蘇州大學

SOOCHOW UNIVERSITY

博士学位论文

论文题目 细粒度情感分析研究

研究生姓名 施寒潇

指导教师姓名 钱培德　周国栋

专业名称 计算机应用技术

研究方向 自然语言处理

论文提交日期 2013 年 5 月

苏州大学学位论文独创性声明

本人郑重声明：所提交的学位论文是本人在导师的指导下，独立进行研究工作所取得的成果。除文中已经注明引用的内容外，本论文不含其他个人或集体已经发表或撰写过的研究成果，也不含为获得苏州大学或其它教育机构的学位证书而使用过的材料。对本文的研究作出重要贡献的个人和集体，均已在文中以明确方式标明。本人承担本声明的法律责任。

论文作者签名：[signature] 日　期：2013.3.23

苏州大学学位论文使用授权声明

本人完全了解苏州大学关于收集、保存和使用学位论文的规定，即：学位论文著作权归属苏州大学。本学位论文电子文档的内容和纸质论文的内容相一致。苏州大学有权向国家图书馆、中国社科院文献信息情报中心、中国科学技术信息研究所（含万方数据电子出版社）、中国学术期刊（光盘版）电子杂志社送交本学位论文的复印件和电子文档，允许论文被查阅和借阅，可以采用影印、缩印或其他复制手段保存和汇编学位论文，可以将学位论文的全部或部分内容编入有关数据库进行检索。

涉密论文□

本学位论文属　　在_____年___月解密后适用本规定。

非涉密论文☑

论文作者签名：方海亮　日　期：2013.3.23

导师签名：钱伟长　日　期：2013.3.23

细粒度情感分析研究
中文摘要

从本世纪初开始，文本情感分析研究渐渐成为信息抽取领域中的研究热点，获得了越来越多的关注。随着互联网的飞速发展，特别是 Web2.0 技术的逐渐普及，广大网络用户已经从过去单纯的信息获取者变为网络内容的主要制造者。同时，随着自然语言处理技术和机器学习技术的不断发展和成熟，对主观性文本进行情感分析也成为可能，并逐渐得到广泛应用。

传统的文本情感分析研究主要面向篇章和句子级别文本，实现相应的情感极性判定。这些研究在一些应用领域，如网络舆情分析、股评分析等，已经体现出较好的应用价值。然而，随着应用的深入，用户提出了更高的要求，比如希望进一步获得评价对象属性所对应的具体情感分析结果。在这方面，传统的情感分析已不能完全满足需求。因此，本文提出了细粒度情感分析方法来迎接这个挑战。通过探索新的研究思路和方法，进一步提高情感分析的准确性和实用性。

本文针对细粒度情感分析中的关键技术展开研究，研究内容主要包括以下三个部分：

1. 研究了情感词极性强度量化方法。情感词的极性判定研究已相对成熟，但为了实现细粒度情感分析任务，还需要进行极性强度的量化计算，以满足情感统计的需求。我们在现有情感词极性强度量化算法的基础上，提出了改进方案：首先，对情感词进行分类；然后，针对不同类型的情感词设计不同的计算规则和方法。该方法的优势在于其能够充分利用了字词之间的关系以及语言学知识。

2. 研究了评价对象属性及其情感表达元素的联合识别。在细粒度情感分析任务中，如何正确识别出文本中的评价对象属性及其情感表达元素具有十分重要的意义。本文结合条件随机场理论，充分利用评价对象属性及其情感表达元素之间的类别关系，提出了序列化联合抽取模型。此外，还分析了基本特征和语义特征的相关知识及抽取方法，特别针对语义特征的抽取进行了技术分析和算法设计。

3. 研究了基于半监督学习的属性分类以及情感计算。针对细粒度情感标注语料的开放资源少、标注工作量大等难点问题，本文在属性分类研究中引入了半监督学习

机制，以减少对标注语料的依赖。首先，研究了自举学习的分层种子选取策略，并与随机种子选取策略在属性分类上进行了实验性能的对比；其次，研究了把分层思想应用到自举过程的每一步迭代之中，探讨了自举迭代的终止条件；最后，针对评论中可能存在情感词缺少对象属性的情况，我们研究通过计算 PMI 值来确定评价对象属性类与情感词之间的关联概率，实现对缺失评价对象属性的情感信息进行合理属性类的指派，使情感汇总计算更为合理有效。

本文的主要贡献总结如下：首先，针对细粒度的情感分析特点，在理论上对情感极性强度的模糊性特点进行了详细分析和研究，充分利用了字词关系和语言学知识，优化了情感词极性强度量化方法，在性能上达到了一定的提升；其次，在评价对象属性及其情感表达元素的联合识别研究中，提出了序列化联合抽取模型，充分利用了评论语句中的基本特征和语义特征信息，并通过调整 CRF 分类器的模板，进一步分析了特征组合以及上下文信息对识别性能的影响，获得了识别效率的提升；再其次，还对细粒度属性分类及文本情感计算进行了相关研究，证明了半监督学习方法在属性分类中的有效性，同时通过设计合理的情感计算方法完成基于属性类的情感汇总，实现了细粒度情感统计的目的；最后，设计了一个基于细粒度情感分析方法的酒店评论意见挖掘系统，有效地实现系统内部核心功能的封装，并提供了友好的用户界面展示。

关键词：情感分析，意见挖掘，情感强度量化，语义角色标注，统计机器学习，自然语言处理

作　　者：施寒潇

指导老师：钱培德
周国栋

Research on Fine-grained Sentiment Analysis

Abstract

Sentiment analysis, as a hot research topic in the research area of information extraction, has attracted more and more attention from the beginning of this century. With the rapid development of the Internet, especially the rising popularity of Web2.0 technology, the network user has become not only the content maker but also the receiver of information. Meanwhile, benefiting from the development and maturity of the technology in natural language processing and machine learning, it becomes possible to widely employ sentiment analysis on subjective texts.

Existing studies on sentiment analysis are mainly focusing on the task of determining document-level and sentence-level sentimental polarity. This task has reflected its valuable applications in some real applications, such as analysis of internet public opinions and stocks review. However, with the deepening of its application, the user puts forward some higher requests. For instance, some users hope to get analysis results of target attribute's sentiment. In this situation, the traditional technologies in sentiment analysis are not capable of meeting such novel demands. Therefore, in this study, we propose the method on fine-grained sentiment analysis to meet the new challenges by exploring new research ideas and methods to further improve the accuracy and practicability of sentiment analysis.

This paper focuses on the key technology in the analysis of fine-grained sentiment and the content includes:

1) The quantification on the polarity strength of a sentiment word. The research of deciding sentiment words' polarity has relatively matured. However, in order to realize the task of fine-grained sentiment analysis, we need to calculate quantified sentiment strength to meet the need of sentiment counting. Based on the existing algorithm on quantifying sentiment strength, we present an improved strategy: First, we classify the sentiment words

into different categories; Second, for each category, we design different calculation rules and methods to quantify the sentiment strength. The main advantegy of the proposed approach lies in its making full use of the relationship between characters and words, as well as the linguistics knowledge.

2) The joint model for recognization of the target attribute and its sentiment expression. In the task of fine-grained sentiment analysis, it is important to correctly recognize the target attribute and the sentiment expression in the text. Combining with the theory of Conditional Random Fields, we make effective use of the class-relation between the target attribute and its sentiment expression and introduce a joint recognition model based on the sequence structure. Additionally, this paper analyzes the related knowledge of the basic and semantic features and the extraction methods. Specifically, this paper analyzes the extraction of semantic features and designs a novel algorithm.

3) The classification method on attributes based on semi-supervised leaning and sentiment calculation. This paper proposes a semi-supervised learning method into the research of attribute classification to reduce the dependence of tagged corpus so as to overcome the difficulties on annotation work on fine-grained sentiment tagged corpus. First, this paper studies the initial seed selection strategy based on stratified sampling and compares it with those not in the performance of the experiment. Second, this paper employs the application of this selection strategy to each step of iteration at bootstrapping process and discusses the termination condition of bootstrapping iteration. Third, for the review which might take sentiment word lacking of target attribute, the PMI is adopted to determine the association probability with target attributes and sentiment words. In this way, the proposed approach is able to realize the reasonable classification of sentiment word lacking of target attribute, and make the sentiment summing more reasonable and effective.

The main contribution of this paper is summarized as follows: First, in theory, it carries on a detailed analysis and research of the characteristic fuzziness on sentiment strength. For the task of sentiment analysis, it makes full use of the relationship between characters and words, as well as the linguistics knowledge, and optimizes the method to

quantify sentiment strength, which achieves some certain improvements in performance. Second, in the research of jointly recognizing the target attributes and the corresponding sentiment expressions, this paper proposes a joint recognition model based on the sequence structure, making full use of the basic features and semantic features in the review. Moreover, by adjusting the template of CRF classifier, this paper further analyzes the effect of feature combination and context information on recognition performance which increase the efficiency in ascension. In addition, this paper also includes the related research of attribute classification and sentiment calculation, verifies the validity of semi-supervised leaning method in attribute classification. Third, by designing reasonable sentiment calculation method, it completes sentiment summing based on attribute classification and realizes the fine-grained sentiment statistic. Finally, this paper designs a fine-grained sentiment analysis application system on hotel review and builds the system with encapsulated internal core function with a friendly user-interface.

Keyword: Sentiment Analysis, Opinion Mining, Quantification of Sentiment Strength, Semantic Role Labeling, Statistical Machine Learning, Natural Language Processing

Written by Shi Hanxiao

Supervised by Qian Peide

and

Zhou Guodong

目　录

第1章　绪　论

作为信息抽取领域中的研究热点，文本情感分析研究从本世纪初开始获得了越来越多的关注，研究成果也日益丰富。本章首先介绍了情感分析的研究背景和意义，接着对情感分析研究相关的各个方面进行了国内外研究现状的阐述，特别针对本文主要研究的细粒度情感分析进行了详细讨论，最后着重叙述了本文的研究重点，给出了本文的组织结构。

1.1 课题背景和意义

随着互联网的飞速发展，特别是 Web2.0 技术的逐渐普及，广大网络用户已经从过去单纯的信息获取者变为网络内容的主要制造者。中国互联网络信息中心发布的《第 31 次中国互联网络发展状况统计报告》[1]的数据显示，截至 2012 年 12 月，我国网络用户总数量已经达到 5.64 亿，网民规模较 2011 年底增长 5090 万人，互联网普及率为 42.1%。如此庞大且快速增长的网络用户群体加上 Web2.0 模式的互联网应用，使网络内容的数量和网络信息的访问量都以前所未有的速度增长，互联网已经成为人们表达观点、获取信息的重要途径。当前互联网上的信息形式多种多样，如新闻、博客文章、产品评论、论坛帖子等等。情感分析就是对这些信息进行有效的分析和挖掘，识别出其情感趋向——高兴、伤悲，或得出其观点是“赞同”还是“反对”，甚至情感随时间的演化规律。

随着电子商务飞速发展，商品评论中的情感倾向性分析逐渐成为当前的研究热点。它的研究目的是利用网络上丰富的顾客评论资源，进行商品的市场反馈分析，为生产商和消费者提供了直观的针对商品各个特性的网络评价报告。目前，一方面情感信息在互联网上呈爆炸式增长，另一方面情感信息对普通消费者，公司组织，和国家政府等各级别的用户都有重要作用，如何帮助用户方便快捷地找到所需的情感信息，成为当前需要迫切解决的问题之一。情感分析任务正是适应这种需求，希望架设一个用户到情感信息的桥梁，使用户能有效获取情感信息。通过对网络上各种信息，特别是主观性文本的倾向性分析可以更好地理解用户的消费习惯，分析热点事件的舆情，

为企业、政府等机构提供重要的决策依据。众所周知，当面对商品评论时，用户更希望了解产品各个方面的情感倾向，这更有利于他们的综合判断和抉择，而传统的情感分析往往是面向篇章和句子的粗粒度分析方法，不能有效解决此类需求，这就需要我们应用细粒度的情感分析方法来实现。

由于情感分析的对象往往具有信息量大、非结构化等特点，所以目前还存在不少问题和难点，如情感词典建设落后，语义特征应用较少，评价对象属性及其情感表达元素的识别效率不高等，这些直接导致现有情感分析的准确率不高，影响了实际使用性能。虽然当前不少网站在提供评论的基础上，还提供了针对评价对象的总体量化评分功能，如 Amazon.com，商品评论进行了五星制评级，5 星为最好，1 星为最差；还有部分网站推出了针对评价对象各个属性（特征）的细粒度评价功能，如 Ctrip.com，它实现了对宾馆的房间卫生、酒店服务、周边环境、设施设备这四方面的 5 分制评级，并给出综合得分。但由于各个网站的评价标准不够统一，再加上原本的细粒度评价结果更多的是建立在人工判定基础之上，所以针对评论的细粒度情感分析是一个重要的研究趋势。我们希望通过本文的相关研究，解决细粒度情感分析中的关键问题，帮助提高分析的准确率，为商业应用推广打下基础。

另外在学术界，近几年来情感分析一直受到研究者的关注，已经成为信息检索和自然语言处理领域的热点研究问题。从近年来在 ACL、WWW、SIGIR、CIKM 等顶级会议上的文章发表情况就可以看出情感分析研究开始吸引越来越多的学者加入，成果也越来越丰富。同时，由于其在企业的商品评价、政府部门的网络舆情监管等方面的应用，已吸引越来越多的 IT 企业参与到该领域的研发之中，如国外的 Google、Autonomy 公司，国内的阿里巴巴、北京拓尔思、北大方正等企业。

总之，研究情感分析具有重要的科学意义和实际应用价值。同时，本文的研究对于推动电子商务的发展，促进政府部门对于互联网的监管等方面也具有十分重要的意义。

1.2 国内外研究现状及分析

情感分析研究的历史并不是太长，2000 年左右开始逐渐成为自然语言处理和文本挖掘领域的热门话题。情感分析，又名意见挖掘，是指通过挖掘和分析文本中表达的情感内容，帮助用户方便快捷的获取所需要的情感信息。本章节首先对情感分析相

关的基础性研究和成熟的情感分析方法进行分析和阐述，如情感词典与情感语料库、文本主观性分类以及篇章、句子级别的情感分析，并总结存在的问题和不足。在随后的介绍中引出细粒度情感分析，并详细阐述其国内外研究现状与趋势，并给出总结。

1.2.1 情感词典与情感语料库

（1）情感词典相关技术。情感词是研究文本情感分析的基础[2]，情感词典研究主要分为情感词获取，情感词极性判定及量化计算，以及情感词的存储管理等工作。无论基于什么粒度的情感分析，都依赖着情感词典这一公共性资源。现阶段的情感词获取工作主要通过现有情感词典、语料库资源进行抽取，或者利用这些资源直接派生出面向具体应用的情感词典，如 Baccianella 等人[3]利用 General Inquirer（GI）词典进行情感词典的构建，Gyamfi 等人[4]则利用 MPQA 语料库建立情感种子词典，并结合 WordNet 实现了情感词典的扩展。而 Devitt 等人[5]和 Esuli 等人[6]都在 SentiWordNet 基础上进行情感极性的识别和观点挖掘。总的来说，以上研究大多通过从已有的情感词典（或其他相关情感语料库资源）中抽取出一定数量的情感种子词，利用语义相似度、点互信息（PMI）等方法进行情感词的获取和极性分类[7]。也有部分研究者开始尝试利用大量标注文档中情感词所处的上下文知识进行机器学习，生成相应的模型，然后对新文本进行情感词的预测和获取。另外，在情感词典的生成方法上，从早期单纯依靠人工建立逐渐向半自动化方法过渡[7-13]，有些系统如果在不考虑过高的精确度要求下甚至可以达到全自动生成[14]。但目前使用的通用情感词典构建方法，大多只考虑了词语间的局部信息而忽略了词语间的全局信息，导致算法准确率不高；或者只在小规模测试集上进行实验，当测试集扩大时，准确率迅速下降（如文献[15,16]中的方法）；或者基准词的数量对实验结果影响很大（如文献[17, 18]中的方法）；或者由于方法自身的很难获得更高的准确率（如文献[19]中的方法）。

相对情感词典构建以及情感词获取、极性分类等工作，情感词的极性强度量化计算方法研究起步相对较晚。极性强度量化计算是对情感词的情感倾向值进行计算，由于情感词的情感倾向值对上层应用中的情感分析、情感统计工作影响越来越大，所以目前也吸引了不少学者进行研究。如 Turney 等人[7]提出了通过计算情感词与正向词“excellent”的 PMI 值减去这个情感词与负向词“poor”的 PMI 值的差值来进行情感词的

极性强度量化计算。Quan 等人[20]分别从篇章、段落和词汇 3 个级别对博客文本进行细粒度标记的规范，并将情感强度分为了 8 个层次。Ku 等人[2]则利用相同汉字往往分布在同一极性的词组之中这一语言学规律，设计出基于字在褒义词典与贬义词典之中出现的概率关系的算法来求解情感词的情感倾向值。

在国内，情感词典的构建已经起步。如董振东等人[21]研制的知网（HowNet）目前已含有情感词典，并对情感词典具体细分为程度级别词语、负面评价词语、负面情感词语、正面评价词语、正面情感词语和主张词语。柳位平等人[22]从 HowNet 发布的情感词语集中提取基础情感词用于情感词典的构建，但其正确度对 HowNet 情感词集的依赖性较强。徐琳宏等人[23]构建了一个情感词汇的本体，并对每个情感词都进行情感类别、强度和极性的描述。

虽然国内外的研究学者们在构建情感字典这一方面投入了很多的精力和心血，但由于工程手段的欠缺以及情感分析的复杂性，导致要使没有感情辨识能力的机器来进行复杂的分析推理是相当有困难的，此外它的作用范围实在很狭窄，要用其处理复杂的情感研究，或者是对某个词进行不同领域的情感辨识就形同虚设，有时甚至可能将文本的情感倾向作其相反的理解。比如说对于一个形容词“长”而言，我们可以说电池的寿命长，很明显这是一个褒义词，且在很多行业领域，例如手机、电脑等数码行业，又或者是电瓶车，蓄电池等机电行业，电池的寿命长就意味产品质量性能的优越，意味着产品市场竞争的优势，相反地，我们又可以说一个程序启动响应时间长，开机的等待时间长，又或者是加油站的排队等待时间长，在此我们又可以明显的感觉出，一个可以带有褒义性质的“长”字不单单只可以是褒义，它也可能是贬义的，当然除了褒贬义之外，“长”又可以是中性的，比如老人们常说的冬至过后，白天的时间开始长了，晚上的时间开始短了。所以有些词或字，只是用现有的基础情感字典来解读，比如说 HowNet、MindNet、Wikipedia 等，它们的确已经不再适用于复杂多变的多领域环境。所以在情感词典的构建中，我们往往需要分成两类进行处理，一类是基础情感词，不管什么领域都不具二义性；另一类是领域情感词，不同领域，情感极性不同。

总的来说，目前的情感词典研究更多集中在情感词的褒贬分类上，而其情感倾向值往往通过人工标注得出，量化计算方法研究相对缺乏。

（2）文本情感语料库建设。文本情感语料库的建设相比情感词典起步较晚。目前更多的情感语料都来源于网上的评论数据，根据情感分析任务分类的不同，设计出

不同的标注方法进行人工标注。由于情感分析任务的特殊性，现有的文本情感语料库基本都是面向领域的。领域语料库有原始语料素材库和标注语料库。影评数据集作为使用较多的原始语料素材[1]，由电影评论组成，广泛应用于词汇和篇章情感倾向研究，但同时其由于未进行细粒度的标注，应用范围有所限制。MPQA 库[24]是由 NRRC Summer Workshop 所开发进行了深度标注的语料库，对论述持有者、对象、极性和强度等进行标注，缺点是规模过小。Quan 等人[20]分别从篇章、段落和词汇 3 个级别对博客文本进行细粒度的标记，建立博客情感语料。将情感分为 8 类，考虑情感主体与对象、情感词和短语、程度词否定词连词、各类修辞和标点符号（如：!），该标注文档最后以 XML 形式输出。语料库的标注程度和精确度直接影响情感倾向分析结果的准确度，所以语料库建设非常关键。

国内也有相关学者进行了情感语料库的研究。汉语情感语料库标注方面的资源则较少，清华大学标注了部分旅游景点描述的情感语料[25]，用来辅助语音合成，但规模也较小。上海交通大学宋鸿彦等人[26]采集了中国汽车网中的评论信息，并利用自行开发的图形化标注工具构建了一个汉语意见型主观性文本标注语料库，但目前标注语料规模比较小。大连理工大学的徐琳宏等人[27]利用小学教材、电影剧本、童话故事、文学期刊等语料构建了一个情感语料库，并且获取了语料库的情感分布，情感迁移规律等统计数据，目前已经标注完成近 4 万句，100 万字的语料。其语料标注的情感以句子为单位，颗粒度较粗。国立台湾大学的古伦维等人[28]开发了一个意见抽取评价语料库，对语料的篇章、句子、词汇级别的情感倾向分别进行了标注，可以区分显式或隐式的意见持有者，但是标注不涉及词法分析与句法分析的信息。

1.2.2 文本主观性分类

在通常的网络文本中，存在大量的客观性文本和主观性文本。客观性文本是一种不带有感情色彩的对个人、事物或事件的一种客观性描述，主观性文本主要描述作者对事物、人物、事件等的个人（或群体、组织等）想法或看法[29]。主观性文本是文本情感分析的主要对象，因此，对大量的网络文本事先进行主客观文本识别非常重要，能够有效地缩小分析范围，减少干扰[12]。就一篇文本来讲，它所表达的情感极性是正

[1] http://www.cs.cornell.edu/people/pabo/movie%2Dreview%2Ddata/

面还是负面往往通过主观性语句体现出来的，如“酒店的服务态度很好!”。但是像“酒店大床房的价格刚好是 150 元整!”这样的客观语句，虽然也有情感词“好”，但在文本情感分类过程中不起任何作用。所以如果能先区分出一篇文本中的主观句和客观句，然后只对主观句进行特征分析和选择，就会很大程度上提高分类的准确率[38]。语言的主观性分析可以用于许多领域，比如：情绪识别[30]，邮件分类[31]，识别说话者角色[32]，文章的风格分析[33]等。

为了在信息抽取任务中有效的识别和抽取事实信息，Riloff 等人[34]设计了多种过滤方法，利用主观性分类器过滤掉篇章中的主观性文本，然后再进行信息抽取，并在 MUC-4 terrorism 数据集上进行了测试，最后得到了多组实验结果，通过对比分析发现，利用分类器首先进行主观性文本选择性过滤能够实现在略微降低召回率指标的情况下，较大地提高实验结果的正确率，整体性能达到最佳。

Toprak 和 Gurevych[35]则在 DEFT’2009 文本挖掘挑战任务中完成了针对英语和法语文档的主观性分类实验，他们主要采用了词语特征、词性特征以及词典信息特征（主要是含有主观表达式的词典，如 SentiWordNet 等），然后利用 SVM 分类器实现基于监督学习模型的文档主观性分类任务。实验结果表明词典信息特征对提高主观性分类任务的召回率有显著帮助，当与词性特征进行组合之后，其性能远好于只选用单个特征的情况。

Remus[36]在研究句子级别的主观性分类中，引入了可读性测量特征，通过对前人研究的总结，选择了六种不同的可读性计算公式，同时通过 Remus 的分析，找出了与可读性相关的三个因素：每个句子中词的平均长度、平均句子长度以及句子中的平均词数，然后设计出了自己的计算公式。通过实验对比，虽然性能上没有比其他方法有太大提高，但 Remus 提出的想法如果能进一步与句法知识进行关联，完善公式的设计，应该能够对性能的提高起到正面的作用。

Wang 等人[37]提出了基于半监督自训练（semi-supervised self-training）的句子主观性分类，在自训练过程中，首先尝试决策树模型作为选择度量方法，但实验效果证明不理想，后来设计了 Value Difference Metric（VDM），并结合 naive Bayes 分类器，在 MPQA 语料库中进行试验，取得了不错的效果。结果证明了 VDM 方法在句子主观分类任务中针对未标注数据选择度量中起到了不错的效果，比基于置信度方法更为理想，另外半监督方法比原来的监督方法要好，这主要因为利用了大量的非标注文档。

