ 0, TF_{ij}\le 0\end{cases} \tag{3.7}$$

其中，i 表示特征编号，在语料中共 M 个，即 $1\le i\le M$；j 表示评论编号，在语料中共 N 条，即 $1\le j\le N$。

(2) 每个候选产品特征标准差的计算如公式 3.8 所示：

$$s_i=\sqrt{\frac{\sum_{j=1}^{N}\left(\omega_{ij}-\overline{\omega_i}\right)^2}{N}} \tag{3.8}$$

其中，$\overline{\omega_i}$ 表示词 T_i 在整个语料中的平均权重，如公式 3.9 所示：

$$\overline{\omega_i}=\frac{1}{N}\sum_{j=1}^{N}\omega_{ij} \tag{3.9}$$

(3) 则每个候选产品特征在领域相关性语料中离差的计算如公式 3.10 所示：

$$disp_i = \frac{\overline{\omega_i}}{s_i} \tag{3.10}$$

2. 然后计算每个候选产品特征C_i的偏差dev_i。

(1) 候选特征C_i在两个语料中分布的差异性$diff_i$计算如公式 3.11 所示：

$$diff_i = \left[P_D(C_i) - P_I(C_i)\right] e^{|P_D(C_i) - P_I(C_i)|} \tag{3.11}$$

其中，$P_D(C_i)$和$P_I(C_i)$分别表示候选特征C_i出现在领域相关性和领域无关性语料中的概率。如果一个候选特征C_i有高的$diff_i$值，表示C_i出现在领域相关性语料中的频率比出现在领域无关性语料中的频率更高，如果候选特征C_i的$diff_i$值比较低，则相对来说，C_i在领域相关性语料中出现的频率较低。

(2) 每个候选产品特征C_i的偏差dev_i计算如公式 3.12 所示：

$$dev_i = diff_i \times lr_i \tag{3.12}$$

其中，lr_i表示C_i在领域相关性语料中的似然比。当$diff_i$和lr的值越大，则候选产品特征和领域相关性语料在词频和文档频率上的相似度都越高。

3. 最后计算每个候选产品特征C_i的领域相关性值rel_i。

候选产品特征C_i的领域相关性值rel_i的计算如公式 3.13 所示：

$$rel_i = disp_i \times dev_i \tag{3.13}$$

其中，$disp_i$通过计算候选产品特征在领域相关性语料中的文档分布来衡量该词在整个文档中的重要性，dev_i通过计算候选特征在不同语料中的分布来衡量该词和领域相关性语料的相关性。在公式 3.13 中，领域相关性rel_i结合了$disp_i$和dev_i，从而反映了该词在整个语料中的分布情况。

最后，本文将领域相关性值大于某一阈值的候选产品特征作为产品特征。

本文产品特征抽取方法的算法描述如图 3.2 所示：

输入：领域相关性语料(Domain review corpus，简称 DC)，领域无关性语料(Domain-independent corpus，简称 DIC)；互信息的阈值 pmi；领域相关性阈值 rel

输出：产品特征集合 F

1. 初始化产品特征集合 F，令 $F=\varnothing$

2. 分别对语料 DC、DIC 进行分词和词性标注

3. 根据公式 3.1 对领域相关性语料 DC 中的相邻名词 N_1、N_2，计算其互信息 $PMI(N_1,N_2)$。如果 $PMI(N_1,N_2)>pmi$，则将 N_1 和 N_2 合并为一个名词

4. 对语料 DC 进行依存分析，并利用句法规则得到候选产品特征集合 C

5. 对每个候选特征 $C_i \in C$：

   (1) 根据公式 3.10 计算 C_i 在领域相关性语料中的离差 $disp_i$

   (2) 根据公式 3.12 计算 C_i 的偏差 dev_i

   (3) 根据公式 3.13 计算 C_i 的领域相关性值 rel_i

   (4) 如果 $rel_i > rel$，则 $F = F \cup C_i$

6. 结束循环，返回 F

图 3.2 SDR 方法的算法描述

3.6 实验与分析

本节介绍了本章产品特征抽取方法所用的实验数据以及对比的基线实验方法，并对实验结果进行了分析。通过与基线实验的对比，验证了本文产品特征抽取算法的有效性和准确性。

3.6.1 实验数据及评价标准

1. 语料：本文进行产品特征抽取方法所用语料选自电子商务网站亚马逊(www.z.cn)中的 1000 条手机评论，由人工标注产品特征。通过使用 Kappa[51]值来计算两名标注者标注结果的一致性， 结果 Kappa 值大于 0.8，表明标注的结果一致性

很高，语料可用，共标注的产品特征有 1529 个。领域无关的语料选自亚马逊美妆评论中的 3943 条评论。

2. 评价标准：本文使用 2.6 节所述的信息检索领域标准的评价准则：准确率 P 、召回率 R 和调和评价值 $F1$ 值，对语料中抽取的产品特征进行评价。

3.6.2 基线方法

为了说明本文提出的产品特征抽取方法的有效性，本章选取了以下几种方法作为基线方法：

1. 基于名词的产品特征抽取方法(A method of opinion targets extraction based on nouns，简称 BN)：选用名词直接作为产品特征。

2. 基于似然比的产品特征抽取方法(A method of opinion targets extraction based on LRT，简称 LRT)：将符合 $N+N$，$J+N$，$N+N+N$，$J+N+N$，$J+J+N$ 规则之一的作为候选特征，再利用文献[18]中的似然比技术筛选得到最终产品特征，其中 J 代表形容词的词性，N 代表名词的词性。

3. 基于内外领域的产品特征抽取方法(A method of opinion targets extraction based on intrinsic and extrinsic domain relevance，简称 IEDR)：文献[50]基于内外领域的产品特征抽取方法。

4. 基于互信息及内外领域的产品特征抽取方法(A method of opinion targets extraction based on PMI and intrinsic and extrinsic domain relevance，简称 PIEDR)：利用互信息对相邻名词进行合并，然后再利用 IEDR 方法进行产品特征的抽取。

3.6.3 实验分析

表 3.2 实验结果对比

实验方法	P (%)	R (%)	F1 (%)
BN	21.04	86.85	33.87
LRT	63.89	46.83	54.05
IEDR	72.73	65.14	68.73
PIEDR	77.78	63.18	69.72
SDR	68.66	72.94	70.74

本文方法和基线方法的实验结果如表 3.2 所示，本文方法的点对互信息和领域相关性阈值分别为 3 和 0.0015。

从表 3.2 可以看出，BN 的实验结果最差，因为 BN 方法只是简单的将名词作为产品特征，由此会引入较大的误差，导致准确率降低。如评论“我给妈妈买了个华为荣耀 7，手感不错”，BN 方法会将所有的名词“妈妈”、“华为”和“手感”作为产品特征抽取出来，而事实上只有“手感”是产品特征，因此会引入较大的噪音。和 BN 方法相比，LRT 方法基于一些已有规则进行候选产品特征的抽取，进而再利用似然比技术，通过计算候选产品特征和领域相关性语料的关联程度筛选出产品特征。该方法仅依靠似然比技术对候选特征进行筛选，得到的词语大部分是产品特征，但是，会过滤掉一些产品特征，使最终得到的产品特征数量减少。因此，LRT 准确率(63.89%)较高，但是召回率(46.83%)较低。由此说明，LRT 能够运用在候选特征的筛选中，但是，由于它的 $F1$ 值较低，所以还有较大的空间来提高该算法的性能。和 LRT 方法相比，IEDR 方法充分利用产品特征在不同领域语料中分布信息的不同来抽取产品特征，取得了较好的结果，$F1$ 值由 54.05%提高到了 68.73%。

正如 3.4.1 节中所述，在一些情况下，产品特征是由几个相邻的名词组成的，而不仅仅是单个名词。但是由于分词的不准确性，以及依存句法分析仅能提供词语间的依赖关系，因此会有不完整的候选特征被抽取出来。为了解决这个问题，本章利用 PMI 来确定是否要将相邻名词合并作为一个新的名词。为了分析引入 PMI 对实验结果的影响，本章增加了基线实验 PIEDR。相比于 IEDR，PIEDR 的 $F1$ 值从 68.73%提高到了 69.72%，证明了使用 PMI 能有效解决分词软件不准确而导致错误的问题。