Lambov 等人[38]研究了跨领域的文本主观分类，提出了利用多视角学习（multi-view learning）方法提高跨主题的文本主观性分类正确率。通过融合高层次（high-level）特征，如情感词的极性强度和名词的抽象程度等，和低层次（low-level）特征，如 TF-IDF 信息，并利用 SVM 和 LDA 分类器进行联合训练（co-training），在四个不同主题的数据集中进行文本主观性分类实验，平均性能比基准方法提高了 9.2%。

另外 Finn 等人[39]通过研究主客观句分类，得出基于词性标注的特征选择方法比词袋效果好的结论。Yu 等人[12]对新闻这类主要讲“事实”的文本进行主客观句子识别，利用 SimFinder 工具计算句子相似度，构造训练集，结合各类词性信息构建贝叶斯分类器，提出多分类器的构建以解决训练集构造的不确定性和训练集质量的问题。Pang 等人[40]利用属性相同的句子位置分布较近的特点，将候选句子构成一幅图，从而将主客观句分类转化为求图的最小割问题，实现 Cut-based 分类器，对主客观句进行分类识别。在中文方面，姚天昉等人[41]着重从一些特殊的特征角度考察了主客观文本，如标点符号角度、人称代词角度、数字角度等。叶强等[42]在 N-POS 语言模型的基础上，利用 CHI 统计方法提取中文词类组合模式，提出主观文本词类组合模式提取方法，建立中文双词主观情感词类组合模式——2-POS 模型，并在语料试验中取得了良好的效果。

总之，主观性文本识别主要以情感词为主，利用各种文本特征表示方法和分类器进行分类识别。该方法定义明确，其根本问题在于特征的选取。因此，尝试使用更深层、更复杂的分类特征也许是这类方法的突破方向之所在。文本在经过主观性句子提取后，能够减少干扰，提高分类准确性。

1.2.3 篇章、句子级别情感分析

早期的情感分析任务大多是面向篇章，主要完成文档的情感分类。根据文献[9]的介绍，对于文档的情感分类研究可以追溯到 Hearst[43]与 Sack[44]，他们采用基于认知语言学的模型对整个文档的整体情感进行判断。

Huettner 与 Subasic[45]依靠手工构造的情感词典，运用模糊逻辑来对文档进行情感分类，该情感词典对情感词进行了详细的领域区分、词性标注、语义分类以及情感

强度量化。Das 与 Chen[46]同样也使用了手工构造的情感词典，他们所进行的工作是研究人们在股评文档中反应出来的情感与股价走势的关系。对于给定的一篇文档，他们首先识别出其中的情感词汇，然后将文档中所有情感词的极性进行累加，其中褒义词为 1，贬义词为-1，中性词为 0，最后得到整个文档的极性，从而实现对文档的情感分类（乐观/悲观/中立）。

Tong[47]提出了一种可生成情感时间曲线（sentiment timeline）的系统。该系统能够跟踪在线的电影评论，通过相应的计算和处理显示出情感时间曲线图，该图反映了随时间变化的带有褒义意见的评论与带有贬义意见评论的数量对比。在 Tong 的实现中情感分类所依赖的短语均为手工选择，这意味着需要为每个领域设计专门的字典，在一定程度上限制了扩展性。

Dini 与 Mazzini[47]提出将信息抽取技术运用到情感分类任务中。该方法的核心是，对文本中各句子进行组块分析，将句子转化为组块序列，然后对组块序列施加匹配规则得到填充模板（是否为意见句，以及极性、主题等）。最后对填充模板中的情感信息进行汇总，得到最终的文本情感倾向。

Turney[7]提出了一种简单的无监督学习算法将评论文章分为正面(“Thumbs Up”)和负面(“Thumbs Down”)。他首先利用相应的规则识别包含形容词和副词的短语作为情感词。每个情感词的情感倾向值是通过计算这个短语与正面词“excellent”的点互信息（Point Mutual Information，PMI）值减去这个短语与负面词“poor”的 PMI 值的差值

$$SO(phrase) = PMI(phrase,"excellent") - PMI(phrase,"poor")$$

（公式1.1）

如果情感词与正面词“excellent”更相关，则其情感倾向值为正值，表明该词为正向情感词。如果情感词与负面词“poor”更相关，则其情感倾向值为负值，表明该词为负向情感词。其中词与词之间的 PMI 值计算如公式 1.2 所示。

$$PMI(word_1, word_2) = \log_2\left(\frac{p(word_1 \,\&\, word_2)}{p(word_1) * p(word_2)}\right) \qquad \text{（公式1.2）}$$

其中，$p(word_1)$是指词 $word_1$ 在相关文档中出现的概率，$p(word_1\&word_2)$是指 $word_1$ 和 $word_2$ 同现的概率。一篇文章的情感倾向值是指出现在这篇文章中所有情感词的情

感倾向值的平均值。如果平均值为正，说明这篇文章为正向；否则文章为负向。

Dave 等人[49]也利用无监督学习方法对评论文章进行情感分类。他们同样采用一系列规则抽取出所有情感词，并计算每个情感词的情感倾向值。最终一篇评论文章的情感倾向值是通过对所有情感词的情感倾向值求和之后的结果进行判断。

$$class(d_i)=\begin{cases} positive & eval(d_i)>0 \\ negative & eval(d_i)<0 \end{cases}$$

$$eval(d_i)=\sum_i score(f_i) \qquad \text{（公式1.3）}$$

其中 $eval(d_i)$ 是指文章中所有情感词 f_i 的情感倾向值 $score(f_i)$ 之和。

Pang 等人[50]利用有监督的学习方法将电影评论分为正面和负面两类。他们分别使用朴素贝叶斯，最大熵模型和支持向量机作为有监督学习算法的分类器。通过实验他们发现对评论文章的情感分类要比基于事实的文章分类更具挑战性。在基于有监督的学习算法中，每篇文章被转换成一个对应的特征向量进行表示，特征选择的好坏对情感分析任务的性能有直接影响。在 Pang 等人的工作基础上，后续的工作主要把情感分类看做是一个特征优化的任务，研究者们通过挖掘各种不同的特征以期望提高情感分类的性能。Chaovalit 等人[51]同样也采用了机器学习方法对电影评论进行观点挖掘。Mullen 等人[52]基于奥斯古德的语义分类理论，利用 WordNet 对形容词进行效力、活力、评价度的判断，由此生成特征。通过加入语义倾向性的类别特征，表明正确率有所提高。Whitelaw 等人[53]关注了具有评价信息的组合（appraisal groups），主要提取了形容词词组作为特征，并结合标准词袋特征，采用 SVM 分类器对电影评论进行测试，达到了 90.2%的准确度。

篇章级别的情感分类是属于粗粒度的情感分析任务，随着研究的深入，逐渐向句子级别的情感分析过渡。

Yu 等人[12]提出了为自动问答系统抽取意见性句子的方法。首先用贝叶斯分类器判断文本是否含有意见，然后运用无监督统计方法（Yu 等人尝试了相似性法、朴素 Bayes、多重 Bayes 等 3 种方法）从文本中识别出主观性句子。对于识别出来的主观性句子，进行情感分类并判别极性。Yi 等人[54]也提出了句子级别的情感分类方法，不过，由于他们使用的是模式匹配的方法，所以其识别能力受到了相应的限制。

Hu 与 Liu[9]提出了基于产品特性的情感摘要任务。在这一工作中，他们也实现了

句子级别的情感分类方法：首先是通过 WordNet 的同义词与反义词关系，得到情感词及其语义方向；然后根据给定句子中语义方向占优势的情感词类，判断句子的极性，从而实现对句子的情感分类。不过，他们在文献[9]中指出该方法对于长句子效果不甚理想。

Fu 和 Wang[55]提出了利用模糊集理论实现中文句子的情感分类。他们定义了三个模糊集来表示三种情感极性类（正面、负面、中立），通过考虑多种情感粒度信息，如情感语素、情感词、情感短语等，设计了由细到粗（fine-to-coarse）的句子情感强度计算方法，最后基于情感强度设计出相应的隶属函数（membership function）计算给定情感句子相对不同模糊集的隶属度，并利用最大隶属（maximum membership）原则判定该句子的最终情感极性。

Johansson 和 Moschitti[56]首先分析了早期意见表达识别及情感极性分类作为两个单独任务的研究情况，然后通过扩展全局结构方法（global structure approach），并引入多种反应句子的极性结构特征，实现意见表达识别与情感极性分类的联合模型（joint models）。在相同的 MPQA 语料库环境下，与 Choi and Cardie[57]的方法进行实验对比，结果表明在性能上，Johansson 和 Moschitti 的方法有较大的优势。

另外国内研究者也进行了一定的研究。徐琳宏等人[58]提出一种结合语义特征和机器学习的汉语文本极性自动识别机制。首先通过 HowNet 计算词汇倾向性，选择极性明显的词汇作为特征值，用 SVM 分类器分析文本的褒贬。为了提高分类准确度，考虑否定副词和程度副词对语义倾向的影响。徐军等人[59]利用朴素贝叶斯和最大熵方法研究新闻及评论语料的情感分类，通过一系列的实验得出各种方法的优劣对比。唐慧丰等人[60]针对分类技术中的关键技术如：特征表示、选择和文本分类方法进行对比实验，采用 BiGrams 特征表示方法、信息增益特征选择方法和 SVM 分类方法，在足够大的训练集和选择适当数量特征的情况下，情感分类能够取得较好效果。张伟等人[61]在分类器中使用了基本核与树核组成的复合核进行句子级别的情感分类，实验结果显示由基本核与树核组成的复合核比线性核具有更加的性能，这是因为复合核的树核函数通过对句法树计算为情感分类提供结构化信息，而复合核中的基本核则可以包含一些无法通过树核函数捕获的信息，这样在复合核中可以引入一些对于情感分类很重要的信息，进而提高情感分类的准确率。

1.2.4 存在问题和不足

从国内外研究现状可以看出，文本情感分析是一项复杂的技术，而不单纯是一个文本挖掘或者文本分类的工作。虽然在以上各方面的研究都取得了长足的进步，但还存在不同的问题：

- **情感词典建设**

目前面向研究的情感词典更多是依赖人工建设，随着互联网的发展，新增词汇的不断涌现，如何设计有效的方法实现自动化情感词典建设和升级是目前有待解决的问题之一，另外目前的情感词典更多还只停留在情感词的褒贬分类，其情感倾向值往往通过人工标注得出，量化计算方法研究相对较少，这对情感分析的量化工作带来了一定的阻碍。

- **文本主观性分类**

文本主观性分类是情感分析的第一步工作。目前更多还停留在利用传统方法进行常规特征的提取，忽略主观性文本语言的多变性、省略性、情感隐蔽性等特点。所以如何挖掘和利用更深层次、更复杂的语义信息将是文本主观性分类任务有待突破的方向。目前针对这方面的研究还相对比较少，尚未形成突破性的解决方案。

- **篇章、句子级别情感分析**

面向篇章和句子的情感分析相对比较成熟，但这方面的研究工作在面临复杂的语言环境和特殊的句子类型时，处理效果还不是很理想。目前已有研究者针对特殊的句型进行情感分析研究，如 Narayanan 等人[84]专门研究了条件句的情感分析。一种最常见的解决方法就是使用一定的模式进行情感判定。但如何自动地选择模式，以及如何评估选中模式的有效性，目前还在进一步的研究中。

总之，以上三个方面的研究在解决粗粒度的情感分析任务时，已经较为成熟。但当需要针对评价对象属性进行分别情感计算的需求时，粗粒度的情感分析方法往往不能胜任，这需要引入细粒度情感分析的研究。

1.2.5 细粒度情感分析现状与研究趋势

如果在单位篇章或者句子中出现评价对象的多个属性（在有的文献中又叫主题或

特征，本文在接下去的叙述中统一称作为属性），前面基于篇章、句子级别的情感分析方法往往不能满足要求，也就是需要更细粒度的情感分析方法支持，以识别出属性词、情感词以及情感修饰词等，并通过相关技术定位他们之间的关系和计算出相应的情感值。所以这几年的情感分析研究更多趋向于细粒度的情感分析。下面我们对细粒度情感分析的国内外研究现状进行介绍和分析。

属性词和情感词等的识别在细粒度情感分析中具有重要作用。利用属性词和情感词的抽取，可以构建领域相关的属性词表和情感词表[62–64]。此外，如果抽取出属性词和情感词，并正确识别属性词和情感词的对应关系。那么可以生成基于属性词和情感词的可视化产品评论摘要[65]。Hu 和 Liu[9]提出抽取属性词和情感词进行评论摘要。他们认为常见（frequent）属性词往往是评论者经常提及的名词或名词短语，因此可以用出现频率来提取。他们使用关联规则的方法抽取最小支持率（minimum support）为 1%的名词或名词短语作为常见属性词。其中最小支持率的阈值是由实验确定的，是指名词或名词短语出现的句子数占总句子数的比例。为了保证属性词抽取的准确度，他们还利用属性词出现的位置和形式来确认其正确性。对于情感词的抽取，他们只认为形容词是候选情感词。当抽取出属性词之后，他们首先定位所有包含属性词的句子，然后在这些句子中抽取修饰属性词的形容词作为情感词。Hu 和 Liu 还根据抽取好的常见属性词和情感词去抽取不常见（infrequent）的属性词，因为他们认为情感词可以修饰常见属性词，也可以用同样的方式修饰不常见属性词。

Popescu 和 Etzioni[66]对 Hu 和 Liu 的属性词和情感词抽取算法做了进一步改进。对于属性词抽取，他们希望过滤掉不是属性词的名词短语。具体改进做法为：首先为每个产品类别，例如扫描仪（scanner），定义了一系列整体关系标识词（meronymy discriminator），例如对于扫描仪类别，标识词包含“of scanner”，“scanner has”，“scanner comes with”等等；然后通过计算名词短语与标识词的 PMI 值来计算名词短语是属性词的可能性，也就是利用名词短语和标识词以及它们的同现在搜索引擎查询返回的次数来计算。这种方法认为一般属性词是这个类别下的组成成分或功能部件。如果名词短语和类别的整体关系标识词同现的次数比较低，则说明名词短语可能不是这个类别下的属性词。Popescu 和 Etzioni 也是利用抽取的属性词来辅助抽取情感词的，不过他们在句法树上定义了十种属性词、情感词句法关系，然后利用这些句法关系和属性词来抽取情感词的。他们采用了 Relaxation Labeling 的方法去识别情感词的情感倾向。这种方法不仅可

以利用情感词的局部依赖关系，而且还可以考虑整个数据集的统计信息。

另外，Kamal 等人[67]通过对评论文章进行语言学和语义分析（主要进行词性标注分析），设计出了相关规则以实现对评论文章中产品“属性-情感对”（“feature-opinion pairs”）的抽取，并生成一个三元组<f, m, o>，其中 f 为对象属性，往往表现为名词短语，o 为对属性 f 的意见表达，往往表现为形容词，m 为相对于意见表达 o 的程度修饰词。同时为了判断抽取生成的“属性-情感对”结果与相关评论文章的可信度，作者应用 HITS 算法构建相应的二部图，其中“属性-情感对”看作 hubs，评论文章看作 authorities，并进行可信度分数的计算。

Fang 和 Huang[68]则提出了主要利用潜在结构模型（latent structural models）实现细粒度的属性分析（aspect analysis）任务。他们首先利用结构学习模型完成了文档级别和句子级别的情感极性分类任务，同时还能够识别出与特定属性相关的句子以帮助实现属性级别（aspect-level）的情感分析。最后通过实验证明了该方法的可行性和优越性。

Filho 等人[69]利用 Ait-Mokhtar 和 Chanod[70]开发的 XIP 分析器对文本进行依存关系分析，通过对其生成的结果进行处理获得相应的语义关系，并最终实现对评论文本的属性级别（feature-based）意见挖掘，同时还利用 SVM 分类器完成了评论文本的情感极性分类。Filho 等人针对最终的分析结果还开发了 XOpin 终端系统，提供友好的图形化界面以方便用户浏览和查看评论及其分析结果的所有信息。

Liu 等人[71]还提出利用基于词的翻译模型（word-based translation model，WTM）[72]实现细粒度的意见挖掘，他们首先分析了两种传统方法，一种是领近方法（adjacent methods），意见对象（opinion targets）的识别仅考虑围绕着意见词在给定窗口范围内进行查找，然而该方法由于窗口大小的限制，特别针对长间隔的意见对象与意见词的抽取，正确率往往不高。第二种方法是基于句法分析方法（syntax-based methods），该方法主要依靠句法分析的正确性，由于在线评论往往是非正式文本，所以句法分析的出错率较高，这直接导致情感分析的正确性。而 WTM 是局限在一个窗口范围内进行意见关系（opinion relations）的识别，同时也没有用到句法分析工具，所以它能够有效识别出长间隔的意见关系，以及避免句法分析错误对后面分析工作的影响。最后通过实验证明，该方法在挖掘意见对象和意见词的关联上比传统方法效果要好。

Miao 等人[73]则对每个产品的评论进行基于四元组（title，help，date，R-content）

的抽取，其中 title 是指评论的题目，help 是指认为该评论是有助于他们的顾客数量，date 是指评论发表的日期，R-content 是指顾客评论的一系列句子集，对于每个句子，采用 Hu[9]提出的方法，即将每个句子表示为一个三元组[属性（feature），倾向（sentiment），句子相关内容（S-content）]。通过对评论信息的结构化抽取建起索引，结合时间信息、认为评论有用的人数和词频信息进行排名权重计算，得出褒贬评论的统计曲线，分别进行褒贬评论对比展示和基于时间序列的图形化展示。该方法实现了评论文本的细粒度情感分析。而 Qiu 等人[63,74]在前几年提出一种半监督学习算法 Double Propagation 来同时抽取属性词和情感词。他们除了定义一系列种子情感词之外，还利用句法树规则定义了属性词和情感词，属性词和属性词，情感词和情感词之间的关系。

另外，Jin 等人[75]还研究了有监督学习的算法对属性词和情感词抽取的工作。他们将属性词和情感词的抽取看作是一个序列标注任务。评论中的每个词都对应一个标签类别。他们提出使用词汇化的隐马尔科夫模型（Lexicalized HMM）来求最有可能的标签序列。Wilson 等人[76,77]也曾研究过基于有监督分类器的方法来识别情感词。实验表明融合多种特征的有监督学习分类器，能大大提高属性词和情感词的抽取性能。以上这些工作说明了有监督学习对属性词和情感词抽取的有效性，但这些方法还存在一定的缺陷。基于隐马尔科夫模型序列标注算法考虑了句子间的顺序关系，但它是一种产生式模型，不能很好地融合各种特征。基于分类的方法，是独立的识别主题词和情感词，忽略了它们之间的类别关系。此外，在如何综合句子的语言学结构，同时利用多种特征仍没有有效的模型。

此外，已有相关研究人员通过引入更多的自然语言处理技术，挖掘更多的语义特征，提高分析效果。Kim 等人[78]利用 SRL 技术实现了从在线新闻文本中提取观点、观点持有者以及主题。Kobayashi 等人[79]利用指代消解技术实现了对评论数据（日文）中评论对象的<属性,值>抽取，首先利用词典找出所有的属性词和情感词，然后针对每个情感词从候选属性词中利用指代消解技术找出相应的属性词，从而确定<属性,值>对。Stoyanov 等人[80][81]提出了利用指代消解技术帮助识别出评论数据中的观点持有者信息，形成相应的指代链，从而提高观点总结的正确性。

国内学者对细粒度情感分析也有一定的研究。刘永丹等人[82]则采用格语法作为语义分析的基础，对文本进行语法和语义分析，提取相应格，然后与事先建立好的基于

语义的过滤模板（该模板对行为受体和行为主体进行考虑）进行匹配，通过匹配距离函数和匹配相关函数计算匹配模板相关度，累加匹配模板相关度，最后与阈值比较得到文本过滤结果。姚天昉等人[83]利用领域本体抽取语句主题以及它的属性，在句法分析的基础上，识别主题和情感描述项间的关系，最终决定语句中每个主题的极性。采用基于经验的语言模式方法，提出一种改进后的 SBV（VOB）极性传递算法，考虑到 SBV 结构（主谓结构）、VOB（动宾结构）和 ATT（定中结构）极性对相关主题词的较准确有效的传递，结合情感词库的建立，实现了合理有效的倾向分析。但该算法没有考虑语气问题。

通过以上介绍与分析，研究者主要实现的是对评价对象属性和情感词进行识别，在情感分析过程中更多只进行情感极性的判定，很少有针对情感极性强度的量化进行研究。为了更好的实现细粒度情感分析任务，我们将在三个方面进行深入研究，分别改进了情感词的极性强度量化计算方法，提出了评价对象属性及其情感表达元素的联合识别模型，实现了细粒度属性分类和情感计算任务。

1.3 本文的研究重点与工作内容

本文以产品评论为对象，主要研究细粒度情感分析中的情感词极性强度量化计算方法、评价对象属性及其情感表达元素的联合识别、细粒度属性分类及其情感计算，最后通过酒店评论作为实例进行应用研究。

（1）情感词极性强度量化计算方法。如果情感词典中只含有褒贬情感分类，缺少情感词的极性强度量化信息，那么在具体应用中，往往有所欠缺。在产品评论中，如果评论都是强烈好评的话，往往是不二之选，相反，如果仅仅一般好评的话，则买家还需要继续货比三家。所以在对评论做情感分析时，识别出评论的情感极性固然重要，如果还能确定情感的极性强度量化值，那对后面的情感计算具有非常大的意义。本文对情感词极性强度量化进行了深入研究，分析并改进了现有的方法，最终实现了基于情感词分类计算的极性强度量化。

（2）评价对象属性及其情感表达元素的联合识别。篇章或句子级别情感分析相对而言是粗粒度的，因为一篇文章或一句话往往不只包含一种情感，评价对象经常被细分为多种主题/属性，这就需要更细粒度的识别才能有效实现情感分析。在细粒度

情感分析任务中，如何正确识别出文本中的评价对象属性及其情感表达元素具有十分重要的意义。本文结合条件随机场理论，充分利用评价对象属性及其情感表达元素之间的类别关系，提出了序列化联合抽取模型；另外还分析了基本特征和语义特征的相关知识及抽取方法，特别针对语义特征的抽取进行了技术分析和算法设计；最后通过调整特征以及特征模板，获到了最佳的实验结果。

（3）细粒度属性分类及其情感计算。与情感词相类似，评价对象属性的描述也是多种多样，同一类对象属性，可以有多种语言表达，如“外观”，相似的描述可以有“外形”、“外表”等。虽然这些词语不相同，但描述的含义、概念是基本相同的。情感量化计算工作之前，评价对象必须确定好属性类别，以方便情感汇总统计。本文通过有监督学习方法和半监督学习方法对属性分类问题进行了深入研究，最后针对评论中可能存在情感词缺少对象属性的情况，通过计算 PMI 值的方法来确定评价对象属性类与情感词之间的关联概率，实现对缺失对象属性的情感信息进行合理属性类的指派，使情感汇总计算更为合理有效。

（4）以酒店评论作为实例进行应用研究。在（1）—（3）的研究内容基础上，我们进行了集成应用研究。本文首先设计相关的爬虫软件实现对目标评论网站数据的进行收集，然后利用相关技术实现对网络评论的预处理及格式化的数据保存，随后采用上述情感分析方法，完成评论数据的对象属性和情感信息联合识别，并利用已有的情感词典及极性强度量化结果，完成以酒店为单位的情感汇总计算，最后给出友好的可视化浏览及查询界面。系统还提供了根据不同区域以及用户关心的酒店属性类别进行排名推荐功能。同时为了方便外部应用的调用，本应用还提供了相应接口帮助实现在线评论的实时处理需求。

1.4 本文的组织结构

本文主要围绕细粒度情感分析的关键技术研究展开，总共分为六章，各章的具体内容安排如下：

第一章 绪论：首先简单介绍了本文工作的研究背景和意义；重点讨论了情感分析的相关国内外研究现状，指出了存在问题与不足；详细分析了情感分析的研究趋势以及细粒度情感分析的特点，从而引出本文的研究重点和工作内容。