在所有实验中，本文方法 SDR 的实验结果均优于基线方法的实验结果。相比于 PIEDR 方法，本文将似然比技术结合起来，不仅能利用领域的差异性，同时也充分利用了似然比技术的优势，使 $F1$ 值从 69.72%提高到了 70.74%。说明了本文方法在计算领域相关性值上的有效性，加上本文方法的代价小，所以在产品特征抽取中，该方法是一个较好的选择。

为了进一步验证两个参数(互信息阈值 pmi 和领域相关性阈值 rel)对本文方法的影响，本章每次固定一个变量进行实验。实验结果如图 3.3 和 3.4 所示。

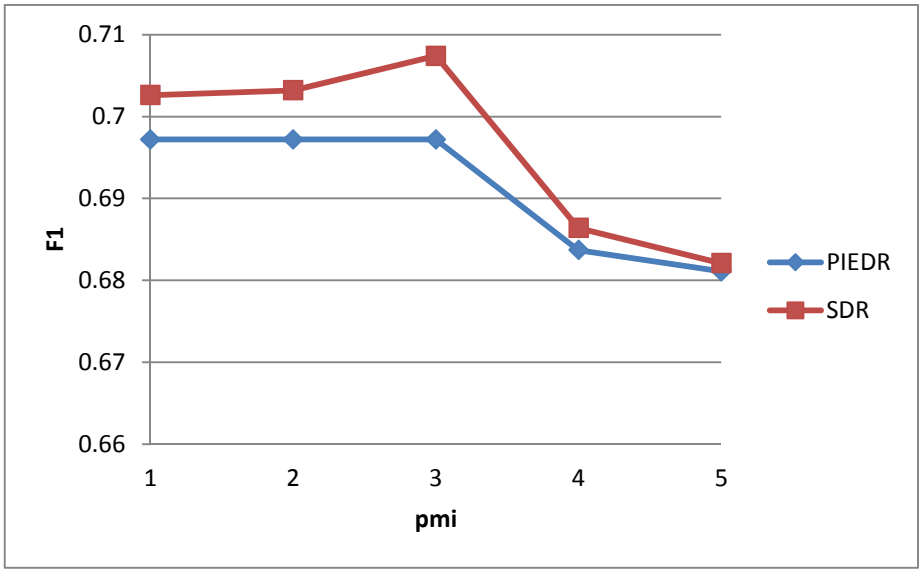

图 3.3 实验结果随互信息阈值 pmi 变化的曲线图

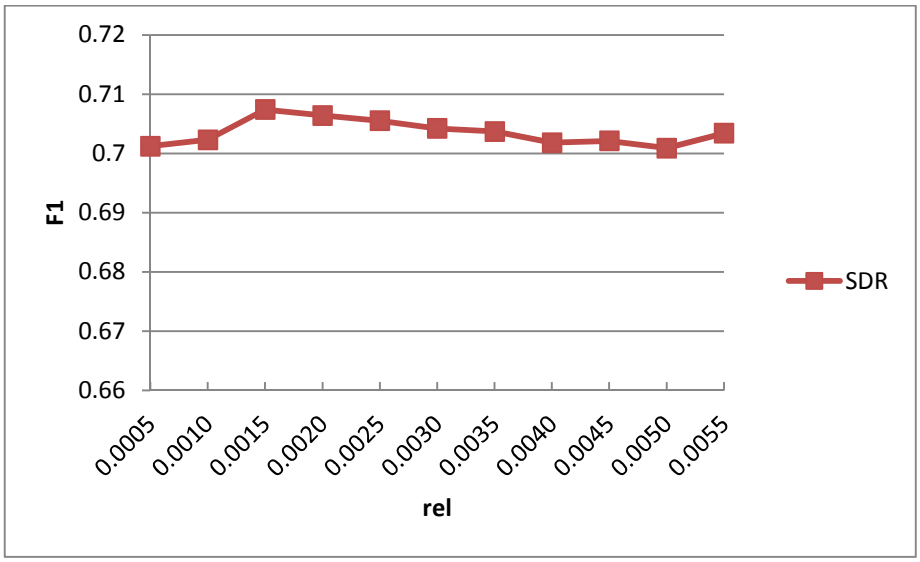

图 3.4 实验结果随领域相关性阈值 rel 变化的曲线图

图 3.3 显示了实验结果随互信息阈值 *pmi* 变化的实验结果。从图中可以看出，随 *pmi* 的变化，SDR 方法的实验结果始终优于 PIEDR，当 *pmi* 为 3 时，实验结果最优。说明随互信息 *pmi* 的变化，SDR 方法的效果较好，具有较好的稳定性。图 3.4 显示了 SDR 方法随领域相关性阈值 *rel* 变化的实验结果，从图中可以看出，随 *rel* 的变化，实验结果仅仅有较小的变化，且当 *rel* 为 0.0015 时，结果最优。由此说明，SDR 方法随领域相关性阈值 *rel* 的变化，实验结果比较稳定。进一步表明 SDR 方法在实验阈值变化的情况下，具有良好的鲁棒性。

3.7 本章小结

基于产品评论的细粒度情感分析研究中，其中最核心任务之一就是产品特征的抽取。本章结合似然比技术和产品特征在不同领域的差异性，提出了基于领域相关性的产品特征抽取方法。首先使用点对互信息确定是否要合并相邻名词，减小了分词不准确引入的误差；然后在内外领域相关性的算法上进行了改进，结合似然比技术以及产品特征在两个领域分布的差异性筛选出产品特征。最后给出了实验验证。实验结果表明，本文提出的方法达到了较好的效果。

第 4 章 基于标签传播的情感词典构建方法

4.1 引言

通过对产品评论文本进行细粒度情感极性地分析，如积极情感和消极情感等，不仅能够为消费者购买某产品提供更全面的参考信息，同时也可以使商家更深入地了解各自产品的优缺点以及用户的需求，为进一步改进产品提供决策支持[52]。

在细粒度情感分析中，情感词典的构建是一个基础任务。情感词典构建的主要任务是通过计算词语的情感倾向性，不断扩充情感词典，最终得到的情感词典中都是已经确定极性的情感词。情感词典的构建不仅可以为文本级的分类提供一定的基础，对词语级的语义理解也起着重要的作用。例如，很多无监督文本情感分类方法首先需要选取一定的种子词，而种子词一般选择情感极性已知的词语。

情感词典的完整性以及正确性会严重影响情感分析的结果。情感词典覆盖率低、情感词分类不正确等这些问题，对情感词典的性能有很大的影响，极大地制约了情感倾向性分析工作的开展。针对以上问题，本文提出了基于标签传播的情感词典构建方法。该方法主要使用 Google 开源工具 Word2Vec 训练语料得到词向量模型，并利用依存句法分析词语的关系，以获得词语之间的相似性，然后利用词语之间的相似性自动对种子词进行扩充，同时利用标签传播算法确定最终构建得到的情感词典。

本文所提出的方法简称为 W2V&CR-LP，其整体思路是：首先分别选取一定数量的积极和消极的情感种子词，然后使用 Word2Vec 训练语料，如果词语 a 和种子词 b 具有相似性，则词语 a 和 b 之间有一条边；同时通过对语料进行依存句法分析，找出和种子词具有连词关系的情感词，如果词语 c 和种子词 b 具有连词关系，则词语 c 和 b 之间有一条边。最后得到一个图，再通过标签传播算法进行极性传播，最终确定所有扩充得到的候选情感词极性。总体流程如图 4.1 所示。