第二章 情感词的极性强度量化计算研究：首先介绍了情感词和情感修饰词的情感极性强度模糊性特点，阐述了情感词极性强度量化的重要性；然后通过介绍现有情感词的极性强度量化方法，并进行了问题分析；最后提出了基于情感词分类计算的极性强度量化方法，针对不同的情感词类型进行不同的计算方法设计，从而提高了情感词的极性强度量化计算正确率。

第三章 评价对象属性及其情感表达元素的联合识别研究：首先介绍了条件随机场模型，对其原理及应用进行了阐述和分析；随后提出了评价对象属性及其情感表达元素的序列化联合抽取模型，详细分析了基本特征和语义特征的相关知识及抽取方法，特别针对语义特征的抽取进行了技术分析和算法设计；最后，利用机器学习方法实现了评价对象属性及其情感表达元素的联合识别，并进行了实验对比和性能分析。

第四章 细粒度属性分类及情感计算：首先介绍了基于监督学习的属性分类，通过特征设计和数据集的构建，实验证明了最理想的特征组合；然后引入半监督学习理论，通过设计和构建多个实验探讨了自举学习的各种情况，以及自举迭代的终止条件；最后介绍了情感计算的方法，实现对商品评论的细粒度情感分析。

第五章 基于细粒度情感倾向性方法的酒店评论意见挖掘系统：首先首先介绍了基于细粒度情感倾向性方法的酒店评论意见挖掘系统架构及功能模块，详细分析了各模块的集成应用特点；然后详细介绍了系统的主要功能界面及使用说明；最后总结了从系统内部集成到用户界面展示这一过程中的问题和特色。

第六章 总结与展望：主要进行了研究工作的总结，分析了整个研究过程中的贡献以及面临的问题；最后对下一步工作进行了设想，提出了一些当前还没有很好解决的挑战性工作任务。

第2章　情感词的极性强度量化计算研究

2.1 引言

通常，文本情感分析时，无论是句子粒度，还是篇章粒度，都强依赖于情感词典。所以情感词典的好坏直接影响情感分析的正确性。但如果情感词典中只含有褒贬情感分类，缺少情感词的极性强度量化信息，那么在具体应用中，往往有所欠缺。在产品评论中，如果评论都是强烈好评的话，往往是不二之选，相反，如果仅仅一般好评的话，则买家还需要继续货比三家。比如用户在选择酒店的时候，看到有评论说“这个酒店非常不错”，往往就会对该酒店留有好印象，至少能把它当作候选。但如果有评论说“这个酒店整体一般”，用户一般不会再对他有兴趣，除非该酒店还有某一方面特别出众，刚好能够迎合该用户的需求。所以在对评论做情感分析时，识别出评论的情感极性固然重要，如果还能确定情感的极性强度量化值，那对后面基于评价对象属性的细粒度情感计算具有非常大的意义。所以本章我们对情感词极性强度量化计算方法进行了深入研究，对现有方法进行优化和完善，设计出更加合理全面的极性强度量化计算方法。

很多研究表明，情感词的极性容易确定，而其极性强度很难量化，主要原因在于情感极性强度模糊性特点。本章首先介绍了情感词和情感修饰词的情感极性强度模糊性特点，分析了模糊性产生的具体缘由，然后研究了早期基于字的情感词极性强度量化计算方法，通过分析和讨论，改进了基于情感词分类计算的极性强度量化方法，针对不同的情感词类型设计了不同的计算方法，并通过实验对比验证了方法的性能和效果。最后针对情感未定词进行了基于特定领域的情感极性判定及极性强度量化计算实验。

2.2 情感极性强度模糊性分析

在情感分析过程中，情感极性强度往往通过分值或者级别来表示，但由于情感极性强度判断具有主观性、普遍性和特殊性，所以在具体确定情感极性强度过程中存在不可避免的模糊性。

2.2.1 情感词的情感极性强度模糊性

情感词的极性强度表现出模糊性。在描述衣服颜色的褒义情感词中，虽然都是正面的描述词语，显然它们表达的情感强度是不同的，在语言理解中，“漂亮”比“好看”所表达的肯定态度要强一些，但增强的程度是一个模糊值。Andreevskaia 等人[85]对 GI-H4 和 HM 两个情感词词表进行了一致性研究，两个词表的一致性仅达到 78.7%。Wilson 等人[86]在 MPQA 语料库中，将情感词标注为 4 个级别：中立、低、中、高。在标注过程中，标注者之间基于不同的理解很难在标注结果上达成一致，在整个语料库中达成一致的标注仅占 61%。以上研究可见情感词的极性强度表现出模糊性。

在情感分析中，情感词的情感极性是确定的，而情感词的情感极性强度是连续变化的，本身具有模糊性。在实际情感分析应用中，特别是细粒度的情感分析，为了便于计算和统计，需要将强度进行量化计算，确定情感词的极性量化程度成为典型的模糊性问题。

细粒度情感分析中，情感词典为情感分析提供支持，在不考虑情感强度的情况下，可以进行文本的情感分析工作，但这丢失了情感词自身表达的情感强度差异，忽略了语言的本质。在考虑情感强度的应用中，情感词强度关于情感极性强度量化的模糊性则成为了普遍问题。关于情感词的极性强度量化计算方法我们将在后面进行详细介绍。

2.2.2 情感修饰词的情感极性强度模糊性

通过对评论语料的分析，情感修饰词往往属于程度副词。程度副词能表示词语或者句子的程度，在情感分析应用中可以对情感词的强度进行强化或者弱化的限制作用。研究者尝试根据其强化或者弱化的程度进行分类，但各家的分类体系不统一、副词的类属不统一，根本原因是其本身的模糊性。

形容词是情感词的主体，而程度副词可以体现形容词的程度量级，程度副词是体现情感强度的重要指标。虽然形容词本身可以表示强度，但大范围内的形容词表示的强度相差不多，更多地通过程度副词表示出来。

程度副词是限制性副词的一种，对形容词的程度量级起了限制作用。这种限制可以分为两类：强化限制和弱化限制。强化限制对被限制词的程度起了强化和增强作用，

比如“非常”、“十分”、“很”等。这类词加在被限制词上，使被限制词的程度量级得到了加强。弱化限制对被限制词的程度起了弱化作用，比如“比较”、“稍微”、“有点”等。这类词加在被限制词上，减弱了限制词的程度量级。

“非常漂亮”和“稍微漂亮”之间的“漂亮”程度显然是不同的，“非常”强化了“漂亮”的程度，而“稍微”则弱化了“漂亮”的程度。在情感分析研究中，情感词“漂亮”表达了正面的意见，在情感极性强度上，“非常漂亮”要强于“漂亮”，“漂亮”强于“稍微漂亮”。

可见，对程度副词的强化作用和弱化作用进行区分，有助于情感分析研究中情感极性强度的计算。但更为复杂的问题随之出现，在强化作用的程度副词之间，强化程度也是不同的。“非常”、“十分”、“很”之间的强化作用并不相同。对程度副词的强化或者弱化程度进行量化是解决问题的简单办法。由于程度副词相对有限，我们通过现有资源选取了 73 个常见的程度副词，进行作用系数人工确定，取值范围为 0.5 至 1.5，步长为 0.2。

通过上面的分析，情感词的极性强度具有模糊性特点，如何进行有效的极性强度量化是当前面临的难题，后面我们将重点探讨情感词的极性强度量化方法。

2.3 基于字的情感词极性强度量化计算方法

词是由字组合而成的，字是词的最小组成单元。国内有许多专家撰文讨论了字与词的关系问题，如韩琳等人[87]从语言学、概念以及文字要素等多方面分析了字与词之间的两者密切关系，魏慧萍[88]也从多个角度分析了汉语字词之间的关系。国外学者 Ku 等人[2]在 AAAI’2006 人工智能顶级国际会议上发表论文阐述了利用汉字的情感统计来计算词的情感方法。从这些文献中，我们可以肯定绝大多数的字与词之间存在着密切的关联，特别在情感词问题上，往往字与词极性分布相似性概率非常大。

我们设计的情感极性强度量化方法主要是以 Ku 等人的方法为基础，下面首先介绍一下他们的算法思路：先利用已有的情感词典通过字频统计的方法，计算出每个字的情感倾向值；然后利用字的情感倾向值设计相应的公式进行词的情感倾向值计算。详细步骤如下。

首先统计每个字的在情感词典中作为褒义词和贬义词的权重，如公式（2.1，2.2）

所示。

$$P_{ci} = \frac{fp_{ci} / \sum_{j=1}^{n} fp_{cj}}{fp_{ci} / \sum_{j=1}^{n} fp_{cj} + fn_{ci} / \sum_{j=1}^{m} fn_{cj}} \quad \text{（公式2.1）}$$

$$N_{ci} = \frac{fn_{ci} / \sum_{j=1}^{m} fn_{cj}}{fp_{ci} / \sum_{j=1}^{n} fp_{cj} + fn_{ci} / \sum_{j=1}^{m} fn_{cj}} \quad \text{（公式2.2）}$$

其中，P_{ci} 为字 *ci* 作为褒义词的权重，N_{ci} 为字 *ci* 作为贬义词的权重。fp_{ci} 为字 *ci* 出现在褒义词表中的频率，fn_{ci} 为字 *ci* 出现在贬义词表中的频率。利用公式 2.1 和公式 2.2 可以计算出每个字作为褒义词和贬义词的权重。*n* 为褒义词表中出现的所有字的个数，*m* 为贬义词表中出现的所有字的个数。为了平衡情感词典中褒义词与贬义词之间的词数差异性，公式 2.1 和 2.2 对每个字在褒贬词表中出现的频率进行了归一化处理。

最后可以利用公式 2.3 计算出字 *ci* 的情感倾向值 S_{ci}。

$$S_{ci} = (P_{ci} - N_{ci}) \quad \text{（公式2.3）}$$

如果 S_{ci} 的值为正数，*ci* 是褒义字，负数则是贬义字，接近于 0 的话，说明 *ci* 趋向于是中性。

在 Ku 等人的方法里，当计算新词 *w* 的情感倾向值时，如果该词由字 $c_1,c_2,\ldots c_p$ 组成的话，只要计算每个字的平均情感值，如公式 2.4 所示，其中 *p* 为词 *w* 中字的个数。如果字 *cj* 没有情感值，则 S_{cj} 取值为 0。

$$S_w = \frac{\sum_{j=1}^{p} S_{cj}}{p} \quad \text{（公式2.4）}$$

如果 S_w 的值为正，说明词 *w* 的情感为褒义，S_w 的值为负则词 *w* 为贬义词，S_w 的值接近 0 的话，说明词 *w* 为中性词，或者叫非情感词。

从 Ku 等人的算法中我们可以看出该方法是一种字袋式的情感值量化统计方法，同等看待每个字的情感值。众所周知，我们用以上方法人工分析情感词时，每个字的

情感权重肯定会有所不同。所以在下面的章节中，我们将提出相应的改进方法，更加合理的进行情感词的极性强度量化计算。

2.4 基于情感词分类计算的极性强度量化方法

根据我们的研究分析发现，Ku 等人的方法在计算字的情感值时，没有对字的特点以及组词特性进行区分考虑。这样容易导致不少字的情感值计算存在问题，根据公式 2.1-2.4 我们不难看出，当如果这个字所在的词存在几乎相同数量的以“不”开头的相反极性的情感词，就会导致该字的情感值趋向为 0。例如，褒义词集里有“好”、“美好”等词，而贬义词集中存在“不好”、“不美好”等词，那样如果利用公式 2.1-2.3 进行情感极性强度量化计算，“好”字的情感倾向值将为 0，因为“好”字在褒义词集和贬义词集中出现的概率完全相同。另外还有一些词含有程度副词，如果简单的利用 Ku 等人的方法，则容易导致极性强度量化不合理的情况，如“很”字，假设在褒义词集和贬义词集中出现的概率差不多，按照 Ku 等人的方法进行计算，“很”字的情感倾向值为 0，如果简单的利用公式 2.4 进行词的计算，那么“很漂亮”的情感值反而低于“漂亮”。

所以情感词的极性强度量化工作完全可以通过分类计算达到更好的效果。我们对情感词分为两大类处理，第一类是基础情感词，也就是首字不含有否定词、程度修饰词且字数不超过 2 个字的情感词；第二类是复合情感词，其首字含有否定词或程度修饰词的情感词或包含 2 个字以上的情感词。在基础情感词的极性强度量化计算工作中，我们首先计算出字的情感值，然后设计相关规则计算出词的情感值；在复合情感词的计算工作中，学习相应的语言学知识，设计相应的规则方法，利用词与词的组合关系进行复合计算。

2.4.1 基础情感词的极性强度量化计算

（1）字的情感值计算

基础情感词的极性强度量化计算首先计算出每个字的情感值。根据我们的统计基础褒贬情感词数为 6473 个，复合褒贬情感词数为 1208 个。为了保证情感词典利用的最大效率，我们在计算字的情感值时还考虑了首字含有“不”等否定词的情感词，把

去除否定词之后字数不超过 2 的情感词去重之后放入到相反极性进行统计。

同样利用公式 2.1-2.3，我们可以算出每个字的情感倾向值。程序的主要算法逻辑思想如图 2.1 所示。

Input: Sentiment lexicon (positive, negative) //输入褒义词典和贬义词典

Output: S_{ci} //输出每个字c_i的情感倾向值

Procedure:

Preprocessing //去除否定词之后字数不超过2的情感词去重之后保存到相反极性的临时词集

Generate characters list $\{c_1,c_2,\ldots c_n\}$ //利用褒贬情感词典中所有基础词表生成相应的字表$\{c_1,c_2,\ldots c_n\}$

Calculate S_{ci} //根据公式2.1-2.3计算字表$\{c_1,c_2,\ldots c_n\}$中每个字的情感倾向值

图2.1 字情感值计算的程序逻辑思想

（2）基础词的情感值计算

分析基础情感词的构造特点，我们不难发现其情感倾向值往往约等于所有字情感倾向值中的最大值。例如，“漂亮”中“漂”的情感值是 0.5，“亮”的情感值是 0.8，我们可以认为“漂亮”的情感倾向值值也等于 0.8，而不应该简单的利用 Ku 等人的平均求值法来计算“漂亮”的情感值。所以在计算基础词的情感倾向值时，我们主要设计了公式 2.5 来实现。

$$S_w = sign(S_{ci}) * Max(abs(S_{ci})) \qquad \text{（公式2.5）}$$

其中 $Max(abs(S_{ci}))$ 为所有字中最大的情感值的绝对值， $sign(S_{ci})$ 则为该字的符号值，如果该字的情感值大于 0，则为+1，如果小于 0，则为-1。

2.4.2 复合情感词的极性强度量化计算

复合情感词的极性强度量化计算相对较为复杂，由于其往往由基础情感词、否定词、程度修饰词等多类词汇组合而成。本项目设计了一种基于组词分类模型的方法解决复合情感词的极性强度量化计算。

针对不同复合情感词的组合特点，我们主要分成了5类：

（1） 基础情感词的叠词，如漂漂亮亮，高高兴兴。这类词我们可以通过寻找词根的方法找到基础情感词的情感倾向值，由于叠词一般来说对原词的情感值影响不大，为了简化问题，我们直接取基础词的情感值。

（2） 基础情感词+基础情感词，如小心谨慎。这类组合词的计算我们采用求平均值的方法实现。

（3） 否定词+基础情感词，如不漂亮。这类组合词的计算可通过对基础情感词的情感倾向值取反运算来实现。

（4） 程度修饰词+基础情感词，如很漂亮。这类组合词的计算可以先获得基础词的情感值，然后根据事先定义的不同程度修饰词的作用强度获得相应的作用系数（取值范围为0.5、0.7、0.9、1.1、1.3、1.5），如“很”的程度系数我们定义为1.3，“比较”的程度系数为0.7。该组合词的最终情感倾向值为两者的乘积，如果该值超出了词的情感倾向值范围[-1,+1]，我们就取最大极值。

（5） 否定词+程度修饰词+基础情感词/程度修饰词+否定词+基础情感词，如不太漂亮/太不漂亮。这类组合词的计算相对比较复杂，前两者之间的位置关系直接影响情感词的倾向值计算。我们利用语言学知识，设计了公式2.6来实现该类复合词的情感倾向值求解。

$$S_w = \begin{cases} (\eta - D_{wi}) * S_{wi} & \text{程度级别词位于否定词与情感词之间} \\ sign(S_{wi}) * (abs(S_{wi}) * D_{wi}) & \text{否定词位于程度级别词与情感词之间} \end{cases}$$

（公式2.6）

其中 S_{wi} 为基础词的情感值，D_{wi} 为程度词的作用系数（范围为 0.5、0.7、0.9、1.1、1.3、1.5），η 为程度词反作用系数，也就是作用系数 D_{wi} 的范围极值之和，所以 η 取值为 2。$sign(S_{wi})$ 为词 S_{wi} 情感值的符号值，如果该词的情感值大于 0，则为+1，如果小于 0，则为-1。$abs(S_{wi})$ 为词 S_{wi} 情感值的绝对值。以“不太漂亮”为例，“漂亮”的情感倾向值为 0.8，“太”的程度作用系数为 1.3，通过公式 2.6 计算“不太漂亮”的情感倾向值为 0.56。而如果计算“太不漂亮”的情感倾向值，同样利用公式，我们可以得到-1。不难看出，这样的结果跟我们主观判断基本一致。

2.5 实验结果及分析

我们首先针对已有的情感词典资源进行整合，如知网的情感分析用语词集[2]、哈工大的同义词林[3]、台湾国立大学的意见词典[4]，通过汇总、去重、筛选、判定等工作，最后选定了 7929 个情感词，具体可以参考网络资源[5]。同时还利用这些资源选定了 73 个常见的程度副词和 23 个否定词。

为了对情感词的极性强度量化计算有更正确的结论，首先安排了三个研究生负责对现有的 7929 个情感词统一进行情感值标注，范围从-1 到+1，步长为 0.1。根据情感词的特点，我们要求标注者只对具有明确褒贬极性的词汇进行情感值打分，如“好”、“差”、“漂亮”等词，因为这些词一般极性都不具二类性。如果具有领域特点的情感词，如“缩短”、“减少”、“长”、“短”等词，虽然安排进入相应的褒贬词集，但统一标注为 0，因为这些词往往要跟具体应用领域对应，我们单从词的概念中很难区分出它的情感极性，如“减少消耗”中的“减少”是褒义，“工资减少了”中的“减少”则是贬义。如果三人不存在情感词极性判定不一致的情况，最后的情感值以三位标注者的平均值为最终标准值，如果存在某人极性判定与另外两位不一致，这由三人讨论后进行修改确定最终的情感值。最终我们对 7929 个情感词划分出了情感明确的词有 7681 个，其中 5432 个贬义词，2249 个褒义词，248 个领域相关情感词，我们暂且定义为情感未知词。另外我们允许极性强度量化计算结果在保证极性判定正确的前提下存在一定误差，因为从人工标注情感值的过程我们可以知道这是一项主观性非常强的任务，也就是不能保证每人都会给出相同的打分，所以后面的实验结果中，我们引入误差系数的概念，只要保证在极性判定正确的前提下并属于误差系数范围之内，我们都认为是正确答案，这也符合我们前面介绍的情感极性程度量化模糊性特点。但同时，如果误差系数过大则将降低极性强度量化计算研究的意义，而过小则与情感值标注的主观性特征相违背。所以通过经验判定和实验数据分析，本文实验中的最大误差系数我们设置为 0.3，也就是说，如果某个情感词的正确情感值为 0.2，而极性强度量化计算结果为 0.5，我们还是认为这个结果是正确，但如果极性强度量化计算结果为-0.1，

[2] http://www.keenage.com/html/c_index.html
[3] http://ir.hit.edu.cn/
[4] http://nlg.csie.ntu.edu.tw/download.html
[5] http://pan.baidu.com/share/link?shareid=520750&uk=1109069842

我们认为计算错误。另外我们还对 73 个程度副词进行作用系数人工确定，取值范围为 0.5 至 1.5，步长为 0.2。

下面我们主要安排了三个实验，第一个实验是基于 Ku 等人方法的情感词极性强度量化基准实验，第二个实验是改进后基于情感词分类的极性强度量化实验，第三个实验是针对情感未定词进行不同领域的极性强度量化统计。第一、二个实验中，只针对有明显褒贬极性的情感词进行，其中基础褒贬情感词数为 6473 个（4712 个贬义词，1761 个褒义词），复合褒贬情感词数为 1208 个（720 个贬义词，488 个褒义词）。训练集和测试集的情感词数量之比为 9:1，为确保数据平衡，我们进行了 10 折交叉验证，实验结果取十次的平均值。第三个实验中，我们主要利用酒店评论进行情感未定词的极性强度量化计算。

2.5.1 情感词极性强度量化计算基准实验

Ku 等人的方法主要利用公式 2.1-2.4 对情感词典中的情感词进行字频统计和情感计算，而没有针对情感词自身特点进行专门处理。为了进行试验结果对比，我们首先还原了 Ku 等人的方法作为基准试验，对情感词进行极性判定和情感极性强度量化计算，其中极性判定是通过对测试集中的情感词进行极性预测，极性强度量化计算就是计算情感词的情感倾向值。实验结果如表 2.1 所示。

表2.1 情感词极性判定及极性强度量化计算基准实验结果

测试集	极性判定正确率	极性强度量化计算正确率（误差系数为 α ）		
		α =0.1	α =0.2	α =0.3
褒义词测试集	0.833	0.353	0.532	0.687
贬义词测试集	0.855	0.363	0.562	0.708
正确率平均值	0.844	0.358	0.547	0.698

从表 2.1 中的极性判定正确率结果可以看出，Ku 等人的方法在贬义词的正确率略高于褒义词，分析原因主要还是在贬义词集数大于褒义词集，一定程度上增加了贬义词的覆盖面积，从而增加了贬义词集中情感词的极性判定正确率。

从极性强度量化计算正确率可以看出，随着误差系数的增大，其正确率也随之提高。这也符合我们的判断，因为在人工标注情感值的过程中，每个人对情感值的判定

都会有不同，所以我们在极性强度量化实验中允许存在误差。当误差系数为α为 0.1 时，由于误差区间给定的非常小，所以正确率普遍不高，而当误差系数α放大到 0.3 时，正确率达到最大。当然如果误差系数继续放大，正确率还将上升，如果放大到 1 的话，相当于求解极性判定正确率。所以过大的误差系数将降低结论的研究价值，本实验中的误差系数我们控制在 0.3 区域范围内。

2.5.2 基于情感词分类的极性强度量化计算

本实验主要针对 Ku 等人方法设计上的不足，研究了分类计算方法来实现情感词的极性强度量化工作。首先把明显褒贬极性的情感词划分成两类：基础情感词和复合情感词，然后分别进行计算实验。

（1）基础情感词的极性强度量化计算结果与分析

根据前面的介绍，基础情感词为首字不含有否定词、程度修饰词且字数不超过 2 个字的情感词，我们针对 4712 个基础贬义词和 1761 个基础褒义词，利用公式 2.1-2.3 和公式 2.5 进行计算实验，最后的实验结果如表 2.2 所示。

表2.2 基础情感词极性判定及极性强度量化计算结果

测试集	极性判定正确率	极性强度量化计算正确率（误差系数为α）		
		α=0.1	α=0.2	α=0.3
褒义词测试集	0.922	0.432	0.684	0.796
贬义词测试集	0.943	0.468	0.702	0.838
正确率平均值	0.933	0.450	0.693	0.817

从表 2.2 的结果中，我们发现褒、贬词集的极性判定正确率分别达到了 0.922 和 0.943，平均正确率达到了 0.932。对比表 2.1 的结果，我们不难发现，极性判定正确率有了明显的提高。这主要原因是本次实验针对的是基础情感词，对象结构比较简单，更符合字词之间的一种原始关联。另外极性强度量化计算正确率的结果也比表 2.1 的结果好许多。随着误差系数α的放大，极性强度量化计算的正确率明显提高，特别当α=0.3 的时候，平均正确率达到了 0.817。

（2）复合情感词的极性强度量化计算结果与分析

复合情感词的极性强度量化计算主要针对 1208 个复合情感词进行，利用 2.3.2

章节中设计的方法进行实验。最后的实验结果如表 2.3 所示。

表2.3 复合情感词极性判定及极性强度量化计算结果

测试集	极性判定正确率	极性强度量化计算正确率（误差系数为α）		
		α=0.1	α=0.2	α=0.3
褒义词测试集	0.911	0.402	0.636	0.746
贬义词测试集	0.924	0.434	0.658	0.778
正确率平均值	0.918	0.418	0.647	0.762