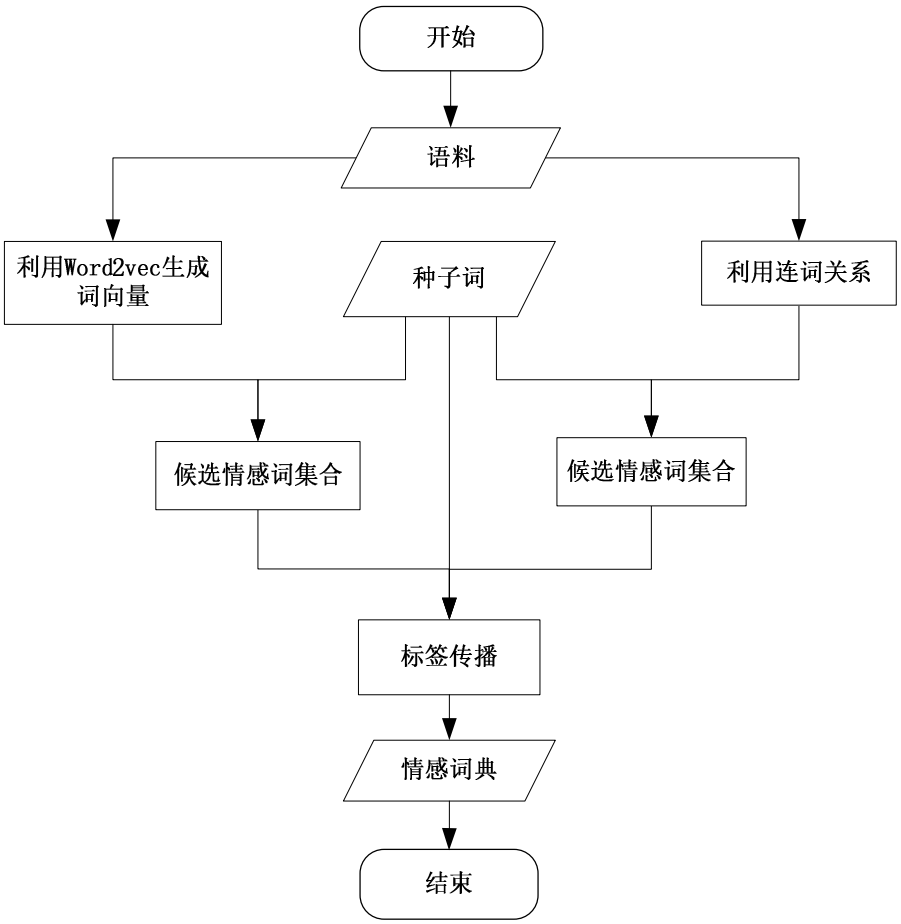

图 4.1 W2V&CR-LP 方法的总体流程图

4.2 种子词的选取

在情感词典扩充中，种子词通常是具有明显情感倾向的词语，而种子词的选取直接决定了情感词典扩充结果的好坏。本文选择 SentiWordNet[53]中部分主观性强的词语作为种子词。“主观性强”的判断是根据文献[53]对 SentiWordNet 中各种词性的情感词打分，分数由高到低进行排列，本章选择分数高且常见的情感词作为种子词。如果同一词语以多种词性出现，保留其中分数较高的词性。本章将极性为积极的情感词放入积极的种子词集合(“Positive”集合，简称“Pos”集合)中，将极性为消极的情感词放在消极的种子词集合(“Negative”集合，简称“Neg”集合)中。如把“good”、“nice”等词语放在“Pos”集合中，将“bad”、“unfortunate”等词语放在“Neg”集合中。所选取的部分种子词如表 4.1 所示。

表 4.1 部分种子词集合

“Pos”集合	“Neg”集合
good	bad
happiness	sad
love	unfortunate
better	poor
nice	worst
…	…

4.3 利用 Word2Vec 抽取候选情感词

本文使用 Google 开源工具 Word2Vec 来训练词向量。Word2Vec 工具中包含了 CBOW 模型和 skip-gram 模型。其中，CBOW 模型是通过上下文来预测当前词，而 skip-gram 模型是通过当前词来预测上下文。由于 skip-gram 模型可以极其高效的对大量非结构化文本训练词向量[54]，因此，本文中选择 skip-gram 作为训练模型，将文本集作为输入，每个词对应的生成向量作为输出。通过生成的词向量，可以计算与用户指定词语之间的相似度。

假设存在词组序列 w_1、$w_2 \ldots w_t$，skip-gram 模型的目标是使目标函数 F 最大化，如公式 4.1 所示：

$$F = \frac{1}{T}\sum_{t=1}^{T}\sum_{-d\le j\le d, j\neq 0}\log P(w_{t+j} \mid w_t) \tag{4.1}$$

其中，d 是表示词组序列中上下文窗口大小的常数，d 越大得出的结果越精确，但是相应训练时间也会增加。由于 Negative Sampling 可以加快训练速度并且对词语的表示更准确[55]，因此，本文使用 Negative Sampling 方法去训练 skip-gram 模型。

本章首先通过 Word2Vec 在语料上训练词向量模型，得到每个词语的词向量，然后分别对积极、消极种子词进行扩充，将与种子词相似度大于设定阈值的词语和种子词放在同一集合，并记录其相似度值。

4.4 利用连词关系抽取候选情感词

一般而言，文本中出现的转折关系的连词会使文本前后情感词极性发生变化，而一般并列关系的连词则不会改变前后情感词的极性。

本文中的连词关系主要分成两种：并列关系和转折关系。本文定义的并列关系的连词有：and，neither...nor，either...or，as well as，not only ...but also...等，转折关系的连词有：but，yet，however，still，while，on the contrary 等。并列关系的文本如例 1)、例 2)，转折关系的文本如例 3)、例 4)。

例 1)“The cellphone is beautiful and durable”

例 2)“The phone is neither cheap nor beautiful”

例 3)“The phone's performance is good, but the price is too expensive”

例 4)“The introduction of this phone was very good, yet in fact it was very disappointing”

考虑到具有转折关系的情感词在评论文本中出现的词语距离较远，运用传统的词汇窗口技术来找出连接的情感词的方法可能失效。因而，本文利用 Stanford Parser[56]对语料进行依存句法分析，来获得词语之间的连词依赖关系，再将和种子词具有并列或者转折关系的词语抽取出来作为候选情感词，种子词的选取如 4.2 节所述。如利用 Stanford Parser 对具有转折关系的例 3）进行依存句法分析，可以得到“det(phone-2,The-1),nmod_poss(performance-4,phone-2),case(phone-2,'s-3),nsubj(good-6,performance-4),cop(good-6,is-5),root(ROOT-0,good-6),cc(good-6,but-8),det(price-10,the-9),nsubj(expensive-13,price-10),cop(expensive-13,is-11),advmod(expensive-13,too-12),conj_but(good-6,expensive-13) ”，由此可以知道“good”和“expensive”具有转折关系。因此，当“good”是种子词时，则可以抽取出候选情感词“expensive”。

当评论文本中的词语和种子词具有并列关系时，将该词语作为该种子词扩充的情感词加入到对应情感词集合中，当集合中已有该词时，则将其数量加 1，且标记为并列关系；当评论文本中的词语和种子词具有转折关系时，将该词语作为该种子词扩充的情感词加入到对应情感词集合中，当该集合中已有该词时，则将其数量加 1，且标记为转折关系。如此在评论文本中对选取的种子词利用连词关系不断进行扩充，直到再没有可以扩充的候选情感词为止。假设“beautiful”和“good”

都是种子词，而例 1)中“beautiful”和“durable”具有并列关系，所以“durable”被抽取出来；例 3)中“good”和“expensive”具有转折关系，所以“expensive”被抽取出来。

在扩充得到的情感词集合中，通过一个种子词扩充得到的具有相同连词关系的词语数量来计算种子词和该词语的相似度。如种子词“good”扩充到三个具有并列关系的词语：“durable”、“beautiful”、“cheap”，数量分别为 3、4、3，则这三个词语与种子词“good”的相似度分别为 3/(3+4+3)，4/(3+4+3)，3/(3+4+3)。

4.5 利用标签传播算法构建情感词典

4.5.1 标签传播算法

标签传播算法(Label Propagation Algorithm，简称 LPA)是由 Zhu[57]等人于 2002 年提出的，它是一种基于图的方法。其思想是通过已知标签的节点预测未知标签节点的标签信息。标签传播算法根据节点之间的关系构建关系图，在图结构完成初始化后，图中节点可以分为已知标签的节点和未知标签的节点。两个节点的相似性主要通过它们的边进行判断，在不断的迭代过程中，未知标签的节点将根据它相邻节点的标签信息来得到自己的标签。

根据标签传播算法的思想，节点之间的标签传播主要通过相似度进行传播。在标签传播的过程中，未知标签的节点通过已知标签的相邻节点信息不断迭代来更新自己的标签。若相邻节点与它的相似度越大，则对其标签影响的权重越大，也就更容易进行标签的传播。标签传播算法不仅简单，而且算法的执行时间较短，可以取得较好的效果。

标签传播算法的过程如下：$(x_1,y_1)...(x_m,y_m)$作为已知标签的节点信息，其中$Y_L=\{y_1...y_m\}\in\{1...C\}$是节点的标签。假设类别$C$是已知的，并且有标签的节点包含了所有的类别。$(x_{m+1},y_{m+1})...(x_{m+n},y_{m+n})$作为无标签的节点信息，其中$Y_U=\{y_{m+1}...y_{m+n}\}$是未知的标签，通常$m<<1$。数据集$X$可表示为$X=\{x_1...x_{m+n}\}\in R_D$。接着在给定的数据集$X$中，通过$Y_U$的学习过程对无标签数据集$Y_U$完成标签地更新。