表 2.3 的结果显示，褒、贬复合情感词集的极性判定正确率分别为 0.911 和 0.924，平均正确率达到了 0.918。同样随着误差系数α的放大，极性强度量化计算的正确率不断提高，当α=0.3 的时候，平均正确率达到了 0.762。从结果上看，虽然复合情感词的极性判定及极性强度量化正确率均比基础情感词的差一些，但相比 Ku 等人的方法，还是要高出不少。

最后对表 2.2 和表 2.3 进行汇总平均，我们可以得出本方法的整体实验结果，如表 2.4 所示。

表2.4 情感词分类极性判定及极性强度量化计算结果

测试集	极性判定正确率	极性强度量化计算正确率（误差系数为α）		
		α=0.1	α=0.2	α=0.3
褒义词测试集	0.917	0.417	0.660	0.771
贬义词测试集	0.934	0.451	0.680	0.808
正确率平均值	0.926	0.434	0.670	0.790

表 2.4 是表 2.2 和表 2.3 的结果汇总，代表了基于情感词分类的极性强度量化计算方法的最终结果，对比表 2.1 基于 Ku 等人方法的基准实验结果，我们可以发现，针对同样的情感词典，基于情感词分类的极性强度量化计算方法在性能上有较大的优势，在极性判定正确率上提高了 8.2%，同时极性强度量化计算正确率也有很大提高，当误差系数α取值为 0.3 时，正确率提高了 9.2%，从而证明了我们方法的优越性。

2.5.3 基于不同领域的情感未定词极性强度量化计算

通过我们的标注及统计，在我们的词典里还存在 248 个情感未定词，由于这些词的极性及其情感倾向值往往跟具体应用领域相关，所以不能给出明确的情感值信息，人工标注时被统一标记为 0。这些情感词只能跟具体应用领域关联，才能给出较为合理的情感值。本次试验主要针对酒店领域进行情感未定词的极性强度量化计算，评论数据采集自驴评网[6]，具体爬虫实现将在第 5 章中介绍，我们针对爬虫结果抽取出其中褒贬极性最显著的 3 万条评论（褒、贬评论各 1.5 万条）。通过统计发现，由于覆盖面的问题，248 个情感未定词中有 19 个词未在 3 万条评论中出现，所以实际实验的情感词数为 229 个。我们首先安排学生对这 229 个情感未定词在酒店领域的情感倾向值进行人工标注，然后利用已区分褒贬的评论进行情感未定词的极性判定及极性强度量化计算。通过分析和研究，我们具体设计了公式 2.7-2.9 来实现情感未定词的极性判定及极性强度量化计算。

$$P_{wi} = \frac{fp_{wi}}{fp_{wi} + fn_{wi}} \qquad \text{（公式 2.7）}$$

$$N_{wi} = \frac{fn_{wi}}{fp_{wi} + fn_{wi}} \qquad \text{（公式 2.8）}$$

其中，P_{wi} 为词 *wi* 作为褒义词的权重，N_{wi} 为词 *wi* 作为贬义词的权重。fp_{wi} 为词 *wi* 出现在褒义评论中的频率，fn_{wi} 为词 *wi* 出现在贬义评论中的频率。利用公式 2.7 和公式 2.8 可以计算出每个情感未定词作为褒义词和贬义词的权重。

最后可以利用公式 2.9 计算出词 *wi* 的情感倾向值 S_{wi}。

$$S_{wi} = (P_{wi} - N_{wi}) \qquad \text{（公式 2.9）}$$

如果 S_{wi} 的值为正数，*wi* 是褒义词，负数则是贬义词，接近于 0 的话，说明 *ci* 趋向于是中性。

利用 3 万条褒贬评论，利用公式 2.7-2.9 对 229 个情感未定词进行实验，结果跟跟人工标注进行比对，如表 2.5 所示。

[6] http://www.lvping.com/hotels/

表2.5 面向酒店领域的情感未定词极性判定及极性强度量化计算结果

极性判定正确率	极性强度量化计算正确率（误差系数为α）		
	α=0.1	α=0.2	α=0.3
0.928	0.412	0.653	0.815

从表 2.5 可以看出，极性判定的正确率达到了 0.928，随着误差系数α的放大，极性强度量化计算的正确率不断提高，当α=0.3 的时候，平均正确率达到了 0.815。对比前面的实验结果，我们发现利用大规模褒贬评论进行极性判定及极性强度量化计算效果比较理想，主要原因有两个：1）大规模的褒贬评论环境下进行统计计算，覆盖面比较广，效果相对比较理想；2）由于这些词本身的情感领域依赖性，所以人工标注时往往标注为偏中间值的情感值，这样在较大误差系数条件下，正确率相对比较高。这种方法对评论语料质量的依赖性比较大，同时前提还必须知道评论语料的褒贬性。

2.6 本章小结

本章首先介绍了情感词和情感修饰词的情感极性强度模糊性特点，分析了模糊性产生的具体缘由。为了解决实际应用中去模糊化的需要，我们针对已有的情感词典资源进行整合，通过汇总、去重、筛选、判定等工作，确定了基于三种情感词类型的情感词典，然后对现有的情感词极性强度量化计算方法进行了介绍，分析其存在的问题，并提出了基于情感词分类计算的极性强度量化方法，针对不同的情感词类型设计了不同的计算方法，通过实验对比验证了方法的性能和效果。从实验结果分析，我们提出的方法更加符合情感计算特点，充分合理的利用了现有的语料资源，使得整体性能有了一定的提升。情感词极性强度量化计算的研究为细粒度情感分析中基于评价对象属性的情感计算奠定了基础。

第3章　评价对象属性及其情感表达元素的联合识别研究

3.1 引言

根据情感分析的粒度，可将情感分析任务分为三个类别：篇章级别，句子级别，词或短语级别。针对篇章和句子级别的情感分析工作开展较早，已有不少学者对此做过许多研究，如 Turney[7]提出了一种简单的无监督学习算法将评论文章分为正向（thumbs up）和负向（thumbs down），还有一些学者[5,87]直接利用情感词典识别情感词及其情感分值来估计文章的情感倾向。基于篇章或句子的情感分析相对而言是粗粒度的，因为一篇文章或一句话往往不只包含一种情感，评价对象经常被细分为多种属性，这就需要更细粒度的识别才能有效实现情感分析。词或短语级别的情感分析是细粒度的，也是本章的主要研究内容。

在词或短语级别的情感分析任务中，最重要的两个任务是识别评价对象属性和情感表达元素。其中，评价对象属性主要是领域对象的属性词，例如酒店领域里，属性词包含“设施”，“服务”，“环境”，“价格”，“交通”等；而在数码产品领域，属性词往往有“外观”、“电池”、“价格”等。而情感表达元素，则主要是情感词及其相关修饰内容，这部分跟篇章和句子级别的情感分析一样，情感词可以分为褒义和贬义两类。褒义词是对情感实体的正面评价，例如“漂亮”、“好”、“周到”、“整洁”等；而贬义词是对情感实体的负面评价，例如“差”、“脏”、“傲慢”、“乱”等。修饰内容则包含程度修饰，例如“非常”、“相当”、“很”、“比较”等，还有否定修饰，例如“没”、“不”、“未”等。

本章的工作内容是识别评论中出现的属性词，情感词及其相关修饰内容。目前已有的工作主要利用无监督和监督学习算法，无监督方法主要是利用规则和词的频率信息来识别属性词和情感词[9]，这类方法一般性能不高。相比较而言，监督学习方法能取得比较好的性能。Wilson 等人[76,77]提出基于分类的方法来识别情感词及其情感倾向。Jin 等人[75]提出基于隐马尔科夫模型的序列标注算法来识别主题词，情感词。基

于分类的方法，是独立的识别属性词和情感词，忽略了它们之间的类别关系。基于隐马尔科夫模型的算法，可以考虑句子间的序列关系，但马尔科夫模型是产生式模型，不适于充分利用高维特征。

属性词及其情感表达元素的识别任务可以看作属性词、情感词及其相关修饰内容的联合识别任务。本章主要实现属性词和情感表达元素的识别任务，对情感词极性判定及极性强度量化计算将在第四章进行讨论。通过分析和研究，我们将识别任务转化为结构化标注任务，并提出使用条件随机场模型框架来联合识别这三类实体。

本章首先介绍了条件随机场模型，对其原理及应用进行了阐述和分析。随后提出了评价对象属性及其情感表达元素的序列化联合抽取模型，详细分析了基本特征和语义特征的相关知识及抽取方法，特别针对语义特征的抽取进行了技术分析和算法设计。最后，利用条件随机场（CRF）分类器实现了评价对象属性及其情感表达元素的联合识别，并进行了实验对比和性能分析。

3.2 条件随机场模型介绍

条件随机场（CRF）由 Lafferty 等人[89]于 2001 年提出，结合了最大熵模型和隐马尔可夫模型的特点，是一种无向图模型，近年来在分词、词性标注和命名实体识别等序列标注任务中取得了很好的效果。条件随机场是一个典型的判别式模型，最常用的是线性链 CRF。若 X=（x_1，x_2，...，x_n）表示一个观测序列，而 Y=（y_1，y_2，...，y_n）表示为状态（标注）序列，在给定一个输入序列的情况下，线性链的 CRF 模型定义为：

$$p_\lambda(Y \mid X) = \frac{1}{Z(X)} \exp(\sum_{t=1}^{n} \sum_{k} \lambda_k f_k(y_t - 1, y_t, X, t)) \qquad \text{（公式3.1）}$$

其中，Y 是串的标注序列，X 是待标记的字符，$f_k(y_t - 1, y_t, X, t)$ 是一个任意的特征函数，λ_k 是对应的特征函数的权值，而 $Z(X)$是归一化因子，使得上式成为概率分布。

对于条件随机场模型的训练，一般是利用最大似然法则获得定义特征的权值。而在预测实例中，则采用 Viterbi 算法来寻找在观测序列 X 条件下，具有最大条件概率的状态序列，即：

$$Y = \arg\max_{Y} P_\lambda(Y \mid X) \qquad \text{（公式3.2）}$$

在本文中细粒度情感信息识别实际上是一个序列标注问题，所以我们采用了 CRF 模型实现。本文使用的 CRF 模型是 CRF++0.53[7]，它是著名的条件随机场开源工具，也是目前综合性能最佳的 CRF 工具之一。其具体使用过程为：（1）准备训练和测试数据；（2）准备特征模板；（3）训练；（4）测试。

其训练和测试文件必须包含多个 token，每个 token 包含多个列，且每个 token 的列数必须一致。token 列的定义可根据具体的任务，如词、词性等。每个 token 必须写在同一行上，且各列之间用空格或制表格间隔。多个 token 序列可以构成一个 sentence，sentence 之间用一个空行间隔。最后一列是 CRF 用于训练的正确标注形式。

例如下面的例子是我们从训练集中抽取出来的一个片段，其中每个 token 包含 4 列，分别为词本身、词性标注（POS）、语义角色标注（SRL），以及人工标注的类别信息。具体如图 3.1 所示。

酒店	NN	ARG0	TP
设施	NN	ARG0	TP
还行	VV	V	SO
，	PU	*	BG
服务	NN	ARG0	TP
超级	AD	*	ADV
差	VA	V	SO
，	PU	*	BG
服务	NN	ARG0	TP
态度	NN	ARG0	TP
也	AD	*	BG
很	AD	*	ADV
不	AD	*	SO
好	VA	V	SO
，	PU	*	BG

图3.1 训练集例子

[7] http://code.google.com/p/crfpp/downloads/list

因为 CRF++设计为一个泛用的工具，使用者必须事先指定特征模板(template_file)，该文件描述了训练和测试时会用到的特征情况。我们在实验过程中设计了多种类型的模板，为了方便说明，以下面模板为例进行阐述，具体如图 3.2 所示。该模板同时也是本章实验所用到的默认模板 T1。

```
# Unigram
#WORD
U01:%x[-1,0]            #上个词
U02:%x[0,0]             #当前词
U03:%x[1,0]             #下个词
U04:%x[-1,0]/%x[0,0] #上个词和当前词
U05:%x[0,0]/%x[1,0] #当前词和下个词

#POS
U11:%x[-1,1]             #上个词的POS
U12:%x[0,1]              #当前词的POS
U13:%x[1,1]              #下个词的POS
U14:%x[-1,1]/%x[0,1]   #上个词和当前词的POS
U15:%x[0,1]/%x[1,1]   #当前词和下个词的POS

#SRL
U21:%x[-1,2]             #上个词的SRL
U22:%x[0,2]              #当前词的SRL
U23:%x[1,2]              #下个词的SRL
U24:%x[-1,2]/%x[0,2]   #上个词和当前词的SRL
U25:%x[0,2]/%x[1,2]    #当前词和下个词的SRL

# Bigram
B
```

图3.2 模板示例

其中，“#”后面的部分属于注释内容，其他说明信息及训练和测试命令可以参考CRF++的官方网站 http://crfpp.sourceforge.net/。

已有不少研究者曾利用条件随机场模型来分析文本中的情感信息。McDonald等人[90]利用条件随机场分类器预测了文章和句子级别的情感类别；Breck等人[91]利用条件随机场模型识别了新闻里的情感表达；Choi等人[92]同样在新闻语料上，识别了情感的表达者。但之前没有任何工作是利用条件随机场模型来同时识别句子中的评论对象属性，情感词及其情感修饰词。

3.3 评价对象属性及其情感表达元素的序列化联合抽取模型

对于评论中的每个句子，我们希望抽取出评价对象属性、情感词以及情感修饰词。下面将具体介绍如何用条件随机场模型框架的图结构来描述句子结构，以更好地利用词与词之间的关系，提高评价对象属性及其情感表达元素同时抽取的性能。

评价对象属性、情感词以及情感修饰词的抽取可以被看作是一个简单的分类工作。每个词看作是一个实例，然后利用支持向量机或隐马尔科夫等分类器，分别独立地判断每个词的类别标签。但这种分类的方法假设词与词之间的类别标签是独立的，而实际上，词的类别标签之间具有很强的相关性。单词所处的上下文类别标签对目标词类别标签的判断也具有十分重要作用。例如，句子的序列化结构关系，对评价对象属性及其情感表达元素的类别标签判断有帮助。在一个句子中有两个连续的词，如果前一个具有情感修饰作用的副词，则它后边的形容词有很大概率属于情感词，例如“酒店的地理位置非常好”中的“非常”和“好”。另外一个例子，如果一个句子中前面的一个词是名词，而后面连续跟着带有情感修饰的副词和带有情感的形容词，则它前面那个名词有很大可能是评价对象属性，例如前面例句中的名词“地理位置”，副词“非常”和形容词“好”这三个词的关系。可见，句子中单词出现的序列化结构关系对情感词和评价对象属性的识别具有重大作用。另外除了词和词性，我们还发现语义角色标注信息对目标词的类别标签判断也具有很大的作用，评价对象属性及其情感表达元素的在句子中的语义角色往往相对固定，例如语义角色“Arg0”、“ArgM-ADV”和“V”往往标注句子中的评价对象属性、情感修饰词和情感词。所以我们在序列化结构下，针对单词的特征集合中，我们还充分利用了语义角色信息。具体语义角色标

注及特征抽取算法将在下面的章节详细介绍。

本文使用线性条件随机场来描述句子中单词出现的序列化结构关系，如图3.3所示。条件随机场模型包含两组结点，其中实心圆表示可观测变量集合，用X表示，是指单词对应的特征集合（包括分词、词性标注、语义角色标注等信息）；空心圆表示隐变量集合，用Y表示，是指要预测的类别标签集合。图3.3中的类别标签说明可以参见3.6.1。在线性条件随机场模型中，我们不难看出每个单词对应的类别标签是按在句子中的位置关系线性相连的，即在统一预测中考虑了相邻单词的类别标签关系。

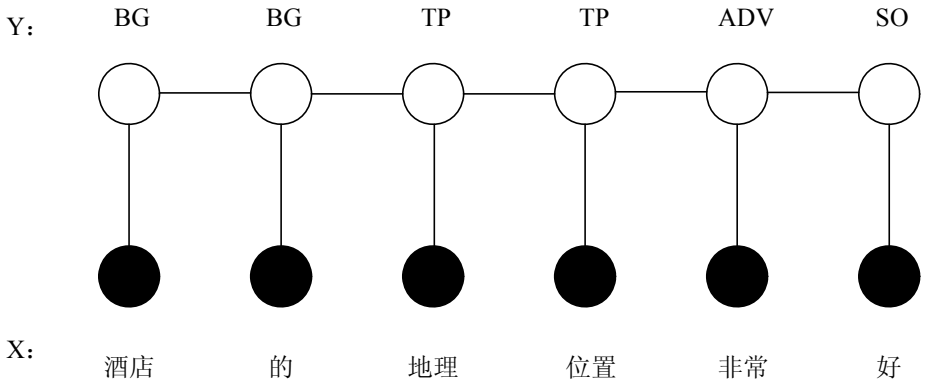

图3.3 线性条件随机场示例

3.4基本特征抽取

3.4.1 词汇信息

在基于监督学习的文本情感分析中，词汇信息特征具有十分重要的作用[93]。词是自然语言中最小的有意义的构成单位，但中文中词与词之间并没有明显的界限，因此，分词是中文信息处理的首要工作。

以往的分词方法，无论是基于规则的还是基于统计的，一般都依赖于一个事先编制的词表（词典）[94]。自动分词过程就是通过词表和相关信息来做出词语切分的决策。与此相反，基于字标注的分词方法实际上是构词方法，即把分词过程视为字在字串中的标注问题。由于每个字在构造一个特定的词语时都占据着一个确定的构词位置（即词位），假如规定每个字最多只有四个构词位置：即 B（词首），M（词中），E（词尾）和 S（单独成词），那么下面句子（1）的分词结果就可以直接表示成如（2）所示的逐字标注形式：

（1）分词结果: / 上海/ 计划/ 到/ 本/ 世纪/ 末/ 实现/ 人均/ 国内/ 生产/ 总值/ 五千美元/ 。

（2）字标注形式: 上/ B 海/ E 计/ B 划/ E 到/ S 本/ S 世/ B 纪/ E 末/ S 实/ B 现/ E 人/ B 均/ E 国/ B 内/ E 生/ B 产/ E 总/ B 值/ E 五/ B 千/ M 美/ M 元/ E。/ S

由于本文对分词系统不做深入研究，所以直接采用了苏州大学自然语言处理研究所自行研发的分词系统。但我们发现该分词系统的训练集主要是时政新闻方面的文章，如果直接用来处理酒店评论则分词出错率较高，如“性价比”这词在训练集中没有出现，所以当对酒店评论进行分词时，往往不会把“性价比”看作一个整体词汇，从而导致后面整个分析的错误。为了减少由于领域变化致使分词出错率上升的情况，我们通过修改训练集，加入一部分酒店评论方面的分词训练数据，重新训练，生成新的分词模型，然后再对酒店评论数据进行分词预测。通过测试发现利用这种方法处理之后，酒店评论的分词性能有明显提高。

3.4.2 词性标注

词性标注（Part-of-Speech tagging，POS tagging）是指对句子中的每个词指派一个合适的词性，也就是要确定每个词是名词、动词、形容词或其他词性的过程，又称词类标注或者简称标注。词性标注是自然语言处理中的一项基础任务，在语音识别、信息检索及自然语言处理的许多领域都发挥着重要的作用。

如果每个单词仅仅对应一个词性标记，那么词性标注就非常容易了。但是语言本身的复杂性导致了并非每一个单词只有一个词性标记，而存在一部分单词有多个词性标记可以选择，如“鼓励”这个词，既可以是动词（“老师鼓励我们好好学习”），也可以是名词（“这是对我们的一种鼓励”）。因此，词性标注的关键问题就是消解这样的歧义，也就是对于句子中的每一个单词在一定的上下文中选择恰如其分的标记。大多数的标注算法可以归纳为三类：一类是基于规则的标注算法（rule-based tagger)，一类是随机标注算法（stochastic tagger），最后一类是混合型的标注算法。基于规则的标注算法一般都包括一个手工制作的歧义消解规则库；随机标注算法一般会使用一个训练语料库来计算在给定的上下文中某一给定单词具有某一给定标记的概率，如基于 HMM 的标注算法；而混合型标注算法具有上述两种算法的特点，如 TBL

（transformation-based learning）标注算法。

本文主要针对中文评论进行处理，标注集采用了 Penn Treebank 标记集[95]，词性标注系统采用了苏州大学自然语言处理研究所自行研发的 POS tagger 系统。

3.5 语义特征抽取

本节主要介绍了语义角色的基础知识以及语义角色在诸多自然语言处理领域的应用研究，并设计出从语义角色标注结果中抽取生成相应的语义特征算法。

3.5.1 语义角色基础知识

语义角色是句子中的名词短语在相应动词驱动下所承担的语义成分，可以利用语义角色标注（semantic role labeling）来获得语义角色。语义角色标注是浅层语义分析（shallow semantic parsing）的一种实现方式。浅层语义分析是指根据句子的句法结构和句中每个实体的词义推导出能够反映这个句子意义的某种形式化表示。语义角色标注不对整个句子进行详细的语义分析，只是针对给定谓词（动词、名词等），标注句子中某些短语承担的语义角色，这些短语作为给定谓词的框架的一部分被赋予一定的语义含义。例如“[委员会 Agent][明天 TMP]将要[通过 V][此议案 Passive] 。”其中“通过”为谓词，“委员会”、“此议案”和“明天”分别是其施事、受事和发生的时间。另外，语义角色标注不考虑时态信息，例如“Thomas hit James.”与“James was hit by Thomas.”，虽然时态不同，但语义角色表示是相同的，Thomas 是施事者，James 是受事者，表示成语义的形式可统一为：“hit (Thomas，James)”。同时，语义角色标注也不考虑谓词改变，但语义不变的情况。例如，“罗贯中写了《三国演义》”与“《三国演义》的作者是罗贯中”，虽然它们语义相同，但由于谓词不同，语义角色标注的结果并不一样。

要进行语义角色标注，也需要语料资源的支持。目前比较著名的英文语义角色标注资源包括 FrameNet[96]、PropBank[97]和 NomBank[98]三种。

FrameNet 是以框架语义为标注的理论基础对英国国家语料库进行标注，试图描述每个谓词（动词、部分名词和形容词）的语义框架，以及这些框架之间的关系。从 2002 年 6 月发布至今，共标注了约 49000 句，每句都标注出了目标谓词、语义角色，

该角色句法层面的短语类型以及语法角色（主语、宾语等）。

PropBank 是宾夕法尼亚大学在 Penn TreeBank 句法分析语料库基础上标注的语义角色标注语料库。它只对动词进行标注，而且只包含 20 多个语义角色。其中核心的语义角色有 Arg0~Arg5 六种，Arg0 通常表示动作的施事者，Arg1 通常表示动作的影响，Arg2~Arg5 根据谓语动词不同会有不同的语义含义。其余的语义角色为附加语义角色，使用 ArgM 表示，例如 ArgM-LOC 表示地点，ArgM-TMP 表示时间等等。PropBank 基于 Penn TreeBank 手工标注的句法分析结果，因此标注结果几乎不受句法分析错误的影响，准确率较高。而且它几乎对 Penn TreeBank 中的每个动词及其语义角色进行了标注，因此覆盖范围更广。

NomBank 是为了弥补 PropBank 仅以动词为谓词的缺点而开发的，它标注了 Penn TreeBank 中的名词性谓词及其语义角色，而且它允许角色出现互相覆盖的情况。

除英文外，许多其它语言也建立了各自的语义角色标注库，其中 Chinese PropBank 是宾州大学基于 Chinese Penn TreeBank[99]标注的汉语语义角色标注资源，标注方法参考了英文 PropBank。