4.5.2 情感词典的构建

首先，将 4.3 节和 4.4 节中由种子词扩充得到的词语和种子词都作为一个节点。如果词语 a 通过 Word2Vec 可以扩充得到词语 b，则 a 和 b 之间有一条边，权重为词语 a 和 b 的相似度；如果词语 a 和词语 c 具有连词关系，则 a 和 c 之间有一条边，权重为词语 a 和 c 的相似度。因此，所有抽取出的词语和种子词被抽象为一张图，如图 4.2 所示。

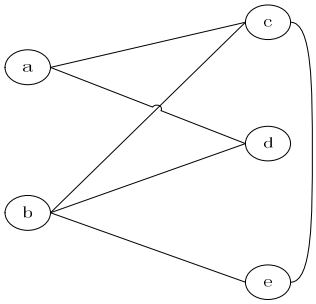

图 4.2 节点抽象的结构图

假设图中共有 m 个节点，则可构建一个 m 维的相似度概率转移矩阵。如公式 4.2 所示：

$$T[i][j]=\frac{SIM(w_i,w_j)}{\sum_{j=0}^{j<m}SIM(w_i,w_j)} \tag{4.2}$$

其中，$T[i][j]$ 表示词语 i 到 j 的相似度转移概率，$SIM(w_i,w_j)$ 表示它们的相似度。

在图 4.2 中，假设 a 和 b 分别是积极种子词和消极种子词，极性分别记为+1 和-1。其余词语的极性未知，记为 0。则词语 $a\sim e$ 初始情感极性如公式 4.3 所示：

$$V=\begin{bmatrix}+1\\-1\\0\\0\\0\end{bmatrix} \tag{4.3}$$

然后利用相似度概率转移矩阵和词语的初始情感极性进行不断地迭代，得到每个未知极性的词语极性。计算方法如公式 4.4 所示：

$$PO[i]=\sum_{j=0}^{j<m}T[j][i]*V[j] \tag{4.4}$$

其中，$PO[i]$表示迭代后的节点i的情感极性，$T[j][i]$表示节点j到节点i的相似度矩阵的转移概率，$V[j]$表示迭代前节点j的初始情感极性。在每轮迭代中，种子词的极性保持不变。经过不断迭代，直到图中词语的极性不再发生任何变化为止，则得到了每个未知标签的词语极性。最终将词语极性的绝对值大于某一阈值的词语作为构建得到的情感词典。

本文情感词典构建方法的算法描述如图 4.3 所示：

输入：评论语料(Review Corpus，简称 RC)，情感种子词SD，利用 Word2Vec 扩充词语的阈值wv，存放情感词的集合U，确定是否是积极情感词的阈值p_{pos}，确定是否是消极情感词的阈值p_{neg}

输出：积极的情感词典(Positive Sentiment Lexicon，简称SL_{pos})和消极的情感词典(Negative Sentiment Lexicon，简称SL_{neg})

1. 初始化情感词典，令$SL_{pos}=\varnothing$，$SL_{neg}=\varnothing$

2. 对语料 RC 进行词性标注

3. 对每个种子词$SD_i \in SD$：

(1) 利用 Word2Vec 训练语料 RC，若 RC 中的词语SW_i和种子词SD_i的相似性大于阈值wv，则$U=U\cup SD_i \cup SW_i$，同时记录词语之间的相似度

(2)对语料 RC 进行依存句法分析，若 RC 中的词语SW_j和种子词SD_i具有连词关系，则$U=U\cup SD_i \cup SW_j$

(3) 计算(2)中词语SW_j和种子词SD_i的相似度

4. 通过步骤 3 得到了所有扩充的词语以及词语之间的相似度值矩阵，再根据公式 4.2 得到词语之间的相似度概率转移矩阵

5. 在整个图中，根据公式 4.4 以及词语的初始情感极性V计算未知极性的词语极性PO

6. 令$V=PO$

7. 重复步骤 4-6，直至整个图中词语的极性不再发生任何变化

8. 得到最终词语极性的向量PO。PO中每个值代表每个词语的情感倾向。如果词语SW_k的极性$PO_k>0$且$|PO_k|>p_{pos}$，则$SL_{pos}=SL_{pos}\cup SW_k$，如果词语$SW_k$的极性$PO_k<0$且$|PO_k|>p_{neg}$，则$SL_{neg}=SL_{neg}\cup SW_k$

图 4.3 W2V&CR-LP 方法的算法描述

4.6 实验与分析

4.6.1 实验数据及评价标准

1. 语料：本章语料来自斯坦福大学所提供的亚马逊网站的相关评论，评论有书籍、手机、衣服、电子产品以及电影评论等，本章选择其中的手机评论[58]作为实验语料，共 194185 条。

2. 评价标准：本文利用人工评判的方法对情感词典进行评估，具体做法是选取情感词典中一定数量的词语，人工判断它们的极性是否正确，然后通过计算情感词典正确率 A 来衡量构建的情感词典性能[36]。

4.6.2 基线方法

为了说明本文提出的情感词典构建方法的有效性，本文选取了以下几种方法作为基线方法：

1. 基于 WordNet 的情感词典构建方法(A method of constructing sentiment lexicon based on WordNet，简称 WN)：直接利用 WordNet 语义知识库对情感种子词进行同义词的迭代抽取，得到构建的情感词典。

2. 基于连词关系的情感词典构建方法(A method of constructing sentiment lexicon based on conjunctive relations，简称 CR)：分析语料，迭代抽取和种子词具有连词关系的词语。和种子词具有并列关系的词语极性和种子词极性相同，和种子词具有转折关系的词语极性和种子词极性相反，得到构建的情感词典。

3. 基于 Word2Vec 的情感词典构建方法(A method of constructing sentiment lexicon based on Word2Vec，简称 W2V)：利用 Word2Vec 在语料上训练词向量，然后迭代计算语料中词语和情感种子词之间的语义相似度，相似度大于某个阈值的词语和该种子词的极性相同，以此得到构建的情感词典。

4. 基于连词关系和标签传播的情感词典构建方法(A method of constructing sentiment lexicon based on conjunctive relations and label propagation，简称 CRLP)：利用连词关系抽取候选情感词后，然后利用标签传播算法确定候选情感词的极性以及构建情感词典。

5. 基于 Word2Vec 和标签传播的情感词典构建方法(A method of constructing sentiment lexicon based on Word2Vec and label propagation，简称 W2VLP)：利用 Word2Vec 在语料上训练词向量后，然后计算语料中词语和情感种子词之间的语义相似度，将相似度大于某个阈值的词语抽取出来作为候选情感词，最后利用标签传播算法确定候选情感词的极性以及情感词典。

4.6.3 实验分析

本文分别选取 20、30、50 个种子词进行实验，种子词的选取如 4.2 节所述。确定是否是积极情感词的阈值 p_{pos} 设为 0.01，确定是否是消极情感词的阈值 p_{neg} 设为 0.02。人工判断构建的情感词典的正确率，实验结果如表 4.2、4.3、4.4 所示。

表 4.2 种子词为 20 个的实验结果

实验方法	积极情感词典的正确率(%)	消极情感词典的正确率(%)	正确率(%)
WN	81.6	60.0	70.8
CR	79.0	62.8	70.9
W2V	74.0	76.6	75.3
CRLP	83.3	62.5	72.9
W2VLP	84.0	68.0	76.0
W2V&CR-LP	88.0	68.0	78.0

表 4.3 种子词为 30 个的实验结果

实验方法	积极情感词典的正确率(%)	消极情感词典的正确率(%)	正确率(%)
WN	80.0	52.0	66.0
CR	80.0	62.0	71.0
W2V	66.0	79.6	72.8
CRLP	83.3	66.7	75.0
W2VLP	84.0	70.0	77.0
W2V&CR-LP	87.0	70.0	78.5

表 4.4 种子词为 50 个的实验结果

实验方法	积极情感词典的正确率(%)	消极情感词典的正确率(%)	正确率(%)
WN	62.0	40.0	51.0
CR	82.0	66.0	74.0
W2V	70.0	74.0	72.0
CRLP	84.2	66.7	75.5
W2VLP	86.0	68.0	77.0
W2V&CR-LP	86.8	70.6	78.7

从表 4.2、4.3 和 4.4 中可以看出，基线实验中，WN 方法所得情感词典的正确率最低。原因在于 WordNet 是一个人工构建的语义知识库词典，在迭代扩充过程中所引入的噪声词比较多，并且覆盖面有限，对于不在 WordNet 中的情感词扩充不到，因而有相应局限。

相比于依赖语义知识库的 WN 方法而言，基于语料库的 CR 方法和 W2V 方法的实验结果均取得了较好效果，说明了基于语料库的情感词典构建方法的优越性。进一步，我们发现在表 4.2、4.3 和 4.4 中，W2VLP 方法的性能均好于 W2V 方法，CRLP 方法的性能也优于 CR 方法，这说明了标签传播算法的有效性。

在 W2V 和 W2VLP 方法中，对于语料中和种子词距离较远的情感词会由于相似度较小而被过滤掉。CR 和 CRLP 方法虽然可以将语料中和种子词具有连词关系的词语抽取出来作为情感词，但是对于其它和种子词没有连词关系的情感词却会被忽略。如“The cellphone is beautiful and durable，I love it very much”这一评论中，若“beautiful”作为种子词，利用 CR 或者 CRLP 方法可以抽取出和它并列的情感词“durable”，但是情感词“love”则因为和种子词没有连词关系而被忽略掉。因此，从上述分析可以看出，基于 Word2Vec 的方法和基于连词关系的方法各有优缺点。

相比于基线方法，本文提出的 W2V&CR-LP 方法在表 4.2、4.3 和 4.4 中的正确率均取得了最高值，相对于 CRLP 和 W2VLP 这两个性能较好的基线方法分别获得了 1.5%~5.1%间的性能提升，可见本文方法对于情感词典的构建有较大的优势。由于 W2V 方法可以抽取出和种子词距离较近的情感词，CR 方法可以抽取出和种子词具有连词关系的情感词，但如上文所述，各自也有一定的不足。W2V&CR-LP 方法结合这两种方法来抽取情感词，能较好的抽取出各自方法所扩充不到的情感词。同时，利用标签传播算法对扩充得到的候选情感词进行极性的确定。由于标签传播算法可以将种子词的情感极性通过边向相邻节点传播，从而使确定情感词极性的问题转化成了图中传播标签的问题。在图中，词语的极性不仅受到近距离种子词的影响，也受到其它种子词极性的影响。在每轮迭代中，距离近的词语通常具有相同的极性，通过多次迭代，标签在不断的传播，可以对极性标注不正确的情感词进行及时地更新，使最终扩充得到的情感词典的正确率更高。因此，本文所提出的 W2V&CR-LP 方法具有一定的优势。

随着种子词数的增加，本文方法的正确率也有所提高，从 78.0%提高到 78.5%，再由 78.5%提高到 78.7%。由于随着种子词数量的增加，可抽取到的和种子词相似度高的词语以及和种子词具有连词关系的词语也越多，同时进行标签传播算法已知极性的种子词数量也越多，因此，正确率会有所提高。

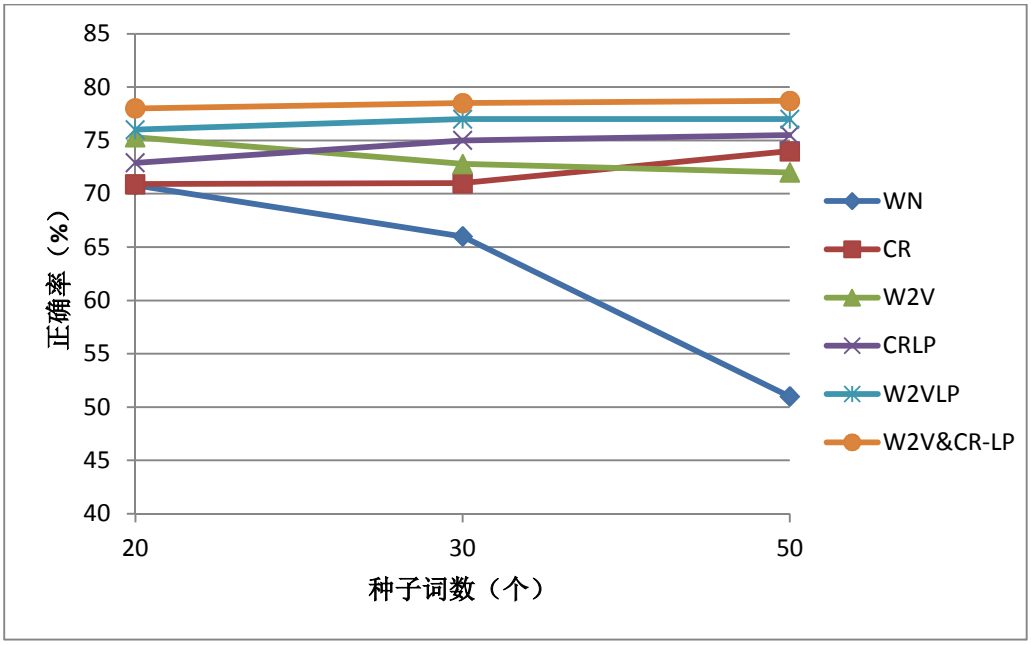

图 4.4　随种子词个数变化的实验结果

此外，本文方法也具有良好的鲁棒性。图 4.4 直观地展示了基线以及本文方法随种子词个数变化的正确率变化情况，从图 4.4 可以看出，本文方法较为稳定，在种子词数量变化的情况下性能均优于其它几种方法，且正确率随种子词数的增加也有所增加。

4.7　本章小结

本文提出了一种新方法进行情感词典的构建，首先选取部分情感种子词，然后利用 Word2Vec 对语料进行词向量训练，找出和情感种子词潜在相似度高的词语加入词语集合 a 中；再基于依存句法分析找出和情感种子词具有连词关系的词语加入词语集合 b 中，并分别计算词语之间的相似度。最后利用标签传播算法对所有扩充得到的词语进行标签传播，更新并最终确定构建的情感词典。实验结果表明该方法优于其它的基线方法。

第 5 章 基于产品评论的细粒度情感分析原型系统

5.1 引言

产品评论反映了用户对产品的真实感受，它不仅可以为潜在消费者购买某产品提供决策，而且可以将用户对产品的真实感受反馈给商家，使商家及时发现自己产品的不足及优势，为进一步改善产品质量提供针对性的依据。因此，对产品评论的挖掘具有重要意义。为了让消费者及商家以较少的时间和精力来查看评论，进一步了解产品特征的优劣，本文将产品特征和评价词语提取出来，并判定评价词语的褒贬，最后进行比较和分类，将产品特征的优势和不足以图形化界面展现出来。

本章首先描述了细粒度情感分析原型系统的系统架构及功能模块，然后对挖掘结果进行了展示。

5.2 系统架构及开发环境

5.2.1 系统架构

产品评论的细粒度情感分析原型系统的主要功能是当用户首先选择某个产品型号时，用户对应的可以选择该产品型号的产品特征，在选择自己想要查看的产品特征后，通过查询，即可直观地展示出用户对该产品特征评价的好坏。

原型系统的总体框架如图 5.1 所示。首先，对于某个特定的产品，本文通过网络爬虫将该产品相关的评论爬取出来，并将这些评论存入数据库中；然后对这些评论进行预处理，通过分词、词性标注以及依存分析等得到相应的结果；再利用基于领域相关性的产品特征抽取算法抽取出评论中的产品特征。然后利用基于标签传播的情感词典构建方法构建中文情感词典，再利用情感词典将距离产品特征最近的情感词抽取出来作为评价词语，同时得到评价词语的极性。最后，将产品特征及对应极性以图形化的界面展示出来。