目前不同的语料库和具体 NLP 任务，对于语义角色集的定义并不统一，只有施事者（Arg0）和受施者（Arg1）这两个语义角色是稳定的。根据 Shen 和 Lapata[100]的统计，Arg0 和 Arg1 占到了各种语义角色总量的 85%以上。而且，目前绝大多数的 SRL 工具对 Arg0 和 Arg1 的标注结果较好，Arg0 的准确率都达到了 85%以上，而 Arg1 的准确率也达到了 78%左右。考虑到情感分析系统的通用性，以及对 SRL 结果的依赖度，再加上情感分析系统中已经通过分词、词性标注子任务对各句子进行了简单分析，因此本文主要考虑了 Arg0 和 Arg1 这两个语义角色和谓词信息。由于在抽取情感表达元素过程中，我们还对情感修饰词进行了识别，所以在 SRL 语义信息抽取过程中进一步考虑了 ArgM-ADV。

3.5.2 语义角色的应用研究

近年来，随着各类语义角色资源库的不断建立，以及从 2004 年开始的与 SRL 相关的各类评测活动的举行，极大的推动了语义角色标注的发展。在语义角色标注性能不断发展的基础上，越来越多的研究者开始关注如何将 SRL 的研究成果应用于其他

各类自然处理领域，并取得了一定的成果。

在问答系统领域，Narayanan 等人[101]首次分析了语义角色对回答复杂问题的重要作用。在问答系统中，他们首先从 PropBank 和 FrameNet 中获得语义角色信息，再将这类信息应用于谓词论元结构的识别，将识别的谓词论元结构与领域特定主题模型相结合计算出各类答案的贴切概率，根据概率获得最佳答案。

在信息抽取领域，Surdeanu 等人[102]给出了一种自动识别谓词论元结构的方法，再利用识别出的谓词论元结构来指导信息抽取，他们进行的各种实验表明，谓词论元结构的加入能极大提升 IE 的效果。

在文本摘要领域，Melli 等人[103]给出了一个文本摘要系统的完整设计。在系统中，他们引入了语义角色信息，从与问题主题相关的一组文档中，基于语义子图选择相应的句子，再生成基于问题的摘要。通过与其他系统的比较发现，语义角色非常有助于文本摘要性能的提升。

而在文本情感分析领域也有一些研究者尝试引入语义角色信息来提升情感分析的性能。Kim 等人[78]利用 FrameNet 的标注数据，并根据他们所提出的问题设计出基于框架语义结构的 SRL 系统，然后通过角色映射抽取出意见、意见持有者和主题。主要设计思想为：第一，收集情感词，并通过情感词找出与情感相关的框架；第二，通过相关的特征设计，抽取出对象词汇（target word），短语类型（phrase type），中心词（head word），分析树路径（parse tree path），位置（position），框架名称（frame name）等特征，开发出基于框架语义的 SRL 系统；第三，利用相应规则系统选择标注为相关语义角色的内容映射成意见持有者和主题。结果显示，Kim 等人所提出的这种方法比简单的以词性推断为基础的基准方法（baseline）在性能上高出许多。但由于 Kim 等人的研究主要是基于 FrameNet 的标注数据，在领域上具有一定的局限性。另外更重要的是在研究任务上这种基于简单的映射方法很难适应本文以商品评论为对象的研究，并且 Kim 等人的方法过多依赖人工判定，所以很难实现情感分析的自动化。

本文将研究 SRL 信息对评价对象属性及其情感表达元素联合识别性能的影响。通过设计相应的抽取算法，合理地提取出 SRL 相关信息作为本系统的重要语义特征，最后通过实验对比验证该特征对提升整个系统性能的效果。

考虑到系统集成的方便，本文使用了由实验室 SRL 小组自行开发的 SRL 工具，

以 Chinese Penn TreeBank 5.1 给定的数据集合为实验语料，通过实验室 SRL 小组已发表的论文来看，本系统性能达到了国际先进水平。

3.5.3 基于语义角色的语义特征抽取

从语义角色标注任务的定义可以看到，谓语动词在语义角色信息的描述中具有非常重要的作用，同一个名词性短语可能是谓词 A 驱动下的 Arg0 角色，同时又可能是谓词 B 驱动下的 Arg1 角色。因此在描述语义角色相关特性时，我们必须设计相应算法从候选角色中选择最合适的角色作为语义特征信息。

我们的任务是从 SRL 标注文档中抽取出合适的语义角色特征信息，主要抽取的角色类型有 Arg0、Arg1 和 ArgM-ADV（简写为 ADV），以及谓词信息（简写为 V）。下面是对我们的评论语料进行 SRL 标注之后的一个片段，具体如图 3.4 所示。

```
酒店  * NN * * (IP(NP* - - - (ARG0* * * * * * *
设施  * NN * * *) - - - *) * * * * * *
还行  * VV * * (VP* - -  还行  (V*) * * * * * *
，    * PU * * * - - - * * * * * * *
服务  * NN * * (IP(IP(NP*) - - - (ARG1* (ARG0*) * * * * *
超级  * AD * * (VP(ADVP*) - - - * (ARGM-ADV*) * * * * *
差  * VA * * (VP*))) - -  差  * (V*) * * * * *
，    * PU * * * - - - * * * * * * *
服务  * NN * * (IP(NP* - - - * * (ARG0* * * * *
态度  * NN * * *) - - - * * *) * * * *
也  * AD * * (VP(ADVP*) - - - * * (ARGM-ADV*) * * * *
很  * AD * * (ADVP*) - - - * * (ARGM-ADV*) * * * *
不  * AD * * (ADVP*) - - - * * (ARGM-ADV*) * * * *
好  * VA * * (VP*))) - -  好  * * (V*) * * * *
，    * PU * * * - - - * * * * * *
.....
```

图3.4 SRL标注文档示例

由于叙述侧重点的关系，针对 SRL 格式的具体说明不在本文中展开，具体可以参见 CoNLL-2005 中的任务说明[8]。针对 SRL 标注文档，我们可以按照 CRF++所规定的格式要求，设计相应的语义特征提取算法，实现相关语义特征的抽取。相对于图 3.4 的 SRL 标注内容，通过应用语义特征提取算法，可以实现相关语义特征的抽取，结果如图 3.5 所示。

酒店	NN	ARG0
设施	NN	ARG0
还行	VV	V
，	PU	*
服务	NN	ARG0
超级	AD	ADV
差	VA	V
，	PU	*
服务	NN	ARG0
态度	NN	ARG0
也	AD	ADV
很	AD	ADV
不	AD	ADV
好	VA	V
，	PU	*
.....		

图3.5 语义特征抽取结果

其中“*”表示不含有语义角色信息。抽取结果分为三列，第一列为词本身，第二列为词性标注，第三列为语义角色标注。另外我们从图3.4中可以看出“服务”这个词根据不同的谓词具有不同的语义角色：ARG0和ARG1，所以在生成最后的特征信息之前必须进行合理的选择。我们通过大量实例发现以下规律：1）由于我们研究的是细粒度情感分析，所以在语义角色选择过程中，选择作用范围最小的语义角色最

[8] http://www.lsi.upc.edu/~srlconll/conll05st-release/README

为合理。2）另外作用范围过大的语义角色信息往往没有实质性的作用，所以我们对标注范围大于一定阈值的角色信息进行了过滤。主要算法思想如图3.6所示。

```
输入：句子的SRL标注结果文档document
输出：句子中每个词对应的唯一SRL标注
for sentence in document
    for word in sentence
        wordTagList:= {SRL tags}   //每个词对应的SRL标记列表
        SRLMatrix.append(wordTagList)   // SRLMatrix是一个矩阵
    Endfor
    if SumSameTagOfAdjacentWords(SRLMatrix)>Threshold
        set tags null
     /*找出所有同一列中标注为相同标记的连续词数大于Threshold的那些词并设置其标记为
     空，SumSameTagOfAdjacentWords(SRLMatrix)为计算SRLMatrix矩阵中同一列标注为相
     同标记的连续词数。*/
    endif
    sort each wordTagList by increment using SumSameTagOfAdjacentWords(SRLMatrix)
   /*  通过排序，使最细粒度的标记放到wordTagList的首位*/
endfor
return wordTagList[0]
```

图3.6 语义特征抽取算法

这里的Threshold值可以预先设定，因为有些句子会出现连续很多词为相同的SRL标记，而真正有用的SRL标记往往是比较细分的。所有我们这里通过设定Threshold的值进行预处理，过滤掉那些粗粒度的SRL标记信息。为了实验比较，我们设定Threshold值为2或3，默认值为2。

3.6训练集构建

3.6.1 标注集

由于训练语料有限，如果标记过于复杂容易出现特征稀疏，所以我们在保证细粒度情感信息有效识别的前提下，尽量采用相对简单的标注集。通过参考Fu等人[104]的定义方法，并结合自身研究目标，定义了四种标记，具体参考表3.1。标注结果序列中如果有出现连续相同标注，我们则判定其为同一对象。

表3.1 标注集及相关说明

标注集	相关说明
<TP>	评价对象特征
<SO>	情感词
<ADV>	情感修饰词
<BG>	其他背景词汇

以一句酒店评论为例进行说明。

“地理位置很好，服务态度一般”其标注结果为：

“地理/TP 位置/TP 很/ADV 好/SO ，/BG 服务/TP 态度/TP 一般/SO”

通过标注，我们可以清楚地分析出这句评论中，评价对象特征为“地理位置”和“服务态度”，分别对应的情感词是“好”和“一般”，同时“很”标注为情感修饰词，“，”标注为背景词汇。

3.6.2 评论语料的人工标注

评论语料我们使用了谭松波老师的酒店评论[9]，褒贬各2000篇，通过取重之后统计得出褒义评论中含有7277个句子，贬义评论含有9978句子。由于人工标注的工作量十分巨大，我们只标注了大约四分之一的语料，并通过去除只含有“BG”标记的句子来实现非主观性句子的过滤。标注工作，我们安排了三个学生进行，对相同的待标注语料进行人工标注。通过设计程序分析三个人的标注结果，发现存在26%的不一致率。为了达到形式上的统一，把标注结果中不一致的26%抽取出来进行集体讨论，达

[9] http://www.searchforum.org.cn/tansongbo/corpus-senti.htm

成统一的结果。最终我们生成了1080句主观性句子作为实验数据集，其中包含了24890个token。

3.7实验结果与分析

3.7.1 实验设置

（1）基本设置

本实验采用有监督学习方法，为避免过学习以及欠学习状态的发生，我们采用10 折交叉验证，即随机将数据集分成 10 份，取其中的 9 份做训练，最后的一份做测试，重复 10 次，最后取平均值。为了保证算法的可比性，所有的实验方法默认都是在同样的 10 折交叉验证中完成的。另外我们还把 1080 句标注训练集分成不同大小的数据集合，都分别进行一次 10 折交叉验证，以观察不同数据集大小条件下的实验结果差异性。

针对 CRF 分类器的选择，我们选用了 CRF++0.53。同时我们只考虑了当前词和上下相邻词的信息以及当前词跟相邻词的组合作为其默认特征模板 T1，如图 3.2 所示。另外我们采用了实验室研发的 SRL 标注工具实现文本的浅层语义分析。

（2）评价指标

我们的评价指标是在词组级别的严格匹配，即只有当词组中的所有类别标签都被正确识别，我们才认为这个词组被正确识别。我们采用准确率 Precision（P），召回率 Recall（R），和 F 值来评价最终的结果

$$Pecision = \frac{Number\ of\ correct\ identified\ phrases}{Number\ of\ all\ identified\ phrases} \quad \text{（公式3.1）}$$

$$Recall = \frac{Number\ of\ correct\ identified\ phrases}{Number\ of\ all\ correct\ phrases} \quad \text{（公式3.2）}$$

$$F - Score = \frac{2 * Precision * Recall}{Precision + Recall} \quad \text{（公式3.3）}$$

3.7.2 基准系统实验结果

我们首先建立一个基准系统（Baseline System），其基本特征的抽取主要参考了Hu 等人[9]所提出的意见，他们认为词性是判断情感信息的重要依据，因此我们在基准系统中只考虑了词汇和词性标注这两种特征信息（Word+POS）。由于没有用到 SRL 特征信息，所以在 CRF++的默认模板中我们应该注释掉 SRL 那部分信息。最后实验在不同数据集条件下进行 10 折交叉验证，其结果如表 3.2 所示。

表 3.2 基准系统实验结果

Data size	Results		
	R	P	F
108	0.707	0.887	0.787
216	0.737	0.917	0.817
540	0.799	0.916	0.854
864	0.829	0.917	0.871
1080	0.842	0.925	0.882

该实验中，Threshold 的值我们设置为默认值 2，为了比较不同 Threshold 值对实验结果的影响，我们还选择 Threshold 值为 3 对前面实验进行了重复，结果如表 3.3 所示。

表 3.3 基准系统实验结果（Threshold=3）

Data size	Results		
	R	P	F
108	0.71	0.891	0.790
216	0.734	0.917	0.815
540	0.799	0.915	0.853
864	0.831	0.918	0.872
1080	0.842	0.924	0.881

表 3.2 和表 3.3 的结果显示当 Threshold 值为默认值 2 时，整体性能还略比 Threshold 值为 3 时要好。该结论同时还证明了我们的判断：连续多词为相同 SRL 标记的粗粒度信息对我们的识别评价对象属性及其情感表达元素的任务并没有正效用，而真正有用的 SRL 标记往往是比较细分的。因此，在后面的实验里，我们继续选择 Threshold 值为 2 作为默认值。从表 3.2 和表 3.3 的结果中还可以发现一些规律，当随着实验数据集不断增大，整体性能也在不断递增，另外实验结果中正确率均要比召回率要好，这说明在识别过程中存在不少被遗漏的识别对象，需要我们引入更多的特征信息进行挖掘。

3.7.3 引入语义特征后的系统实验结果

在基准系统的基础上，利用 3.5.3 节设计的算法抽取出相应的 SRL 信息，作为第三组特征信息加入，这样我们系统的主要特征升级为：Word+POS+SRL。同样在相同的 CRF 默认模板 T1 下进行实验，采用不同的数据集进行 10 折交叉验证。其结果如表 3.4 所示。

表3.4 引入语义特征后的系统实验结果

Data size	Results		
	R	P	F
108	0.730	0.897	0.805
216	0.764	0.902	0.827
540	0.804	0.919	0.858
864	0.834	0.922	0.876
1080	0.852	0.932	0.890

表 3.4 的结果与表 3.2 进行对比，在各个相同的数据集条件下，性能均有一定的提升，结果表明 SRL 特征信息能有效帮助系统提高评价对象属性及其情感表达元素的联合识别性能。由于引入 SRL 特征信息，使得训练过程中，系统充分考虑了关键几个语义角色的作用，生成后的模型在一定程度上能帮助系统做出更正确的预测。

3.7.4 不同模板条件下的系统实验结果

通过上面两个实验我们发现，引入 SRL 语义特征信息之后，总体系统的识别性能得到了提高。同时我们还发现在小数据集中，SRL 特征信息对实验结果的影响比较大，随着数据集增大到一定程度，性能影响相对变小。我们针对特征列表（Word+POS+SRL）设计了一个条件随机场新模板 T2，该模板充分考虑了 POS 特征信息和 SRL 特征信息的组合，在模板 T1 的基础上，增加了 U31:%x[0,1]/%x[0,2]。同时为了对比不同模板条件下的实验结果，我们把图 3.2 所示的默认模板 T1 中有关上下文信息的条件项去掉，也就是只保留 U02:%x[0,0]，U12:%x[0,1]，U22:%x[0,2]这三项，生成一个新的模板 T0。实验在相同数据条件、不同特征模板下进行，具体实验结果如表 3.5 所示。

表3.5 不同特征模板条件下的实验结果对比

Data size	Results（T0）			Results（T1）			Results（T2）		
	P	R	F	P	R	F	P	R	F
108	0.647	0.854	0.729	0.730	0.897	0.805	0.734	0.895	0.807
216	0.670	0.846	0.745	0.764	0.902	0.827	0.775	0.907	0.836
540	0.749	0.867	0.801	0.804	0.919	0.858	0.809	0.924	0.863
864	0.774	0.879	0.823	0.834	0.922	0.876	0.84	0.928	0.881
1080	0.797	0.884	0.837	0.852	0.932	0.890	0.855	0.939	0.895

结果显示 T2 模板和 T1 模板在召回率（Recall）指标上基本相同，但精确度（Precision）指标上，T2 模板要略微领先。而 T0 模板的性能最差，因为我们在模板 T0 中没有考虑上下文信息。表 3.5 的实验结果再次验证我们提出的评价对象属性及其情感表达元素序列化联合抽取模型的合理性。在模板 T1 和 T2 的设计过程中，我们考虑了上下文信息，由于单词所处的上下文的类别标签对目标词类别标签的判断具有十分重要作用，所以实验结果也相对更好。

综合分析上述所有实验的结果，如果我们只考虑 F 值的情况，可以通过图 3.7 来

进行汇总表示。从图 3.7 的结果中，我们可以清晰的看出随着模板的改进，系统性能不断提高，另外在模板 T2 中还考虑了 SRL 特征信息与 POS 特征信息的组合，这对性能提升起到了一定的作用，同时说明当 POS 信息和 SRL 信息一起作用的时候对情感信息的识别效果能够达到最佳。

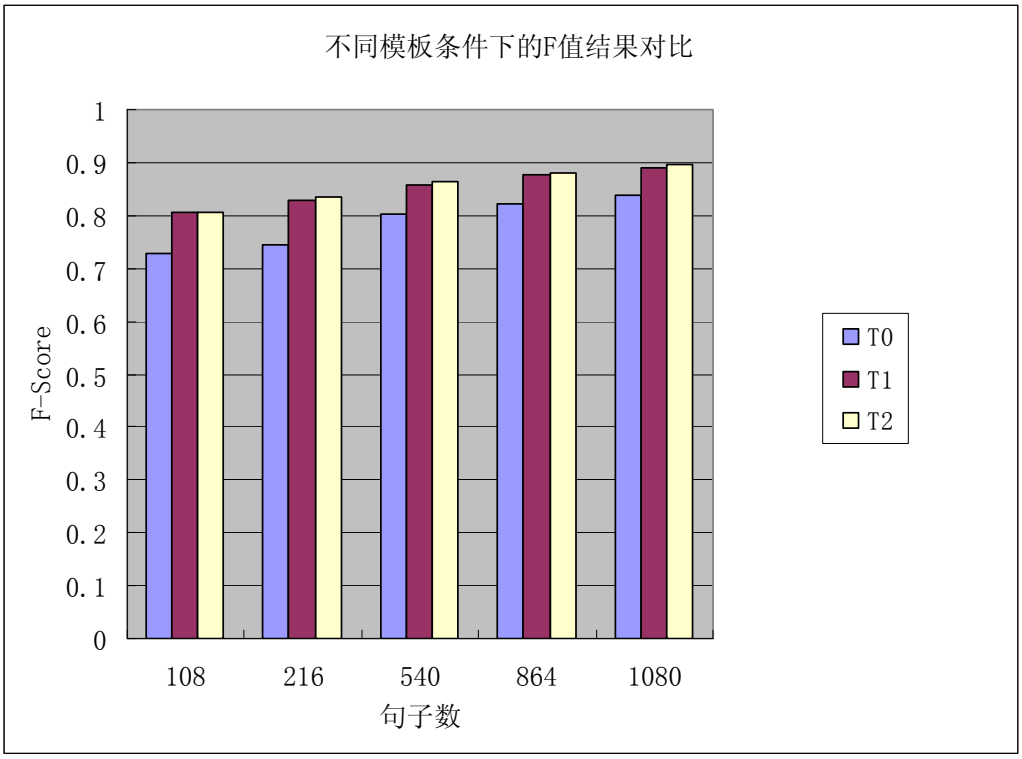

图3.7 不同模板条件下的F值结果对比图

最后我们通过分析原始评论及其标注结果，发现了影响实验结果的两个主要问题：

（1）分词错误。分词出错的原因有多方面，主要因为评论信息更多的是由用户生成的，存在不可避免的错别字现象，另外数字和汉字的混合、以及标点符号不正确标注都有可能导致分词结果出错。

（2）POS、SRL错误。分词结果出错将直接影响到POS和SRL的标注结果，另外句法分析出错也会影响SRL的结果，从而影响语义特征的抽取性能，最终对情感信息的识别性能产生级联影响。

所以如何有效提高主观性文本的分词性能是目前情感分析任务中最基本的问题，因为它直接影响后面的工作。

3.8本章小结

将评价对象属性，情感词转化为结构化标注任务，有以下三个好处：第一，能有效利用多种特征，来识别这三类实体。主题词和情感词的识别是一个比较复杂的任务，与其关联的因素有很多，比如词汇原型，词所在的句子，词的情感先验等等，因此我们需要综合多种特征来完成实体识别的任务。隐马尔科夫模型是产生式模型，一般在综合各种特征时假设特征之间是独立的，限制了其应用空间；而条件随机场模型是区分式模型，能有效利用多种特征，取得较好的性能。第二，结构化序列标注模型能有效利用句子的语言学结构；条件随机场模型框架是概率图模型的一种，它可以利用概率图的结构有效表达句子的语言学结构，充分利用句子的各种语言学结构来提升实体识别的性能。第三，它可以充分利用评价对象属性词与情感词、情感修饰词之间的关系。在结构化标注模型中，评价对象属性词、情感词以及情感修饰词的识别不是独立的，而是同时识别，这样可以有效利用它们之间的类别关系来提升系统的性能。

本章首先介绍了条件随机场模型，对其原理及应用进行了阐述与分析。随后提出了评价对象属性及其情感表达元素的序列化联合抽取模型，详细分析了基本特征和语义特征的相关知识及抽取方法，特别针对语义特征的抽取进行了技术分析和算法设计。最后，利用机器学习方法实现了评价对象属性及其情感表达元素的联合识别，并通过实验对比，证明了方法的优越性。

第4章　细粒度属性分类及情感计算

4.1引言

与情感词相类似，评价对象属性的描述也是多种多样，同一类对象属性，可以有多种语言表达，如“外观”，相似的描述可以有“外形”、“外表”、“表面”等。虽然这些词语不相同，但描述的含义、概念是基本相同的。细粒度情感计算工作之前，评价对象必须确定好属性类别，以方便情感汇总统计。所以属性分类（attribute classification）工作对细粒度情感分析十分重要，尽管已有 WordNet（英文）、哈工大的同义词林等资源能够在一定程度上帮助属性分类，但由于存在领域相关性、资源局限性等情况，要在实际应用中实现有效的属性分类还比较困难。所以如何有效、正确的进行属性分类是细粒度情感计算、情感汇总的首要工作。

已有不少研究者进行了产品属性分类的相关研究。Guo 等人[105]提出了利用多层潜在语义关联技术（multilevel latent semantic association，mLSA）实现产品属性描述的归类，mLSA 属于非监督方法，通过设计相关规则以及进行统计计算获得关联信息，从而实现产品属性描述的归类。Zhai 等人[106]提出了弱约束（soft-constrained）期望最大化（expectation maximization）算法（简称 SC-EM）结合半监督学习方法实现产品属性的归类（feature grouping）。他们的方法一定程度上很好的实现了属性分类，但缺乏对方法本身的深入分析，比如半监督学习在属性分类问题上，肯定存在种子集选择、种子集大小、迭代终止条件设计和迭代时间等问题。我们将在后面的半监督学习方法中进行深入探讨和分析。

本文主要以酒店评论为例，参考相关网站（ctrip.com、qunar.com 等）以及相关论文介绍，主要划分了 7 大类属性，分别为：“设施”，“服务”，“环境”，“价格”，“餐饮”，“交通”，“整体”。另外根据统计，在前面章节实验中用到的 1080 个评论句里，含有 366 个不同的属性词语描述，累计出现了 1558 次。

本文首先通过有监督学习方法对属性分类进行研究，在特征设计时主要利用了属性识别结果及其上下文词汇信息，以及属性识别结果对应的词性标注信息及其上下文词性标注信息。由于细粒度情感标注语料的资源少、标注工作量大的特点，本文在属

性分类研究中还引入了半监督学习方法，以减少对标注语料的依赖。首先研究了自举学习（一种半监督学习方法）的分层种子选取策略，并与随机种子选取策略在属性分类上进行了实验性能的对比；另外还研究了把分层思想应用到自举过程的每一步迭代之中，探讨了自举迭代的终止条件；针对评论中可能存在情感词缺少对象属性的情况，我们研究通过计算 PMI 值来确定评价对象属性类与情感词之间的关联概率，实现对缺失评价对象属性的情感信息进行合理属性类的指派，使情感汇总计算更为合理有效。