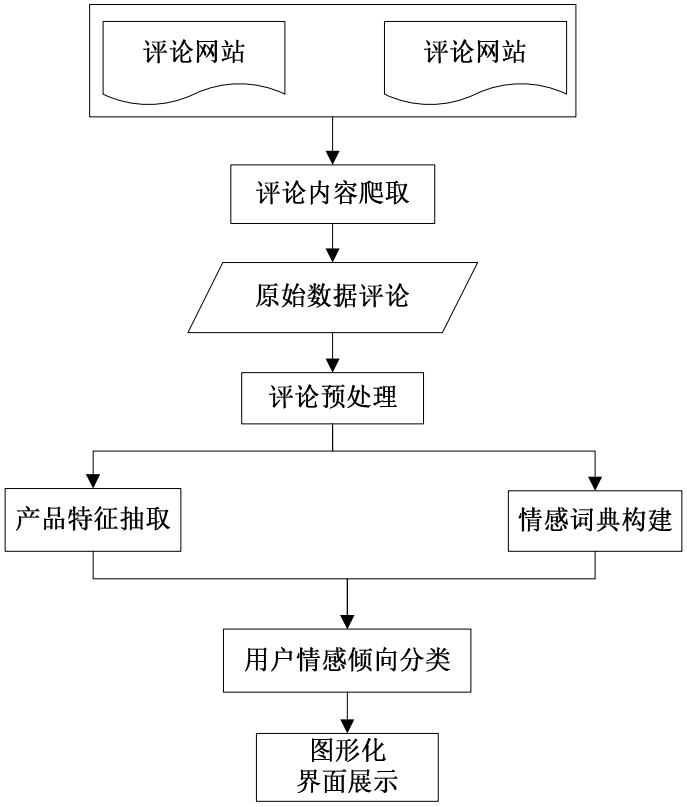

图 5.1 原型系统的总体框架

5.2.2 开发环境

原型系统的开发及运行环境为：

硬件环境：Intel(R) Core(TM) i5-4570 CPU @3.2GHz，8G 内存。

软件环境：Windows 7 操作系统，IntelliJ IDEA 2016，MySql 数据库。

5.3 系统实现与展示

5.3.1 系统实现

在实现原型系统前，需要抽取产品特征和评价词语。本文首先利用点对互信息合并该值大于某一阈值的相邻名词，然后根据制定的句法规则来抽取候选特征。对于每个候选特征，利用两个不同领域的语料来计算其领域相关性值，该值大于某一阈值的候选特征作为最终产品特征。本文将产品评论的预处理结果保存在表(word)中，其数据库表设计如表 5.1 所示。

表 5.1 产品评论预处理结果表

字段名称	类型	长度	注释
id	int	11	id
name	varchar	12	词语
attribute	varchar	12	词性标注
dependency	varchar	12	依存分析
file_name	varchar	12	所属文本
all_count	int	11	所有文本中出现的次数
all_txt	int	11	出现的本文数
pmi	varchar	15	互信息

对得到的预处理文本，利用领域相关性方法将抽取得到的候选产品特征存放在表(candidate)中，进一步通过计算领域相关性值得到产品特征。候选产品特征的数据库表设计如表 5.2 所示。

表 5.2 候选产品特征结果表

字段名称	类型	长度	注释
id	int	11	id
name	varchar	12	候选产品特征
attribute	varchar	12	词性标注
dependency	varchar	12	依存分析
file_name	varchar	12	所属文本
mrel	varchar	15	领域相关性值

然后将产品评论中距离产品特征最近的情感词抽取出来作为对该产品特征的评价词语。抽取的评价词语存放在表(sentiment)中，评价词语的数据库表设计如表 5.3 所示。

表 5.3 评价词语结果表

字段名称	类型	长度	注释
id	int	11	id
file_name	varchar	12	所属文本
comment	varchar	255	产品评论
feature	varchar	12	产品特征
sentiment	varchar	12	评价词语
polarity	varchar	6	情感极性

5.3.2 系统展示

对产品评论进行细粒度情感分析以后，本文得到了产品特征及其对应的极性，为了减少用户查看产品评论所花费的时间和耗费的精力，同时也为了更直观的向商家提供用户对产品的反馈信息，本文对产品特征进行了原型系统地展示。产品特征主要反映了产品评论中用户主要提及的产品特征有哪些，评价词语反映了用户对产品特征的主观感受，对该产品特征的评价是好的还是坏的。本文以图形化界面的方式将产品评论中对产品特征的评价直观展示出来，使用户及商家一目了然。

首先，用户或商家可以选择自己感兴趣的产品型号，在选择产品型号后，会出现对应的产品特征；然后，对哪个产品特征感兴趣，就选择哪个产品特征进行查询，系统则会显示出用户对所选产品的产品特征评价的好坏。产品型号的选择如图 5.2 所示。

图 5.2 产品型号选择界面

在选择产品型号后，可选择该产品型号对应的产品特征，如图 5.3 所示。

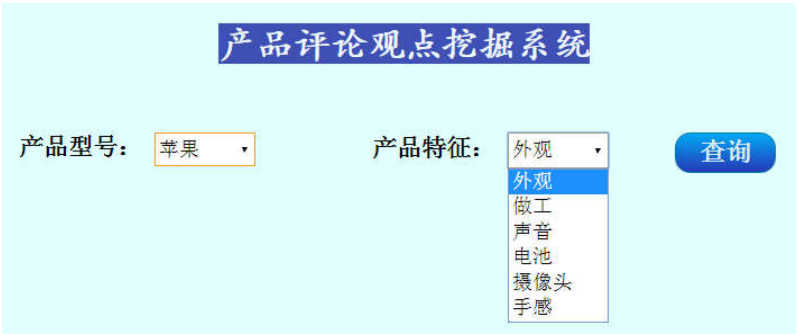

图 5.3 产品特征选择界面

在选择了需要查看的产品型号及对应产品特征后，通过点击“查询”按钮，即可查看评论者对该产品特征评价的极性。本文将产品评论分为积极和消极两大类，原型系统展示如图 5.4 所示。

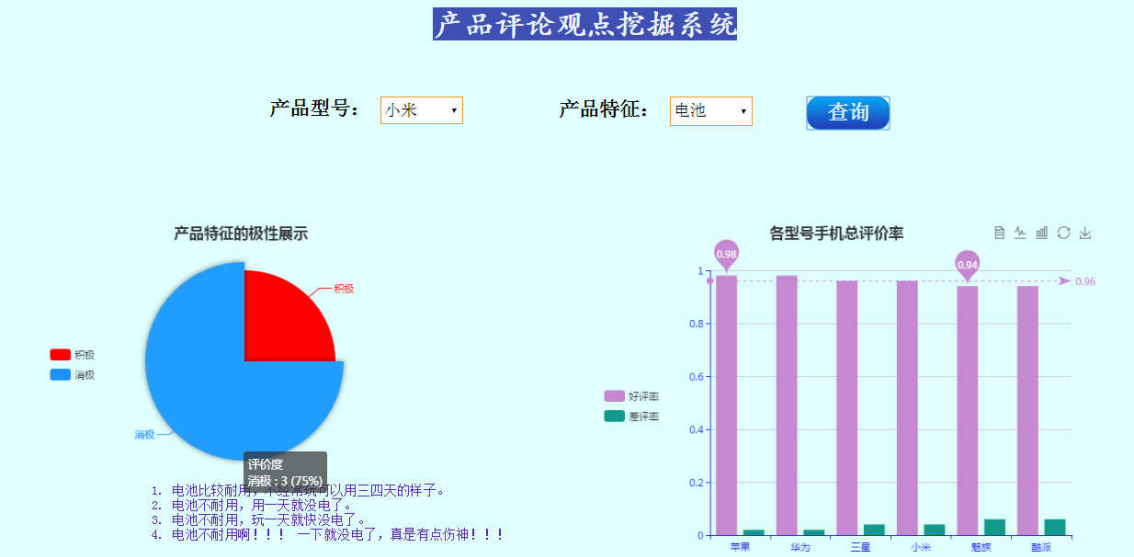

图 5.4 原型系统的图形化界面展示

在图 5.4 中，饼形图中红色表示产品特征的极性是积极的，蓝色表示极性是消极的。图中饼形图是对评论中产品“小米”及其对应产品特征“电池”的评价倾向判断，当鼠标移动到饼形图时，可以看到对小米电池的负面评价有 3 条，占 75%；正面评价有 1 条，占 25%。图中柱形图展示了各型号手机的总体评价率，如苹果手机的好评率达到 98%。

5.4 本章小结

为了对产品评论进行自动化分析，本章有效利用产品评论实现了基于产品评论的细粒度情感分析原型系统。该系统可以找出产品评论中产品型号对应的产品特征、产品特征的评价词及对应极性。本文首先通过领域相关性算法抽取产品特征，然后利用情感词典将距离产品特征最近的情感词语抽取出来作为评价词语，同时确定了该评价词语的极性。最后将这些产品特征及对应极性以图形化界面展示出来，使用户和商家一目了然地了解自己感兴趣的产品。