4.2 基于监督学习的属性分类研究

分类器的种类多种多样，如 Bayes 分类器、MaxEnt 分类器、CRF 分类器和 SVM 分类器等等，它们的性能各不相同，要求也大不一样。为了对比监督学习中分类器性能，我们还选用了最大熵分类器进行实验。

4.2.1 最大熵模型介绍

最大熵分类器的理论基础是最大熵模型，1992 年，Pietra 等人[107]首先把最大熵方法应用于自然语言处理。最大熵模型是一个比较成熟的统计模型，适合于解决分类问题。其基本思想是，给定一已知事件集，在已知事件集上挖掘出潜在的约束条件，选择一种模型，而把所有未知的事件排除在外。这个模型必须满足已知的约束条件，同时对未知事件，尽可能使其分布均匀。假设 d 表示某一具体事件，c 表示该事件被分类的结果。那么，如何表示从事件集上得到的约束条件呢？研究者引入了特征函数（有时简称为特征）的概念。特征函数一般为二值函数，对于分类问题，可选择“特征——类别”对作为一个特征函数，比如对于特征 w 和类别 c'，它的特征函数如公式（4.1）所示：

$$f_{w,c'}(d,c)=\begin{cases}1 & c=c' \ \&\ d \text{ cotains } w\\ 0 & otherwise\end{cases} \qquad \text{（公式4.1）}$$

给定特征集合后，首要的任务是基于训练集合计算每个特征的期望值，每个特征的限制条件都要求这个经验期望（empirical expection）与模型中的理想特征期望值相同。在所有满足限制的概率分布模型中，选取满足使熵值最大化的分布。

利用最大熵模型得出在特征限制条件限制下具有最优的概率分布，即概率值 $p(c|d)$。根据最大熵原理，概率值 $p(c|d)$ 的取值符合公式（4.2）的指数模型：

$$p_\lambda(c \mid d) = \frac{1}{Z_\lambda(d)} \exp(\sum_i \lambda_i f_i(d,c)) \quad \text{（公式4.2）}$$

其中 $Z_\lambda(d)$ 为范化常数，$Z_\lambda(d) = \sum_c \exp(\sum_i \lambda_i f_i(d,c))$；$f_i$ 为特征函数，λ_i 表示特征函数 f_i 的权值，即特征函数 f_i 对于模型的重要程度。根据最大熵原理可知，求最优的概率分布模型转化为求参数 $\lambda_i (1 \le i \le n,$ 其中 n 为特征函数总数) 。在最大熵模型中，估算参数 λ_i 常采用的方法是 Darroch 和 Ratcliff 提出的通用迭代缩放算法（Generalized Iterative Scaling, GIS），或 Della Pietra 提出的改进迭代算法（Improved Iterative Scaling, IIS）。

属性分类中，在预测一个属性描述词是否为某一属性类过程中会涉及各种因素，假设 X 就是一个由这些因素构成的向量，变量 y 的值为语义角色类型。p(y|X)是指系统对某个属性描述词预测为某一属性类的概率。这个概率可以用上述思想来估计。最大熵模型要求 p(y|X)在满足一定约束的条件下，必须使公式（4.3）定义的熵取得最大值：

$$H(p) = -\sum_{X,y} p(y \mid X) \log(y \mid X) \quad \text{（公式4.3）}$$

最大熵模型的一个最显著的特点是其不要求具有条件独立的特征。因此，研究者可以相对任意地加入对最终分类有用的特征，而不用顾及它们之间的相互影响。另外，最大熵模型能够较为容易地对多类分类问题进行建模，并且给各个类别输出一个相对客观的概率值结果，便于后续推理步骤使用。同时，最大熵的训练效率相对较高。上述优点使其成功应用于包括分词、词性标注、句法分析、机器翻译和信息抽取等多个自然语言处理领域。

本文在属性分类实验中采用最大熵分类器与前面介绍的 CRF 分类器进行对比，所使用的最大熵分类器工具包原型为 maxent-2.5.3[10]。该工具包为开源软件，支持多元分类，同时支持特征向量值为字符型。

4.2.2 特征设计

属性分类任务主要针对不同的属性描述进行有效的归类。在特征设计时主要利用

[10] https://sourceforge.net/project/showfiles.php?group_id=5961

现有的属性识别结果及其上下文词汇信息，以及属性识别结果对应的词性标注信息及其上下文词性标注信息。以属性识别结果作为基础特征，通过多个实验比较，验证上下文窗口大小对其属性分类性能的影响，以及引入词性标注信息之后的性能变化。

（1）属性描述

属性识别任务在本文第三章中已进行了相关介绍，并通过实验验证了相应的工作任务。由<TP>标记的数据对象就是属性描述特征信息，虽然语言描述呈现出多样化特性，但人们往往可以由其字面信息找到对应的属性类别。这也就是本实验的基本特征信息。

（2）词性标注信息

在属性分类工作中，已识别出来的属性对象所对应的词性标注信息往往具有一定的规律性，所以当词性标注信息跟属性描述以及相应的上下文相结合，分类效果将有可能达到更好性能。本实验通过引入词性标注特征，并结合属性描述特征及其上下文信息，验证性能改善效果。

（3）上下文信息

上下文信息是主要通过分析当前属性描述词或其词性标注在一定窗口范围内的信息，通过抽取相应词汇或词性标注信息，生成特征，以期提高属性分类性能。

4.2.3 训练集构建

本节主要利用监督学习方法实现属性描述词的归类，由于没有直接与属性分类任务相关的数据集，所以我们利用通过对第三章中1080句评论语料分析后抽取出来的总共1558个属性描述词。本实验将对1558个属性描述进行相关特征抽取以及利用人工标注完成属性归类，从而实现训练集的构建。

由于实验结果对比需要，我们利用不同的上下文特征，构建了多个训练集。实验中，针对属性对象及其POS信息，都选取了上下文窗口1、3、5、7、9，通过不同组合测试，获得结果最优的特征组合集。

针对所有1558个属性语料集，我们首先安排了两位研究生进行属性类的人工标注，同时针对不同的特征组合，设计相关程序自动生成相应的1558个训练单元。属性分类实验中主要计算了分类正确率，考虑到数据的分布不均匀性，我们均采用了10

折交叉验证。

4.2.4 实验结果及分析

4.2.4.1 实验设置

本实验主要针对情感对象的属性进行分类，通过抽取不同的特征组合进行性能对比，从而验证出最理想的特征表示。

在特征组合上，为了验证词性标注特征信息的作用，首先只选用了属性词及其上下文信息作为基础特征组，并通过10折交叉验证，计算出正确率。然后引入词性标注信息及其上下文，抽取生成相应的特征表示，通过10折交叉验证，再次计算出正确率。分析前后实验结果，作出相应的实验结论。

我们首先在“只有属性描述词及其上下文特征”的属性分类实验中选择最大熵分类器与条件随机场分类器进行对比实验，并作出相应的结果分析，从而确定性能最好的分类器作为后面半监督学习实验的分类器工具。

4.2.4.2 属性分类实验结果

（1）只有属性描述词及其上下文特征

首先利用最大熵分类器进行构建，通过前面的分析，我们首先利用属性描述词及其上下文信息，针对1558个属性描述词建立训练集，然后进行10折交叉验证，分别计算出相应的正确率。结果如表4.1所示。

表4.1 基于最大熵分类器的属性分类正确率统计表

属性上下文窗口	正确率
1	0.768
3	0.779
5	0.781
7	0.742
9	0.698

从表4.1的结果看出基于最大熵分类器的属性分类正确率一开始随着上下文窗口

的放大而随之提高，但当窗口放大到7的时候，正确率开始下降。这个结果说明了，一定大小的上下文窗口在一定程度上确实提高了分类正确率，但当上下文窗口大于某个阈值的时候，正确率反而开始下降。

另外，我们利用CRF分类器进行了相同的实验，结果如表4.2所示。

表4.2 基于CRF的属性分类正确率统计表

属性上下文窗口	正确率
1	0.838
3	0.885
5	0.915
7	0.936
9	0.908

表4.2的结果显示了随着上下文窗口的放大，属性分类正确率随之提高，但当窗口放大到9的时候，正确率开始下降。该结果再次证明了一定大小的上下文窗口在一定程度上确实提高了分类正确率，但当上下文窗口大于某个阈值的时候，正确率反而开始下降的结论。图4.1清晰地给出了基于CRF分类器的上下文窗口与分类正确率之间的关系。

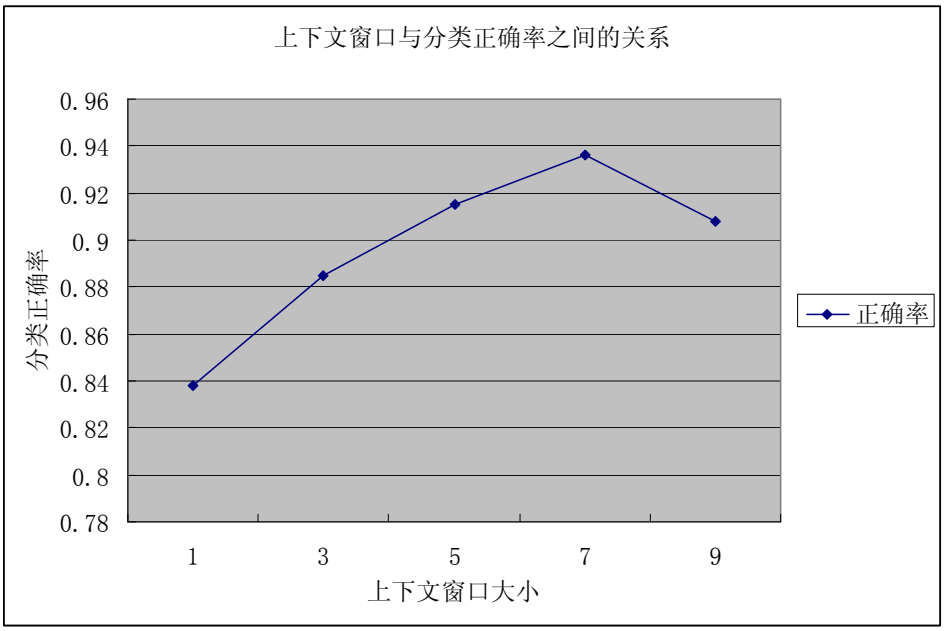

图4.1 基于CRF分类器的上下文窗口与分类正确率之间的关系图

通过表4.1和表4.2的结果对比，我们不难看出在性能上CRF分类器比最大熵分类

器要好许多，所以在后续的实验中，包括半监督学习实验，我们均以CRF分类器作为机器学习工具。

（2）属性描述词、POS及其上下文特征

本次实验在前面实验的基础上再引入属性描述词对应的POS信息作为新特征，通过两者的上下文窗口的不同取值，找出最理想的特征组合。通过前面的实验，我们认为上下文窗口不能过大，所以本次实验中，我们选用最大的上下文窗口为7。同样针对1558个属性词建立训练集，然后进行10折交叉验证，分类器选用CRF++。结果如表4.3所示。

表4.3 基于属性描述词、POS及其不同上下文窗口的分类正确率统计表

属性上下文窗口	POS 上下文窗口	正确率
1	1	0.834
1	3	0.855
1	5	0.857
1	7	0.854
3	1	0.881
3	3	0.901
3	5	0.908
3	7	0.908
5	1	0.834
5	3	0.874
5	5	0.939
5	7	0.943
7	1	0.880
7	3	0.893
7	5	0.968
7	7	0.967

从表4.3的结果中我们可以发现，在相同属性上下文窗口条件，随着POS的上下文窗口放大，分类正确率基本呈现出先同步快速上升，后基本不变甚至有所下降的趋势，

如图4.2所示。对比表4.2的实验结果，说明在一定程度上，基于一定窗口大小的POS信息确实提高了属性分类的性能。

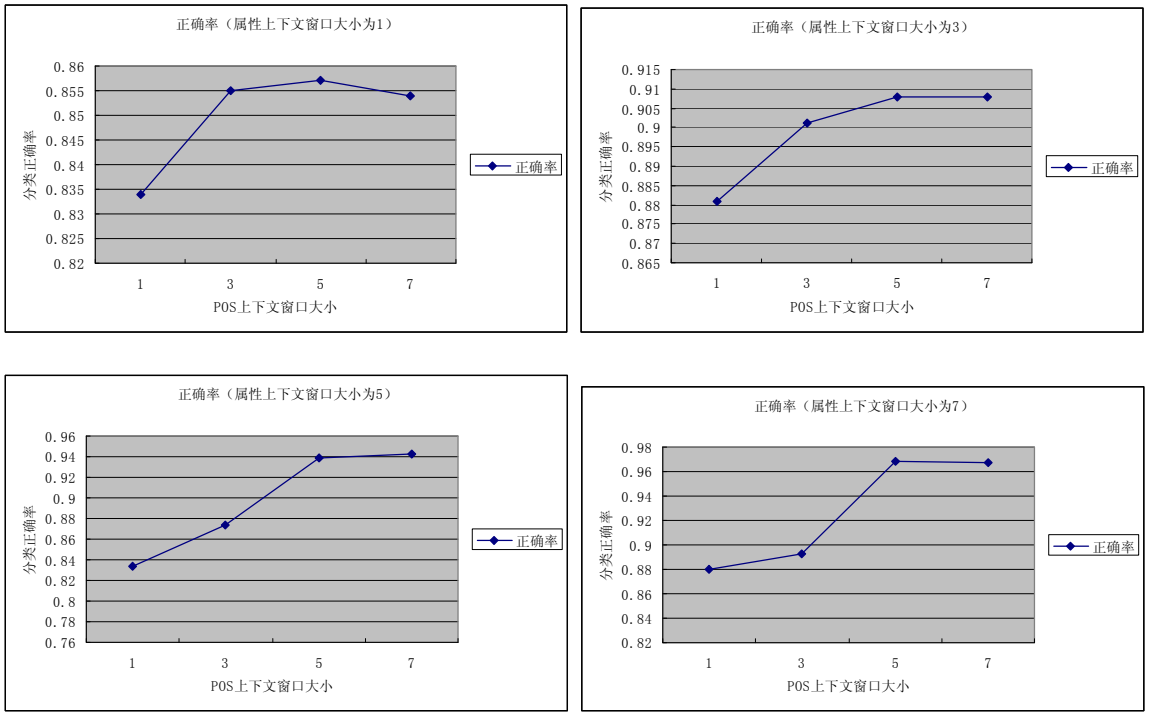

图4.2 不同属性、POS上下文窗口大小与分类正确率之间的关系图

对比表4.2和表4.3，我们还可以发现，无论什么属性上下文窗口条件下，当引入较小窗口的上下文POS信息后，整体分类性能都没有提高，往往反而有所下降。这进一步说明，只有一定程度大小的上下文信息才能提升整体的分类性能。

通过观察表4.3，我们发现当属性上下文窗口为7，POS上下文窗口为5时，性能达到最优，因此在后面的半监督学习实验中，我们采用这样的特征组合生成1558个属性训练语料。

4.3基于半监督学习的属性分类研究

随着统计自然语言处理技术的发展和普及，一个长期困扰该领域乃至机器学习研究的关键问题是语料库特别是标注实例的创建和收集成本。相比传统粗粒度情感分析，在细粒度情感分析中情感语料的标注工作更加费力、费时。本章节探索利用半监督学习方法实现细粒度情感语料的自动扩展，以减少对标注语料的依赖。

半监督学习（semi-supervised learning），有时候也叫弱指导学习（weakly-supervised

learning），就是能利用大量的未标记实例来辅助对少量有标记实例的学习方法。本章提出了面向自举学习的分层种子选取策略，根据统计学中的抽样原理，先将所有实例按属性类别分为若干个不同的级层，然后再分别从每个级层中按该层的实例数量占所有实例总数的比例来抽取样本。另外，我们也把这种分层思想应用到自举过程的每一步迭代中，即在扩展训练集时也尽可能地按一定的比例进行，避免训练集中的实例被少数类型的实例所控制。最后，我们也探讨了自举迭代的终止条件，提出了利用绝对熵值或差分熵值的方法来判断迭代是否可以停止，试图克服自举过程漫长的问题。

4.3.1 半监督学习方法

半监督学习的方法主要有以下几类[108]：生成式模型（Generative models），协同训练（Co-training），基于图的方法和自举学习（Bootstrapping）等四种，有时候也把直推方法归入半监督学习。生成式模型[109][110]利用大量未标记数据来帮助分类器建立高斯混和模型的各个成分，并采用EM算法来进行标记估计和模型参数估计。为了避免一个分类器强化自己的错误，协同训练[111][112]要求在两个独立的并且包含足够信息的视角上，建立两个分类器，两个分类器每次互相标记一部分置信度高数据给对方，然后重新训练，迭代到没有更多合适的未标记数据加入。直推方法（Transductive SVM，TSVM）[113][114]是通过加入约束项使得未标记数据落在Margin之外，即使得分类的超平面避开数据密度高的区域。基于图的方法[115]通过相似度度量将标记和未标记数据联系起来，同时假设相似结点的标记相近，并根据相似度的大小将标记传播到邻近结点。自举学习[116]使用一个分类器利用已有训练样本建立的模型对未标记样本进行分类，选出置信度高的样本加入训练集中重新训练，迭代这个过程。总之，半监督学习方法在信息抽取、文本分类等方面具有很大的潜力，因为它可以大大减少学习过程中所需要的标注语料库规模，不过其存在的主要问题是如何进行初始种子集的选取、如何控制迭代过程中的噪音干扰以及进一步探索新的半监督学习方法。

自举学习主要针对以下一组问题：给定了一个较小的标注数据集和一个较大的未标注数据集，要求产生出一个较高性能的分类器。在统计语言学领域，由于我们面临的是大量的未标注的自然语言数据，而标注的数据只占很小一部分，因此自举学习自然而然成为目前自然语言处理领域的一种非常热门的方法。因此本系统主要尝试通过

自举方法来实现半监督学习，同时引入分层抽样模型，优化初始种子集，采用不同的训练集样本扩展算法，最后得出结果对比。

4.3.2 基于分层抽样的自举属性分类方法

本节首先介绍统计学中的分层抽样模型，然后描述了基本的自举属性分类方法，接着详细讨论了如何在自举属性分类采用分层采样策略来选取初始种子集，最后是对该方法的实验结果进行分析。

4.3.2.1 分层抽样模型

分层抽样是统计学中经常使用的一种抽样调查方法[117]，目的是为了减少不同采样单位之间的差异性。其主要思想是：先将总体的采样单位按某种特征分为若干级次（或称为“层”），然后再从每一层内进行单纯随机抽样，组成一个样本。分层抽样可以提高总体指标估计值的准确率，它可以将一个内部变异很大的总体分成一些内部变异较小的层。每一层内个体变异越小越好，而层间变异则越大越好。分层抽样比单纯随机抽样所得到的结果准确性更高，组织管理更方便，而且它能保证总体中每一层都有个体被抽到。这样除了能估计总体的参数值，还可以分别估计各个层内的情况，因此分层抽样技术常被采用。

分层抽样的特点是：由于通过划类分层，增大了各类型采样单位内的共同性，容易抽出具有代表性的调查样本。该方法适用于总体情况复杂，各单位之间差异较大，单位较多的情况。

各层样本数的确定方法主要有3种方法：

①分层定比法。即各层样本数与该层总体数的比值相等。例如，样本大小n=50，总体N=500，则n/N=0.1 即为样本比例，每层均按这个比例确定该层样本数。

②奈曼法。即各层应抽样本数与该层总体数及其标准差的积成正比。

③非比例分配法。当某个层次包含的个案数在总体中所占比例太小时，为使该层的特征在样本中得到足够的反映，可人为地适当增加该层样本数在总体样本中的比例，但这样做会增加推论的复杂性。

总体中赖以进行分层的变量为分层变量，理想的分层变量是调查中要加以测量的

变量或与其高度相关的变量。分层的原则是增加层内的同质性和层间的异质性。在社会统计中，常见的分层变量有性别、年龄、教育、职业等。分层随机抽样在实际抽样调查中广泛使用，在同样样本容量的情况下，它比单纯随机抽样的准确率高，此外管理方便，费用少，效率高。

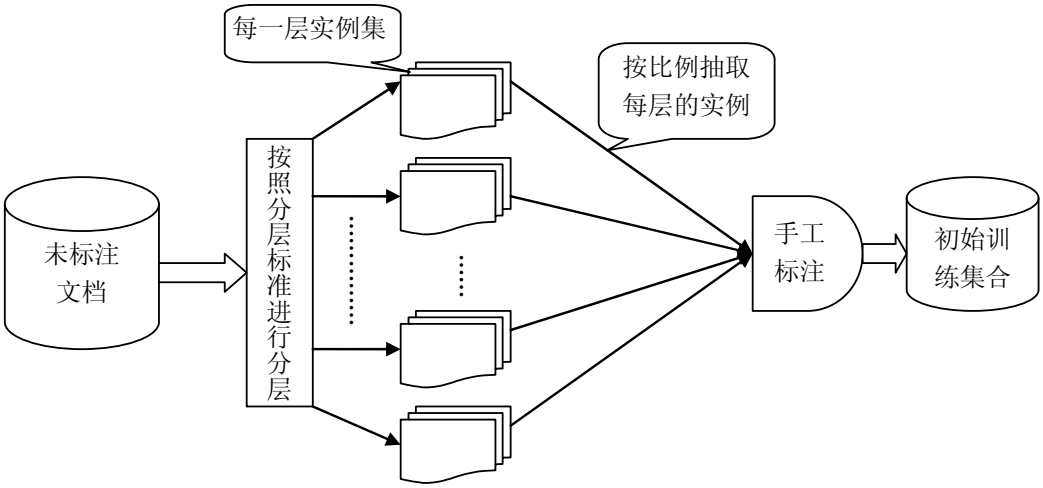

图 4.3 使用分层策略进行初始数据集选择过程

如图4.3所示，使用分层策略进行初始种子集选择的基本思想如下：在未标注文档中，我们首先对所有实例按照一定标准进行分层，然后再在每层中按比例选择实例组成初始数据集，进行手工标注，对于没有被抽取到的实例类别，我们还需要再进行添加，保证每个类别的实例至少在初始数据集中出现一次。通过这种选择方法，一方面保证了所选择的实例具有较高的代表性，另一方面也考虑了各个类别在初始种子集上的数据分布均衡。

4.3.2.2 自举属性分类算法

基于半监督的自举属性分类系统主要由两个部分组成：一是一个基本的有监督属性分类模块，如前文中所讨论的基于特征向量的属性分类方法；二是建立在有监督学习算法上的自举模块。

本文采用最基本的自举学习方法来研究半监督学习过程中的若干关键问题，如初始种子集的产生、训练数据集的扩展以及迭代终止条件等。图4.4为自举学习算法的基本流程。

Algorithm 自举学习算法

Require: 标注种子集 L
Require: 未标注数据集 U
Require: 每次迭代加入的数据集大小 **S**
Repeat
在L上训练有监督分类器,得到分类模型
用分类模型对U进行预测分类
在U中找出最多S个分类器具有最高预测值的实例
将这些实例加入L中
Until: 所有实例均已加入到训练集中或迭代终止条件已满足

图 4.4 自举学习算法基本流程

自举学习算法首先从语料库中选择具有代表性的一部分关系实例进行标注，这一部分数据集称为标注种子集L，其余大量的实例组成未标注数据集U，种子集最常用的选择方法是随机选择，以使种子集中的实例在语料库中具有一定的代表性。然后把标注种子集作为训练语料库，训练一个有监督分类器（如CRF、SVM分类器）并得到一个分类模型。再用该分类模型对未标注数据集U进行预测，找出最可靠的S个实例加入到标注数据集中，继续该过程直到所有的未标注数据均已加入或终止条件已满足。

4.3.2.3 分层策略的自举属性分类

作为半监督学习方法中的一种，自举学习方法在很多领域具有一定的积极作用，但是其中仍有一些关键问题没有被很好地研究，下面分三个方面结合自举属性分类任务进行详细讨论。

（一）初始种子集的选取

在基于半监督的属性分类中，初始训练集合的选取具有非常重要的作用，一个合适的初始种子集才有可能使得自举过程成功地执行下去，反之自举学习则会使性能越来越低。初始种子集的选取主要涉及到两方面的问题：