第 6 章 总结与展望

6.1 全文总结

情感分析作为自然语言处理的一个重要研究方向，可以挖掘出大量有价值的信息。而对产品评论进行细粒度的情感分析，抽取归纳出消费者对产品特征的情感倾向，无论对于潜在消费者还是商家，都具有极其重要的意义。

本文主要针对产品评论进行细粒度情感分析研究，着重研究了其中最主要的两个问题：产品特征的抽取和情感词典的构建，实现了产品评论的细粒度情感分析原型系统。下面对论文的主要工作进行总结：

1. 分析了对产品评论进行细粒度情感分析的必要性，研究了近年来情感分析中产品特征抽取以及情感词典构建所用的方法，总结各方法的优缺点，为进一步的研究做理论研究和技术准备。

2. 提出了基于领域相关性的产品特征抽取方法。该方法首先通过一定的句法规则抽取候选特征；然后利用候选特征在领域相关性和领域无关性两种领域语料中的差异性，再结合似然比技术，得到每个候选特征的领域相关性值；最后通过领域相关性值来确定产品特征。利用领域相关性抽取产品特征，不仅可以减少人工的工作量和计算量，而且可移植性强。实验表明，该方法取得了较好的结果。

3. 提出了基于标签传播的情感词典构建方法。该方法首先选取一定数量的情感种子词，然后利用 Word2Vec 找出和种子词具有潜在相似性的词语，同时利用依存句法分析找出和种子词具有连词关系的词语；再通过标签传播算法对扩充得到的词语不断迭代进行标签的传播，最终确定扩充的情感词语及其对应极性。实验结果证明了该方法的有效性。

4. 结合细粒度情感分析的实际需求，设计并实现了一个基于产品评论的细粒度情感分析原型系统。用户可以在该系统上选择自己感兴趣的产品型号及对应的产品特征，进而查看评论者对该产品特征评价的好坏以及所占百分比，以可视化图形进行展示。

6.2 工作展望

本文主要针对产品评论中产品特征抽取以及情感词典的构建进行了研究，提出了基于领域相关性的产品特征抽取方法以及基于标签传播的情感词典构建方法，然后对所提方法以及其它基线方法进行实验，并做了实验结果对比以及结果分析。细粒度情感分析里有很多方面值得研究和探索，由于作者的研究时间及能力有限，本文的研究并不是很全面。在未来的工作中，可进一步研究的方向有：

1. 针对产品特征的抽取，本文只考虑了显式的产品特征，并未考虑隐式特征。如在产品评论“苹果手机不错，但是有点贵”中可以看出，“贵”是用来评价“价格”的，但是在评论中并没有“价格”这个产品特征，所以它是隐式的产品特征。下一步将考虑隐式产品特征的抽取。

2. 针对情感词典的构建，本文方法结果较好，但是还有提升空间。在下一步工作中，提高情感词典的准确率以及情感词典的覆盖面，同时将本文方法运用在不同领域语料的情感词典构建中，是研究的重点。

3. 由于时间的限制，本文所实现的细粒度情感分析的原型系统不是很完善。因此，扩充该系统的功能也是下一步要做的工作。

参考文献

[1] 中国互联网络信息中心. CNNIC 发布第 39 次《中国互联网络发展状况统计报告》[J]. 中国信息安全, 2017: 24.

[2] 翟东升, 徐颖, 黄鲁成等. 基于产品评论挖掘的竞争产品优势分析[J]. 情报杂志, 2013, 32(2): 45-51.

[3] Cambria E, Schuller B, Xia Y, et al. New avenues in opinion mining and sentiment analysis[J]. IEEE Intelligent Systems, 2013, 28(2): 15-21.

[4] 赵妍妍, 秦兵, 刘挺. 文本情感分析[J]. 软件学报, 2010, 21(8): 1834-1848.

[5] Liu Bing. Sentiment analysis and opinion mining[J]. Synthesis lectures on human language technologies, 2012, 5(1): 1-167.

[6] Quan Changqin, Ren Fuji. Target based review classification for fine-grained sentiment analysis[J]. International Journal of Innovative Computing, Information and Control, 2014, 10(1): 257-268.

[7] 张奇. 细颗粒度情感倾向分析若干关键问题研究[D]. 上海: 复旦大学, 2008.

[8] Jin W, Ho H H, Srihari R K. OpinionMiner: a novel machine learning system for web opinion mining and extraction[C]//Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining. New York : ACM, 2009: 1195-1204.

[9] Hu Minqing, Liu Bing. Mining opinion features in customer reviews[C]// American Association for Artificial Intelligence. San Jose: AAAI, 2004, 4(4): 755-760.

[10] Popescu A M, Etzioni O. Extracting product features and opinions from reviews[M]//Natural language processing and text mining. London: Springer, 2007: 9-28.

[11] Blair-Goldensohn S, Hannan K, McDonald R, et al. Building a sentiment summarizer for local service reviews[C]//WWW workshop on NLP in the information explosion era, 2008, 14: 339-348.

[12] Long Chong, Zhang Jie, Zhu Xiaoyan. A review selection approach for accurate feature rating estimation[C]//Proceedings of the 23rd International Conference on Computational Linguistics: Posters. Association for Computational Linguistics. Stroudsburg: ACL, 2010: 766-774.

[13] 邱培超. 基于特征的观点挖掘中的若干关键问题研究[D]. 上海: 复旦大学, 2011.

[14] Somasundaran S, Ruppenhofer J, Wiebe J. Discourse level opinion relations: An annotation study[C]//Proceedings of the 9th SIGdial Workshop on Discourse and Dialogue. Stroudsburg: ACL, 2008: 129-137.

[15] Qiu Guang, Liu Bing, Bu Jiajun, et al. Opinion word expansion and target extraction through double propagation[J]. Computational linguistics, 2011, 37(1): 9-27.

[16] 唐晓波, 肖璐. 基于依存句法网络的文本特征提取研究[J]. 现代图书情报技术, 2014, 11: 31-37.

[17] 赵妍妍, 秦兵, 车万翔等. 基于句法路径的情感评价单元识别[J]. 软件学报, 2011, 22(5): 887-898

[18] Ferreira L, Jakob N, Gurevych I. A comparative study of feature extraction algorithms in customer reviews [C]// 2008 IEEE International Conference on Semantic Computing. Denver: IEEE, 2008: 144-151.

[19] Khan K, Baharudin B, Khan A. Semantic-based unsupervised hybrid technique for opinion targets extraction from unstructured reviews[J]. Arabian Journal for Science and Engineering, 2014, 39(5): 3681-3689.

[20] Boer N D, Leeuwen M V, Luijk R V, et al. Identifying explicit features for sentiment analysis in consumer reviews[C]//International Conference on Web Information Systems Engineering. Greece: Springer, 2014: 357-371.

[21] 郁亚辉. 产品评论挖掘中特征同义词的识别[J]. 中文信息学报, 2016, 30(4): 150-158.

[22] Hofmann T. Probabilistic latent semantic indexing[C]//Proceedings of the 22nd Annual International SIGIR Conference on Research and Development in Information Retrieval. New York: ACM, 1999: 50-57.

[23] Steyvers M, Griffiths T. Probabilistic topic models[M]. Handbook of latent semantic analysis, 2007, 427(7): 424-440.

[24] Mei Qiaozhu, Ling Xu, Wondra M, et al. Topic sentiment mixture: modeling facets and opinions in weblogs[C]//Proceedings of the 16th international conference on World Wide Web. New York: ACM, 2007: 171-180.

[25] Zhao W X, Jiang J, Yan H, et al. Jointly modeling aspects and opinions with a MaxEnt-LDA hybrid[C]//Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing. Cambridge: ACM, 2010: 56-65.

[26] Sauper C, Haghighi A, Barzilay R. Content models with attitude[C]//Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies. Portland: ACL, 2011: 350-358.

[27] 刘羽, 曹瑞娟. 基于观点挖掘的产品特征提取[J]. 计算机应用与软件, 2014, 31(1): 81-84.

[28] Zhang Shu, Jia Wenjie, Xia Yingju, et al. Opinion analysis of product reviews[C]// 2009 Sixth International Conference on Fuzzy Systems and Knowledge Discovery. Washington: IEEE, 2009: 591-595.

[29] 王山雨. 面向产品领域的细粒度情感分析技术[D]. 哈尔滨: 哈尔滨工业大学, 2011.

[30] 刘丽, 王永恒, 韦航. 面向产品评论的细粒度情感分析[J]. 计算机应用, 2015, 35(12): 3481-3486.