(1)初始种子集的规模。即选取多少实例较合适，这个问题不但和半监督学习方法本身有关，还同应用领域有着密切关系。就属性分类而言，属性类型可以根据对象的不同方面划分成多个类型。如果初始种子集数量选择太少，分类器就无

法较好地捕获每个类型的特征，因而产生的初始训练模型较差，导致最终结果不太理想；初始种子集数量也不能选择太多，因为标注一个属性分类实例需花费一定的时间和人力，种子集中的属性分类实例过多会导致标注工作量的猛增，从而违背了半监督学习方法的原则。基于此，同时结合初始试验结果，本文选择了 200 个实例作为初始种子集的大小。

(2) 初始种子集的选择。目前普遍采用的语料库均存在着数据稀疏问题，即某些类别的实例特别少。在选择初始种子集时应该充分考虑到各个类别之间的均衡问题，同时应该选择那些具有代表性的实例，从而产生一个性能较好的初始种子集模型。

目前基于半监督学习方法的相关系统在选取初始训练集合时都采用了随机抽样的方法，如Zhang [118]在ACE 2003上进行半监督语义关系抽取实验时，首先从未标注文档中随机抽取100个实例作为初始种子集，然后使用自举方法进行语义关系分类，结果显示随机抽取的初始种子集在各个类别上分布极不均衡，而且抽取性能很不稳定，严重影响了半监督语义关系抽取的整体性能。

由于使用随机抽取所选择的初始训练数据不具有较高的代表性，同时在各个类别上分布极不均衡。因此本文使用了统计学中的分层抽样原理：先将总体的单位按某种标准分为若干次级（层），然后再按照比例从每一层内进行单纯随机抽样，组成一个子样本。由于通过划类分层，减小了同类型样本之间的差异性，增大了不同类型样本之间的差异性，容易抽出具有代表性的调查样本。该方法适用于总体情况复杂，各单位之间差异较大，单位较多的情况。

在使用分层策略抽取初始种子集时，首要工作是选择一种有效的分层标准，将未标注文档划分成相互之间具有较好区分度的各层。本文实验工作主要在前面章节生成的1558个属性训练语料上进行，该语料库已对各属性描述词进行了标注，而我们将这些标注文档仍然作为未标注文档看待。因此我们可以将属性类别作为分层标准，虽然在实际的应用中，由于未标注文档的属性类别信息本身并不可知，不过仍然可以把属性类别信息作为一个参照标准，假设在属性类别已知的情况下，基于分层的自举算法能取得什么样的性能。

因此，在本文中我们比较了两种初始种子集的选取方法：

（1） 随机抽取。参考 Zhang [118]的方法，从整个训练集（含有 1558 个实例）中

随机抽取出 200 个实例作为初始种子集。

（2） 按实例的类别进行分层选取。虽然在实际应用中，我们事先无法获得未标注实例的属性类别信息，但是为了说明基于属性类型的分层策略对自举属性分类性能的影响，我们仍将实例的类别作为划分层次的标准。首先根据某一类实例占总体实例的比例确定每一层次的选取样本数，然后在每一层内采用单纯随机抽样，同时保证每一层的实例至少要有一个被抽取。

（二）训练数据集的扩展

自举算法的另一个关键问题是在训练数据集的扩展中可能会加入一些错误标注的实例。事实上，由于分类器对未标注的实例进行预测时，不可能做到100%准确，因而必定会有少量标注错误的实例加入。所以在每一次的迭代过程中，这些错误信息都会累加。这样，随着训练数据集的不断扩展，分类器的性能可能会呈现一种下降趋势。同时，由于实例类型分布不均衡，某些类型的实例数较多，在迭代的过程中这些类型的实例被不断加入，因而分类器逐渐会被这几个甚至一个类型的实例所控制。本文采用两个方法尝试解决这些问题。

（1） 尽量提高每一步中加入实例的分类准确率。由于我们采用的是 CRF++分类，因此使用与 Zhang[118]类似的方法来计算实例的分类准确率，因为通过设置 CRF++相关的参数同样可以获得不同分类的概率值。在每一次的迭代过程后，分类器对每一个未标注实例都会给出它属于某一类别的概率值。据此可计算出它的熵值 H，并选择 S 个熵值最小的未标注实例加入到训练集中，再进行下一轮的迭代。按照熵值的定义，熵值越小，确定性越好，实例的分类可信度也就越高，其被正确分类的可能性也越大，因而将其加入到训练集中进入下一轮的迭代，所带来的风险也就较小。

实例的熵值H的计算公式如下：

$$H = -\sum_{i}^{n} p_i \log p_i \qquad \text{（公式4.4）}$$

其中n表示了属性类别的个数，而 p_i 代表了当前实例被分到第i个类的概率。

（2） 按比例加入各类型的实例。在每次迭代时，由于分类可靠性高的实例往往集中于少数类别，从而使得训练集被某些类别的实例所控制。为了解决这个问题，我们将分层的思想引入到训练集的扩展中，在每次扩充训练集时，首先选出置信度较高的 m 个实例（m >=50），然后使用分层的方法从中再选择 50 个实例加入到训练集

中进行下一轮迭代。这样做的出发点是为了让更多类别的实例也能加入到训练集中，从而使训练集中各个类型的实例数量比较均衡，有利于后续的迭代过程能取得较好的性能。

（三）迭代终止条件

在半监督属性分类中，当分类器的性能往往一开始呈现增长趋势，到达一定阶段后趋于稳定，最后由于迭代过程中标错的实例不断加入，性能又不断下降，这样的过程可能会反复多次。为了取得最高的自举分类性能，最简单的方法是在迭代过程中将所有的未标注实例全部加入到训练集中，再从中找出最好的性能。但是，这样会导致自举时间变长，而且在实际使用中，未标注实例的数量往往是惊人的，要全部加入也是不现实的。

我们比较了两种方法来判断迭代的终止条件是否满足，一种是根据熵值的变化趋势；一种是根据熵值的绝对值大小。

（1）根据熵值的变化趋势。熵值表示分类器对实例进行分类的置信度。首先定义一个综合熵值 H_i:

$$H_i = H_{avg} + H_{\min} \qquad \text{（公式 4.5）}$$

其中 H_{avg} 是加入到训练集中实例的平均熵值，$H_{\min}$ 是其中的最小熵值，因此综合熵不仅考虑了实例的平均熵，也考虑了其中的最小熵（即分类最确定的实例）。然后再定义迭代终止条件：

$$H_{i+1} - H_i \le p \qquad \text{（公式 4.6）}$$

其中p值是一个经验值，可以用测试集的方法进行估计。即当加入到训练集中的实例综合熵趋于稳定时（即前后两次迭代过程的综合熵之差小于p），迭代过程就终止，此时取得的性能应该是最好的。这样做的前提是我们假设：自举刚开始的时候，由于训练集较小但实例标注较准确，分类器对实例分类结果的可信度忽高忽低，因此综合熵值的波动比较明显。随着训练集的增加，可信度逐步稳定，综合熵值因此趋于稳定。随后，由于大量错误标注的实例进入训练集，综合熵值也会再次波动。

（2）根据熵值的绝对值。在自举迭代过程中，加入训练集的实例的平均熵值的大小呈现出波浪形的变化规律。但是我们知道，熵值越小，分类可信度越高，因此我们假设：当熵值低于某一阈值时，自举过程已取得了最好性能。这时的终止条件可表述为：

$$H_i <= p \qquad \text{（公式 4.7）}$$

此时H_i为每次迭代加入到训练集中的实例的熵值，p的值可通过测试集的方法进行估算。

4.3.2.4 实验结果和分析

自举属性分类实验主要是在前面生成的含有1558个属性词的数据集基础上，特征类别主要有7类，分别为：“设施”，“服务”，“环境”，“价格”，“餐饮”，“交通”，“整体”。表4.4列出了训练集和测试集中各类的实例数量及其占训练集和测试集总数的百分比。本章主要针对分类正确性进行比较，所以本实验过程中只计算正确率。

表 4.4 训练集和测试集中的实例数量及其百分比

属性类别	训练集	%	测试集	%
设施	322	25.84	81	25.96
服务	226	18.14	56	17.95
环境	325	26.08	82	26.28
价格	79	6.34	19	6.09
餐饮	85	6.82	21	6.73
交通	98	7.87	24	7.69
整体	111	8.91	29	9.29
	1246	100.00	312	100.00

实验采用类似于有监督学习中的5折交叉验证法，在总的实例中划分出的训练集有1246个实例，测试集有312个实例。从表中可以清楚地看出，各类之间的实例数量相差都很大，分布极不均衡。例如，属性类“环境”具有最多的实例数量，其次是“设施”，而“价格”则最少。

自举学习用的初始种子集**L**取自训练集，训练集中剩余部分作为未标注实例**U**。每次迭代后，将从L中学习到的分类器模型应用于测试集进行分类预测，并计算出相应的分类性能作为该次迭代的测试性能。

本文在进行半监督属性分类时使用条件随机场的CRF++作为分类器，因为它支持多类分类任务，而且更重要的是它还可以提供每个实例属于各类的概率值，便于计算

每个实例的信息熵值（也就是衡量每个实例可信度的指标）。当需要获得分类预测的所有概率，只需要通过执行命令：“crf_test -v2 -m model test.data”来实现。部分结果如图4.5所示。

图 4.5 带有概率值的预测结果实例图

图4.5显示了两个属性描述词及其POS信息的上下文特征信息，以及属性分类结果的所有概率值数据。我们不难看出，属性描述词“服务态度”的属性分类最大可能是“服务”，因为其概率为0.966485，而“地理位置”的属性分类最大可能是“交通”，其概率为0.902684。通过这些概率值，我们就可以方便的进行信息熵值的求解。

（一）不同初始种子集选取策略的比较

每一次实验的初始种子集为 200，每次迭代均加入 50 个置信度最高的实例，直到所有扩展训练集数据用完。首先，我们列出两种不同种子集选取策略对应的自举性能情况，其中自举性能主要是衡量自举学习方法对属性分类性能的影响。

1） 随机选取

参考 Zhang[118]的方法，我们选择每一次实验的初始种子集为 200，每次迭代均加入 50 个置信度最高的实例，直到所有训练数据集用完。总共进行 20 次抽样实验，分别记录初始种子集的性能以及最高性能，并将实验结果列入表 4.5 中。表 4.5 统计了 20 次抽样中，种子集的初始性能以及迭代过程中能取得的最高性能。从表中我们发现，尽管每次测试的结果都有一定的差距，但自举学习确实提高了属性分类的性能。P 值的增幅平均值达到了 18.4，这有力地说明了自举学习方法在属性分类中的有效性。

表 4.5 随机选取初始种子集的自举性能

抽样次数	初始值 P(%)	最高值 P(%)
1	41.3	64.1
2	33.0	50.9
3	46.7	66.6
4	45.5	59.9
5	44.8	63.4
6	42.6	54.1
7	41.6	54.1
8	44.8	59.2
9	39.4	57.6
10	42.6	62.5
11	44.2	54.4
12	41.9	56.4
13	37.5	60.8
14	42.9	57.0
15	41.9	62.1
16	50.0	71.1
17	41.0	66.9
18	47.4	71.1
19	45.1	62.8
20	37.5	65.7
平均值:	42.6	61.0

为了分析一次自举学习过程中的变化情况，图 4.6 显示了第 3 次随机采样中，把训练集中的所有实例逐步加入到种子集后的测试性能 P 的变化趋势。由于训练集中的实例有 1246 个，初始种子集大小为 200，每次迭代扩展 50 个，所以经过 20 步迭代之后自举过程结束。虽然每次采样后的自举结果会有差异，但迭代趋势却能说明一定的问题。

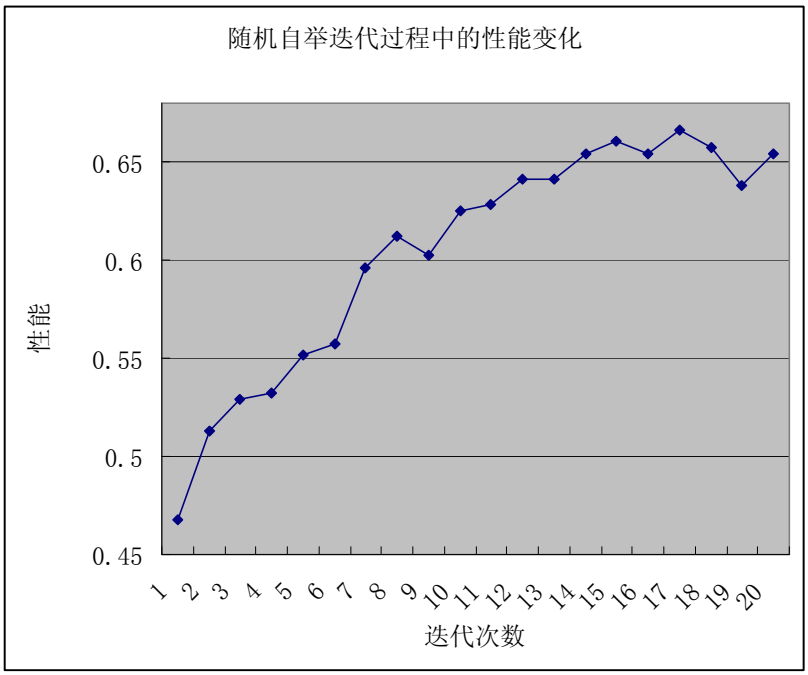

图 4.6 随机自举迭代过程中的性能变化

从图 4.6 中我们可以看出，在迭代刚开始时，自举性能逐步上升。此时，扩展训练集的语料比较大，扩展加入到临时训练集的 50 个语料的正确率较高，于是自举性能逐步上升。但是，随着迭代次数的变多，扩展训练集剩余的语料渐渐变少，但为了保持加入扩展 50 个，所以一些不正确的语料也被加入，不过在临时训练集中还是正确的语料占了多数。所以到了第 8 次迭代时，自举性能第一次出现了下降，但随后又经历上升下降的波动，说明性能已经达到了最高，开始处于平稳波动状态。

2） 按实例类别进行分层选择

通过上面基于 Zhang[118]的种子集随机选取实验的结果分析，发现初始性能普遍不高，我们提出了按实例类别进行种子集的分层选取方案。初始种子集分层选取就是按照各特征在扩展训练集中数目的比例进行选取，我们选取初始总量为 200 个，当某一类别比例过小时，在选取时确保该类至少有一个实例被选中，表 4.6 显示了在分层策略下各特征在初始种子集中的数目。

表 4.6 分层选取种子集中的特征分布

特征	个数
设施	51
服务	36
环境	52
价格	13
餐饮	14
交通	16
整体	18
总数	200

我们以分层策略生成种子集，每次迭代时选取熵值最小的 50 个实例加入训练集，直到所有数据用完，进行了 20 次抽样试验，每次实验进行 20 次迭代，统计出每次抽样中种子集的初始性能和在迭代中能获得的最高性能，结果如表 4.7 所示。从表中我们看出，相比表 4.5 随机策略生成种子集的性能，在分层策略下，种子集的初始性能有了明显的改善。

表 4.7 分层选取种子集的自举性能

抽样次数	初始值 P(%)	最高值 P(%)
1	69.8	69.8
2	48.3	66.0
3	69.8	69.8
4	68.5	68.5
5	35.9	42.1
6	34.6	34.6
7	61.5	61.5
8	61.5	61.5
9	45.9	62.0
10	60.2	60.2

11	69.8	70.5
12	60.2	65.7
13	25.9	47.1
14	33.6	41.3
15	49.3	49.3
16	36.7	48.0
17	61.5	61.5
18	25.9	45.1
19	45.3	48.3
20	57.3	57.3
平均值:	51.1	56.5

我们也发现在分层策略下，极端情况变得更加极端。有的抽样初始性能仅为25.9%，有的却高达 69.8%。推测原因是：

在分层的条件下，各特征只能在各自类别的范围内进行选取特定的数目，选取的范围和数目都受到限制。而我们的种子集的大小本身就小，只有 200 个，在这样的情况下如果刚好选取了代表性的特征，就会出现比较理想的性能；如果，刚好选择了那些不具代表性的特征，则会出现极端差的现象。这个问题可以通过扩大初始集规模数得到一定的解决。同时，我们发现大部分初始性能都在 60%左右。因此，在分层选取种子集的策略下的效果应该是要比随机选取好的。

同时也可以预测，如果我们用分层策略去生成一个数目更加大的种子集，那么效果会更加理想。在后面“初始种子集的大小对自举性能的影响”章节中就可以看到具体实验效果。

（二）训练数据集的扩展方法比较

在前面的实验中，每一轮迭代扩展训练集的时候均采用单纯信息熵值进行扩展，即在未标注的所有实例中挑选出熵值最小的前 50 个实例加入训练集中。但是从前面的分析中我们知道，由于训练实例数量较少，分类器往往会被某几类实例所控制，因而熵值最小的前 50 个实例通常集中在个别类型上，从而使训练集中的实例分布极不均衡。为了缓解这个问题，本文采用分层策略来扩展训练集，具体步骤为：

- 首先把每轮迭代所加入的实例数仍然固定在 50 个，但把挑选的范围扩充到前 n（$n \geq 50$）个实例中，即所有未标注的前 n 个实例按熵值递增进行排序；
- 由于前 n 个实例中的类型分布同种子集中的类型分布也不匹配，因而上一步按照分层策略抽取出的实例总数通常小于 50，数量不足部分再按照熵值从 n 个实例中进行抽取。

特别地，当 n=50 时，分层扩展方法退化为单纯按熵值扩展方法。本实验初始种子集的大小为 200，考虑了随机选取初始种子集和分层选取初始种子集两种情况，分别来比较当步长为 50、100、150、200、250、300、350、400、450、500 时，每 20 次抽样平均的初始种子集性能和平均最高值性能。

实验结果如图 4.7 所示，横坐标表示扩展步长，纵坐标表示在相应的步长下，进行 20 次抽样实验后的平均性能。从图 4.7 中，我们可以看到两方面信息：不同种子集按信息熵值和分层两种方式扩展的性能变化；分层扩展的性能随着扩展步长的变化。当步长等于 50 时，实际变成了按熵值扩展，同时还可以看出随着扩展步长的增加，分层扩展方式的效果越来越明显。因此如果要比较分层扩展和熵值扩展的性能，只需要比较步长等于 50 与步长大于 50 时每个点的性能情况。

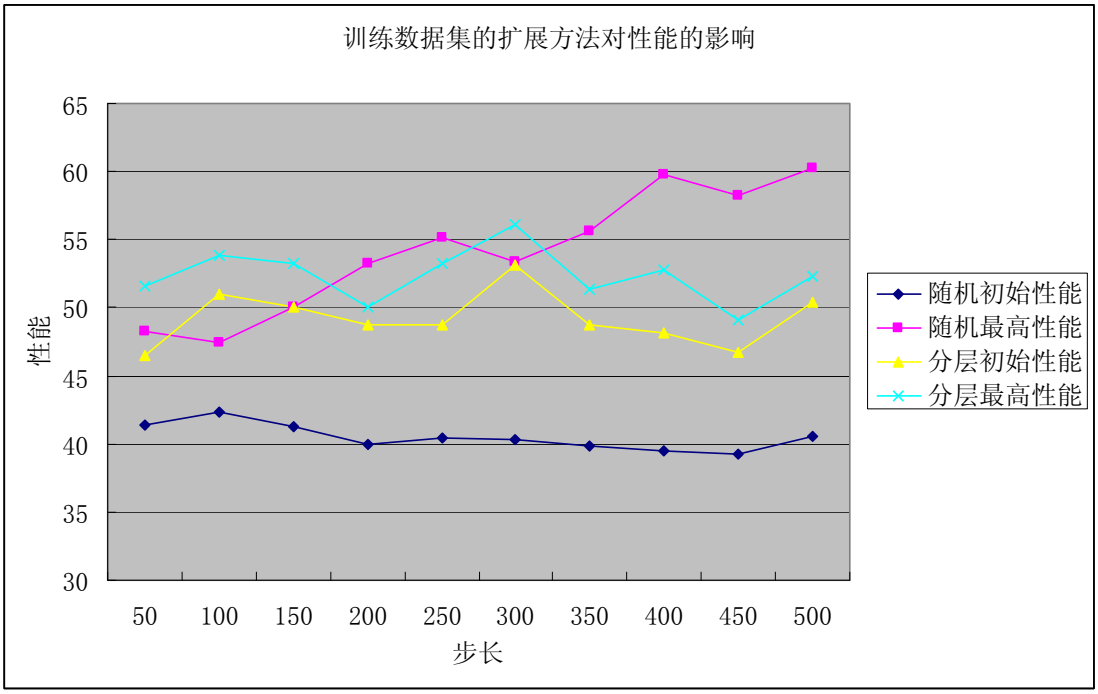

图 4.7 训练数据集的扩展方法与性能的关系图

从图 4.7 中我们还可以看出，在初始性能方面，分层种子集的平均性能要高于随机

种子集的性能。这主要是因为，分层生成种子集可以合理地分配各个特征在种子集中的数量，不会出现某些特征过多或者过少的情况，从而在初始预测时就能取得比较高的性能。随着步长的增加，我们发现分层和随机的最高性能都有所变化。随机最高性能是随着步长的增加而稳步提高，而到了步长 500 时，其实就相当于步长接近无限了。而分层的最高性能在开始阶段要比随机最高性能要大，并会波动上升，但是到了 300 之后处于稳定波动状态，分层最高性能反而小于随机最高性能了。分析原因主要是因为分层扩展虽然按照严格的比例分配特征，但与此同时，使得有部分占比例较小的特征在预测该类特征时不能非常的准确，从而导致随着迭代次数增加后，错误的语料也渐渐越来越多地加入，所以，性能不会随着步长的增加而一直增长，反而在最后趋于平稳了。

（三）迭代终止条件的比较

为了寻找合适的终止条件，首先对自举过程中熵值变化情况进行观察，然后依据绝对熵和差分熵的变化趋势来判断迭代是否终止。

1）自举过程中熵值的变化情况

我们首先了解一下在自举过程中熵值的变化趋势，图 4.8 描述了在随机采样的种子集自举过程中（分层采样数据，训练集扩展采用 Top 50 策略），加入到训练集中的前 50 个实例的平均熵、最小熵和综合熵的变化情况（其中，在 18 步测试性能取得最佳值）。从图中可以看出：

- 在迭代开始时，综合熵较高，而随着迭代次数的增加，综合熵渐渐减低，伴随有一定的波动，但在迭代快要结束的时候，综合熵又开始上升了。产生这样的原因是，刚开始迭代时临时训练集比较小，因此预测的性能比较低，熵值较高；而随着迭代的进行，临时训练集渐渐变大，所以预测的性能也会上升，因此，熵值也变小了。但迭代进行到最后时，虽然临时训练集已经变得很大，但是在迭代过程中，扩展训练集的操作也使得把越来越多的错误语料加入了，影响了预测的效果，所以，性能在最后会略微下降，使得预测的性能在达到某一峰值后，呈现出上下波动的现象，因此，在迭代的最后，熵值开始有所回升。
- 最小熵的变化趋势与综合熵基本相似，平均熵的波动相对比较小。因而，综合熵的总体变化趋势基本上由最小熵决定。