[31] 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): 56-62.

[32] 戴敏, 王荣洋, 李寿山等. 基于句法特征的评价对象抽取方法研究[J]. 中文信息学报, 2014, 28(4): 92-97.

[33] 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.

[34] 张玥. 面向产品评价的细粒度情感分析技术研究[D]. 哈尔滨: 哈尔滨工业大学, 2013.

[35] 丁晟春, 吴婧婵媛, 李霄. 基于 CRFs 和领域本体的中文微博评价对象抽取研究[J]. 中文信息学报, 2016, 30(4): 159-166.

[36] 王科, 夏睿. 情感词典自动构建方法综述[J]. 自动化学报, 2016, 42(4): 495-511.

[37] Neviarouskaya A, Prendinger H, Ishizuka M. SentiFul: A lexicon for sentiment analysis[J]. IEEE Transactions on Affective Computing, 2011, 2(1): 22-36.

[38] Hassan A, Abu-Jbara A, Jha R, et al. Identifying the semantic orientation of foreign words[C]//Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics. Portland: ACL, 2011: 592-597.

[39] Kanayama H, Nasukawa T. Fully automatic lexicon expansion for domain-oriented sentiment analysis[C]//Proceedings of the 2006 conference on empirical methods in natural language processing. Sydney: ACL, 2006: 355-363.

[40] Huang Sheng, Niu Zhengdong, Shi Chongyang. Automatic construction of domain-specific sentiment lexicon based on constrained label propagation[J]. Knowledge-Based Systems, 2014, 56(3): 191-200.

[41] 郗亚辉. 产品评论中领域情感词典的构建[J]. 中文信息学报, 2016, 30(5): 136-144.

[42] Peng Wei, Park D H. Generate adjective sentiment dictionary for social media sentiment analysis using constrained nonnegative matrix factorization[J]. Proceedings of the 5th International AAAI Conference on Weblogs and Social Media, 2004, 51: 61801.

[43] Rao D, Ravichandran D. Semi-supervised polarity lexicon induction[C]//Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics. Athens: ACL, 2009: 675-682.

[44] 杨阳, 刘龙飞, 魏现辉等. 基于词向量的情感新词发现方法[J]. 山东大学学报(理学版), 2014, 49(11): 51-58.

[45] Hinton G E. Learning distributed representations of concepts[C]//Proceedings of the eighth annual conference of the cognitive science society. Chicago: Amherst, 1986, 1: 12.

[46] WordNet 简介[OL]. URL: [2015-3-17].http://wordnet.princeton.edu.

[47] Church K W, Hanks P. Word association norms, mutual information, and lexicography[C]// Meeting on Association for Computational Linguistics. Vancouver: ACL, 1989: 76-83.

[48] Zhao Qiyun, Wang Hao, Lv Pin, et al. A bootstrapping based refinement framework for mining opinion words and targets[C]//Proceedings of the 23rd ACM International Conference on Conference on Information and Knowledge Management. Shanghai: ACM, 2014: 1995-1998.

[49] Zhao Li, Huang Minlie, Sun Jiashen, et al. Sentiment extraction by leveraging aspect-opinion association structure[C]//Proceedings of the 24th ACM International on Conference on Information and Knowledge Management. Melbourne: ACM, 2015: 343-352.

[50] Hai Zhen, Chang Kuiyu, Kim J J, et al. Identifying features in opinion mining via intrinsic and extrinsic domain relevance[J]. IEEE Transactions on Knowledge and Data Engineering, 2014, 26(3): 623-634.

[51] Cohen J. A coefficient of agreement for nominal scales[J]. Educational and psychological measurement, 1960, 20(1): 37-46.

[52] 杜嘉忠, 徐健, 刘颖. 网络商品评论的特征-情感词本体构建与情感分析方法研究[J]. 现代图书情报技术, 2014, 30(5): 74-82.

[53] Baccianella S, Esuli A, Sebastiani F. SentiWordNet 3.0: An enhanced lexical resource for sentiment analysis and opinion mining[C]// Proceedings of the International Conference on Language Resources and Evaluation. Melbourne: LREC 2010, 10: 2200-2204.

[54] Mikolov T, Chen K, Corrado G, et al. Efficient estimation of word representations in vector space[J]. arXiv preprint arXiv: 1301.3781, 2013.

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

[56] Wu Yuanbin, Zhang Qi, Huang Xuanjing, et al. Phrase dependency parsing for opinion mining[C]//Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing. Singapore: ACL, 2009: 1533-1541.

[57] Zhu Xiaojun, Ghahramani Z. Learning from labeled and unlabeled data with label propagation[J]. 2002.

[58] McAuley J, Targett C, Shi Q, et al. Image-based recommendations on styles and substitutes[C]//Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval. Santiago: ACM, 2015: 43-52.

致谢

时间过的好快，不知不觉，我的研究生生活就要结束了。研究生三年，另加本科四年，在重庆邮电大学的七年时间，是我一生最难忘的日子。我在这里学习知识，收获友情，在母校的光辉下快乐成长。现在即将毕业，有太多的不舍。

研究生三年，实验室成了我生活和学习的中心，每天两点一线，往返于宿舍和信科。无论在学习，还是生活上，实验室的老师和同学们都教会了我很多。现在能够顺利毕业，真的特别感谢帮助和支持我的每一个人，是你们对我无私的帮助，才有了现在的我。

首先，我必须要感谢我的导师张璞老师。是您，以身作则，教会了我很多。您教导我们，在科研上要一丝不苟，对自己感兴趣的方向，要认认真真做实验，并要善于总结，发现新的问题并解决。无论在学习还是生活上，不管我们有什么问题向您请教，您总是以谦和的态度及时帮我们解答。我以我有您这样的导师感到自豪，您也是我这一生学习的榜样。

同时，对我们同一个团队的熊安萍老师以及蒋溢老师提出深深地感谢。在研究生期间，两位老师给我们提供了舒适的实验室环境，并给我们提供了参与项目开发的实践机会，让我的动手能力大大提升。熊老师您对工作的负责态度以及严谨的教学作风对我未来的工作产生着巨大的影响；而蒋老师您与人为善、快乐幽默的生活作风对我的人际交往以及团队合作上有巨大的帮助。

我要感谢师姐冯浩、王运萍，我也要感谢实验室的王浪、杨方方、朱恒伟、汤婷婷、夏玉冲、罗宇豪、黎海青、印振宇、王英豪、刘杰、王志迁等同学，你们在生活、学习以及工作上给了我很大的支持和鼓励，也带给了我很多快乐。我还要感谢室友王梅和王文斌，你们在三年的研究生生活中帮助了我很多。

我要感谢我的父母和亲人，是你们对我无私的付出，让我每一天都过的这么美好。对你们我无以为报，只有自己努力学习和工作，尽我所能来让你们享受幸福生活。

最后对百忙之中评阅论文和参加答辩的各位专家表示衷心的感谢。

再次真挚感谢所有关心和帮助过我的人。

攻读硕士学位期间从事的科研工作及取得的成果

参与科研项目：

[1] 面向产品评论的细粒度观点挖掘方法研究(cstc2015jcyjA40025), 重庆市科委(重庆市前沿与应用基础研究(一般)项目), 2015.8-2018.7.

[2] 面向 Web 评论文本的情感分析方法研究(KJ1600440), 重庆市教委科技项目, 2016.8-2018.7.

[3] 产品评论观点挖掘系统的研究(CYS15171), 重庆市研究生科研创新项目, 2015.9-2017.6.

发表及完成论文：

[1] Pu Zhang, **Junxia Wang**, Yinghao Wang. A statistical approach to opinion target extraction using domain relevance[C]. Proceedings of 2016 2rd IEEE international conference on computer and communications, 2016.

[2] 张璞, **王俊霞**, 王英豪. 基于标签传播的情感词典构建方法[J]. 计算机工程.(已录用)

[3] **Junxia Wang**, Pu Zhang, Yinghao Wang. A mixed approach of expanding sentiment lexicon based on word2vec and phrase dependency tree[C]. Proceedings of 2017 the 7th international workshop on computer science and engineering, 2017.(已录用)

[4] Pu Zhang, Yinghao Wang, **JunXia Wang**, Xianhua Zeng, Yong Wang. An improved term weighting scheme for sentiment classification[C]//Proceedings of 2017 IEEE 2nd advanced information technology, electronic and automation control conference, 2017:462-466.