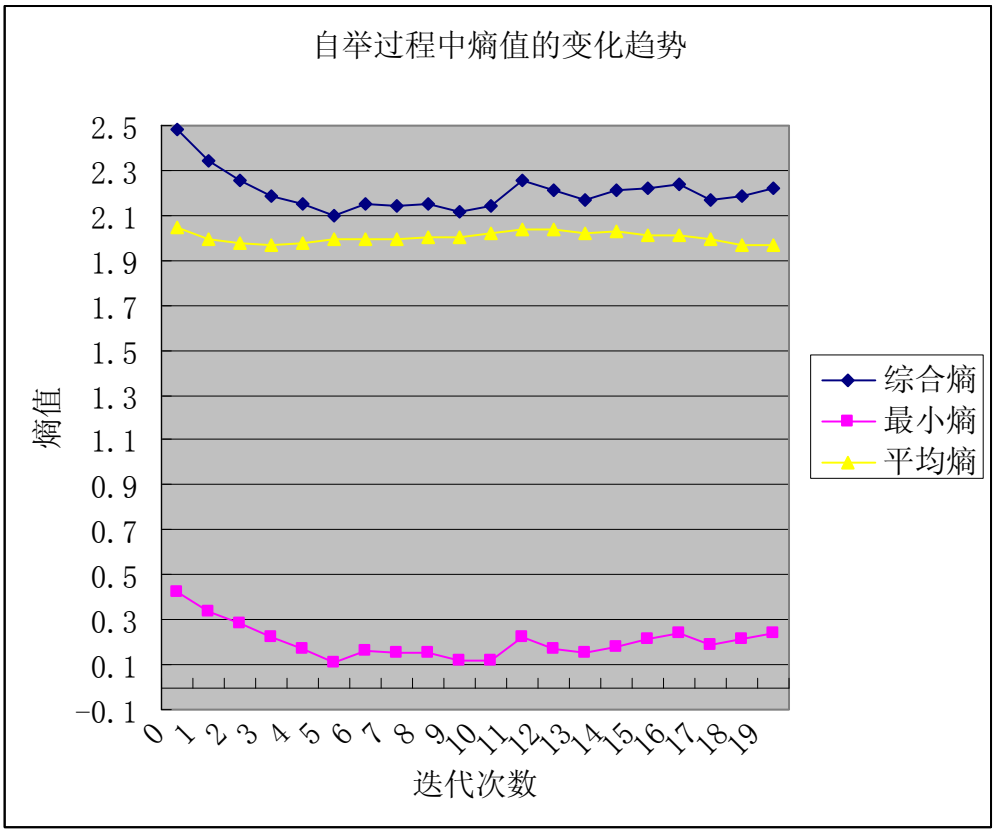

图 4.8 自举过程中熵值的变化趋势

2）基于绝对熵值的迭代终止条件

基于绝对熵值的迭代条件关键在于什么样的熵值才是最低点，由于无法用分析的方法得出，我们采用实验的方法进行确定。由于平均熵、最小熵和综合熵的变化范围不同（即最小熵<平均熵<综合熵），因此最低点的确定也不相同。

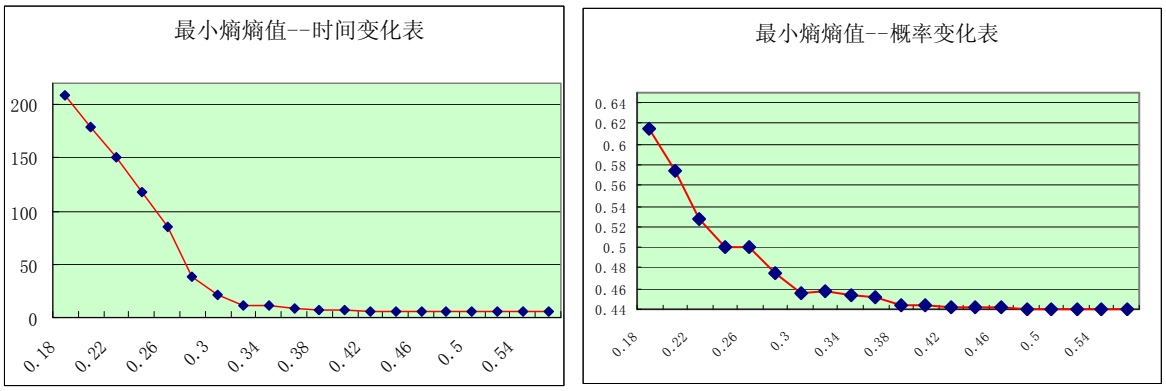

图 4.9 最小熵阈值对性能和自举时间的影响

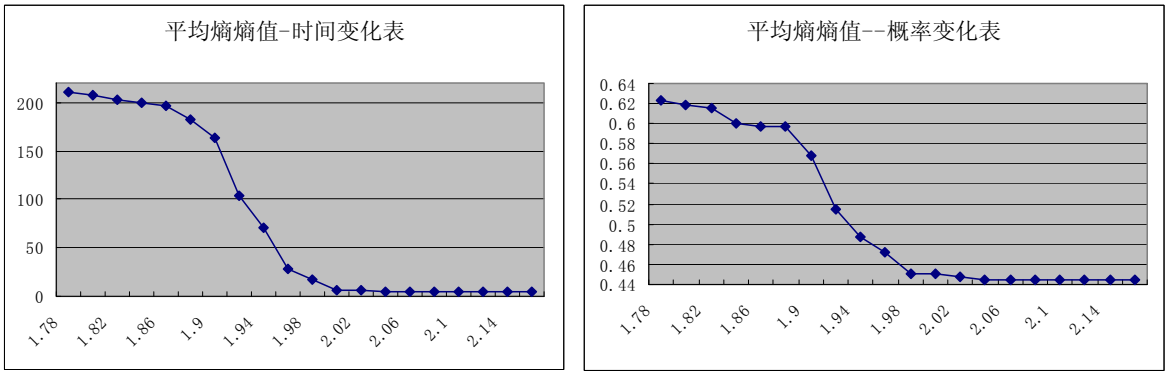

图 4.10 平均熵阈值对性能和自举时间的影响

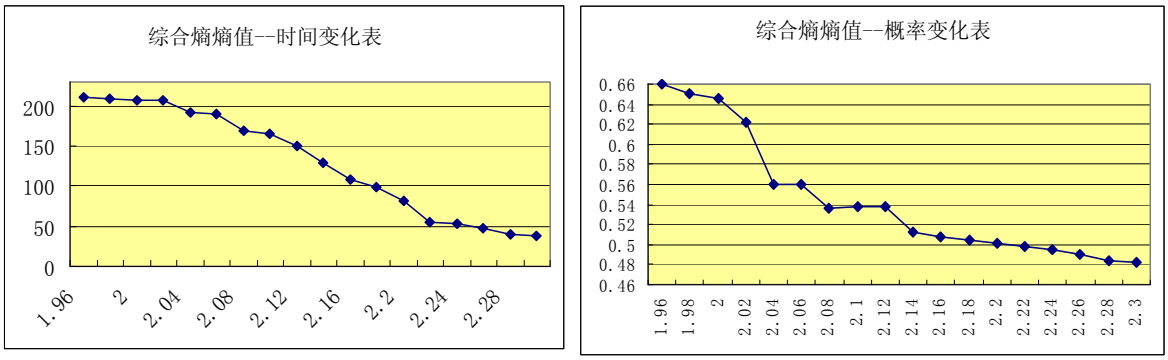

图 4.11 综合熵阈值对性能和自举时间的影响

上图 4.9-4.11，比较了采用不同绝对熵方法（最小熵、平均熵和综合熵）来判断迭代终止条件时，绝对熵阈值对自举时间（秒）和自举性能（预测的正确率）的影响（随机选取种子 200，采用步长无限扩展策略）。其中自举性能为 20 次实验的平均值，而自举时间为 20 次实验的总和。左边的图描述了该种熵的绝对熵阈值对自举时间的影响，由于自举时间对绝对熵阈值的变化很敏感，因此绝对熵阈值的变化步长为 0.02。相对应地，右边的图则描述了该种熵的绝对熵阈值变化对自举性能的影响。结合每种熵值的两张图可以看出：

- 左边图中最小熵、平均熵和综合熵的变化范围分别为 0.18~0.56、1.78~2.16 和 1.96~2.34，即熵值在这些范围的左边时，训练时间同没有终止条件是一样的（迭代终止条件退化为用完所有的可用训练集中的实例），而在右边时，训练时间接近于 0（即小于 1 分钟）。相应地，右边的图列出的是在该变化范围内所对应的 F 值。
- 随着 p 值（临界熵值，也就是图中的横坐标）的增加，自举性能缓慢地下降，而自举学习所需要的时间却显著降低，即使当自举时间不小于 1 分钟时，而性能仍然比初始性能好。这是由于，熵值越高，迭代过程越早结果，取得最佳性能的可能性越

小。同时意味着，我们能够找到一个折衷的阈值，在略微损失性能的前提下，显著降低自举学习所需要的时间。

- 训练时间随绝对熵阈值的变化趋势中，相对而言，随最小熵阈值的变化最快。但是最小熵阈值的自举时间不会在最后趋向于稳定，因此，最后随着性能的提升，自举时间也会一直快速的增加，这意味着，采用最小熵阈值来判断迭代终止条件，将难以清晰地进行判断。而平均熵和综合熵的自举时间和性能在最后将趋向于稳定，所有可以利用这两种熵值进行迭代终止条件的判定。

- 对于平均熵而言，熵值为 1.88 应该是个较好的选择，该点的时间为 183 秒，性能为 0.597。因为此该点较熵值 1.86 相比，性能没有下降，但是自举时间却节省了 14 秒(7.1%)。对于综合熵而言，熵值为 1.96 应该是个较好的选择，该点的时间为 211 秒，性能为 0.66。因为此该点较熵值 2.02（时间为 207 秒）相比，性能提高了 0.038(6.1%)，但是自举时间却只增加了 4 秒(1.9%)。但是由于平均熵达到该阈值（1.88）所花的时间（183 秒）比综合熵（211 秒）少了 13.3%，而性能只少了 9.5%，这说明采用平均熵阈值的迭代停止条件，可以用较小的性能损失作为代价，换取自举时间的大幅下降。

3）基于差分熵值的迭代终止条件

该实验条件设置上和“基于绝对熵的迭代终止条件”相同，即进行 20 次抽样，时间取 20 次的总和，概率取 20 次的平均。该实验分别记录差分熵与时间和概率的关系，我们根据之前“基于绝对熵终止条件”实验中迭代前后的熵值差计算得出，差分熵的范围为：0~0.95，差分熵的步长为 0.005。

图 4.12 记录差分熵与时间的关系，横坐标表示差分熵的数值，纵坐标表示在 20 次抽样迭代过程中达到该差分熵值所需要的总时间。

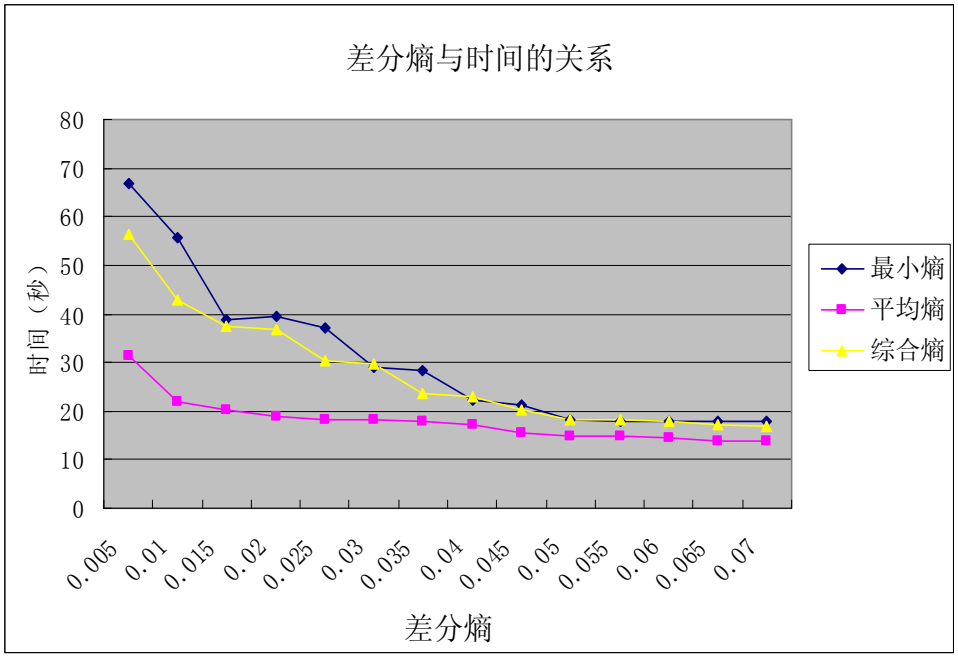

图 4.12 差分熵与时间的关系

从图中可以发现，达到差分熵值越小所用的时间越长，说明在迭代刚开始的时候前后熵的差值都是比较大的，随着迭代的进行差值越来越小。最小熵时间随差分熵的变化最明显，而平均熵相对而言要平稳一些。

另外，图 4.13 记录了差分熵与性能的关系，横坐标表示差分熵的数值，纵坐标表示在 20 次抽样迭代过程中达到该差分熵值的平均性能。

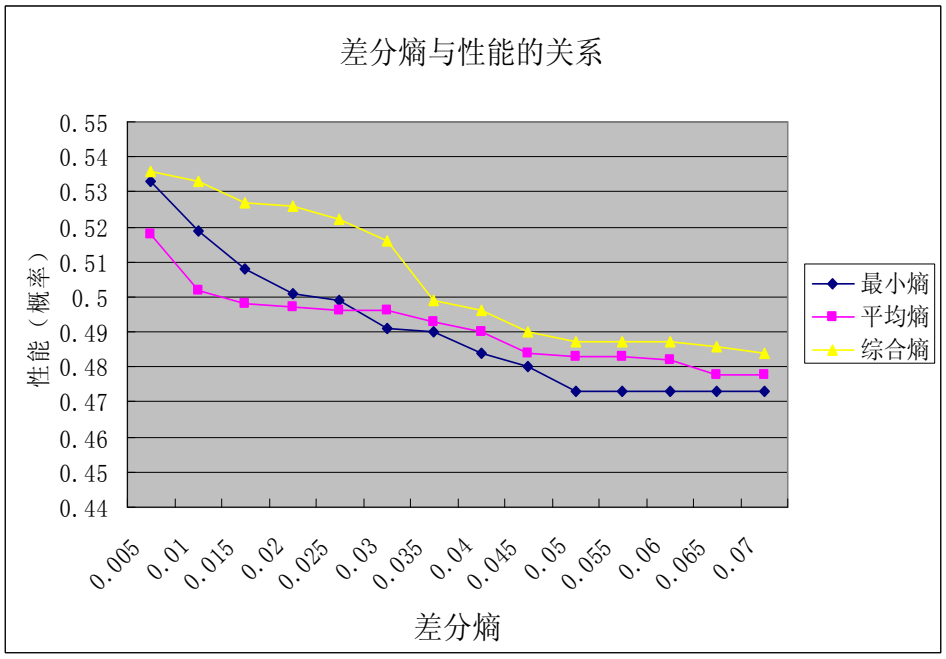

图 4.13 差分熵与性能的关系

从图中可以发现，自举性能随着差分熵的增大而减小。这说明，在迭代过程中当前后熵的差值比较大时，性能也不高；而随着熵渐渐趋于稳定，差分熵慢慢变小，性能也不断提高，最后达到一个最高值。

结合以上两张图，对于综合熵而言，我们发现当差分熵为 0.025 时，所花费的时间只有在迭代取得最高性能时的一半，然而性能却只降低了 0.014。所以以 0.025 作为综合熵的差分迭代终止条件较为合理，达到了花费较小时间代价的同时，取得较高性能的目标。

（四）初始种子集的大小对自举性能的影响

为了考察初始种子集大小对自举性能的影响，我们在不同的种子集上进行了自举实验。种子集规模从 100 开始，每次增加 100，并采用随机和分层两种策略选取初始种子集，在训练集扩展时采用 Top 50，步长无限的策略。由于随机种子集数量的增加，随机抽取的各个种子集之间的方差也不断缩小，因此在 100~500 之间安排随机抽样 20 次，在 600~1000 之间时抽样 10 次。

采用随机和分层两种策略选取初始种子集，计算每次实验的平均初始性能和最高性能。实验结果如图 4.14 所示。

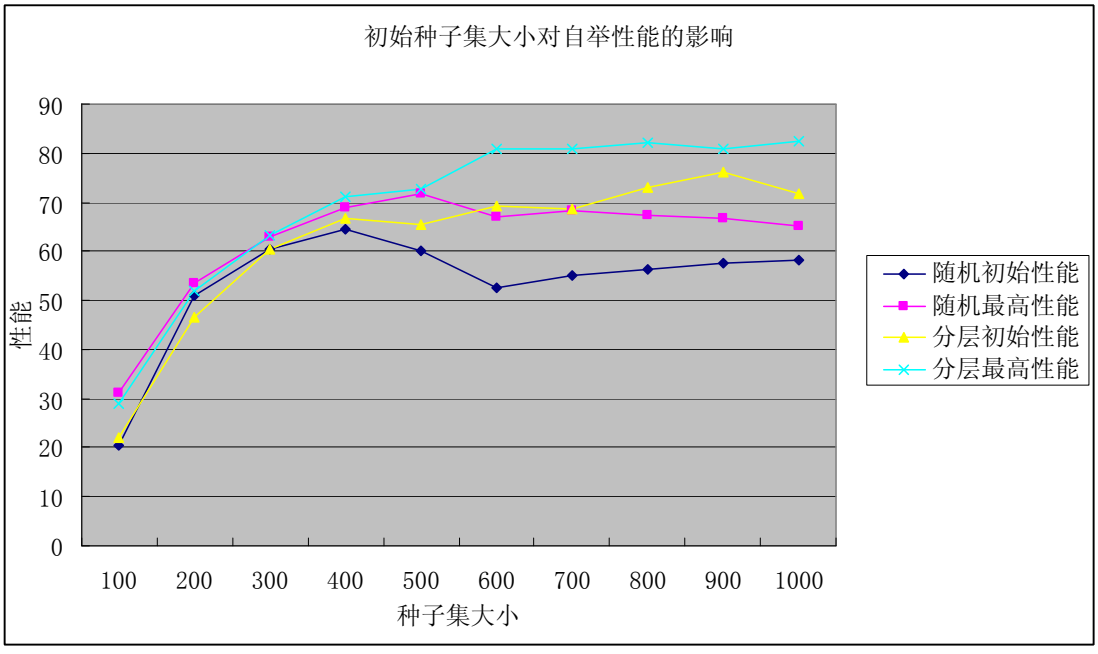

图 4.14 初始种子集大小对自举性能的影响

图 4.14 显示了自举性能随初始种子集大小的变化情况，从图中可以看出：

- 随着初始种子集数量的增加，自举学习的初始性能也呈单调递增。特别地，当初始种子集规模较小时（如 100~400），初始性能的增加较明显，随后增长趋势放慢。一般认为，机器学习的性能和训练集的大小呈对数线性关系。这是由于，当训练集过小时，训练所得到的分类模型不稳定，这时候加入的训练实例对模型的帮助很大；而当训练集本身较大时，分类模型已经较稳定，此时再加入的实例对它帮助就很小；
- 当初始种子集数量增加时，自举所获得的性能提升呈现出小幅上下波动，但就其总体趋势而言，性能提升越来越小。例如，种子集从 100 增加到 500 的过程中，折现的斜率慢慢变小。可以认为，随着种子集规模的扩大，分类模型越来越稳定，对自举所添加的实例控制能力越强，因而未标注实例的帮助作用越小。
- 从初始种子集选取方式的角度，我们也可以看出，不管是初始性能还是最高性能，分层选取种子集的情况在开始的时候性能不如随机选取，但是都会后来居上。原因是，在种子集较小的情况下，种子集分层生成会导致部分特征比例不够，这样去预测这类特征时性能自然不高。而随着种子集的增大，临时训练集中的特征严格按照各自比例分配，使得某些比例小的特征也达到可以高准确预测的需求后，预测性能稳步提高了。

4.4情感计算研究

本章提出的情感计算主要针对评论中所涉及的情感词进行情感倾向值确定，然后根据情感词所在位置与其相关的评价对象属性进行关联，最后对所有评论进行基于属性分类的情感倾向值汇总统计。针对有些评论中存在情感词缺少对象属性的情况，我们利用 PMI 方法来确定评价对象属性类与情感词之间的关联概率，使情感汇总计算更为合理有效。

4.4.1 属性类别与情感词的关联挖掘

有时在情感表达过程中存在评价对象属性缺失的情况，如果在情感汇总过程中丢弃那些没有对象属性的情感信息，那将是非常大的损失。例如，有一酒店评论，“太漂亮了！当我一下出租车，就仿佛进入了城堡。”，在这句评论中，“太漂亮了！”属于

情感句，但缺少了评价对象的属性描述词，所以按照传统方法，我们可能会丢失这个情感信息，这样的话，对整个情感计算是一种非常大的浪费。所以我们将通过挖掘对象属性类与情感词之间的关联性，实现对缺失对象属性的情感信息进行合理属性类的指派。首先通过自行开发的爬虫软件获取一定数量的酒店评论，我们主要对驴评网进行评论爬虫，总评论数为 251132 条，涉及 3535 个酒店。驴评网上的评论采用评分值，5 分最好，0 分最差，为了使得本实验效果更理想，我们在两个分值端点进行数据采集。通过对数据库中的酒店评论评分统计，我们可以发现评论分值分布不平均，小于 2 分的评论数为 17537 条，而大于 4.5 分的评论数居然有 86619 条，分析原因，我们不难理解 2 分以下的评论基本都属于非常差评，而非常差评的酒店数肯定大大少于好评的酒店。所以我们抽取了 2 分以下评价中最差的 1.5 万条贬义评论和 4.5 分以上评论中最好的 1.5 万条褒义评论作为训练集。然后采用第三章中提出的评价对象属性及其情感表达元素的联合识别模型对数据进行处理。同时利用本章前面部分提出的属性分类方法对所有属性描述进行归类，最后采用 PMI（点互信息）方法[119]实现情感词与属性类的关联计算，如公式 4.8 所示。

$$PMI(TPi, Sw) = \log(\frac{p(TPi \,\&\, Sw)}{p(TPi) * p(Sw)}) \qquad \text{（公式 4.8）}$$

其中 $p(TPi)$ 为属性类 TPi 在训练语料集中出现的频率，$p(Sw)$ 为情感词 Sw 在训练语料集中出现的频率，$p(TPi \,\&\, Sw)$ 为属性类 TPi 和情感词 Sw 在训练语料集中同时出现的频率。

通过对训练集中所有出现的情感词与属性类进行点互信息计算，然后选择点互信息值最大的作为其最可能的关联对象属性类，也就是当该情感词出现，但没有对应的属性描述词存在，我们就把该情感倾向值汇总到该关联对象属性类上进行计算。

4.4.2 情感汇总计算

情感汇总是针对所有评论中相同评价对象属性类的情感值进行汇总统计。由于情感信息的描述多样化，可能存在多种修饰词，这对情感值的计算提出了较高的要求，但我们不难发现这跟 2.4.2 章节中复合情感词的极性强度量化计算类似，所以在求解以句子为单位的情感倾向值的过程中，针对识别出来的情感词以及修饰词信息，我们

利用 2.4.2 章节中提出的方法实现情感倾向值的求解。

由于本文没有考虑评论发表者的权重情况，所以我们认为所有评论的权重相等，情感汇总计算任务也就转变为计算某个属性类对应的情感平均值，如公式 4.9 所示。

$$\overline{S_{c(i)}} = \frac{\sum_{j=1}^{n(c(i))} S_{(c(i),j)}}{n(c(i))} \qquad \text{（公式 4.9）}$$

其中 *c(i)*为属性类 i，*n(c(i))*为评论中属性类 *c(i)*出现的总次数，$S_{(c(i),j)}$ 为评论中第 j 次出现的属性类 *c(i)*对应的情感倾向值，$\overline{S_{c(i)}}$ 为所有评论中属性类 *c(i)*所对应的平均情感倾向值。

4.4.3 实验结果及分析

我们首先在所有驴评网酒店评论中抽取出存在评价对象属性缺失的酒店评论，由于人工标注代价的原因，我们从中随机只挑选出 420 条评论作为实验数据，进行属性类别与情感关联正确率实验和情感计算实验。通过人工标注分析，这 420 条评论中有情感描述但缺失对象属性的情况有 489 处，我们首先利用公式 4.8 完成了 3 万句评论中出现的所有情感词与属性类之间的关联计算和最可能属性类的确定，然后利用这个结果对 489 处有情感描述但缺省对象属性的评论进行关联指派，结果如表 4.8 所示。

表 4.8 情感词的属性类关联指派实验结果

缺失对象属性描述的情感词个数	正确关联指派数	无法关联指派数	错误关联指派数	关联正确率
489	433	2	54	0.885

从表 4.8 的结果显示，关联正确率达到了 0.885，同时分析关联指派错误的情况发现，主要原因是，这些情感词在原先的关联计算中，就存在与不同属性类的关联概率区分度不大的情况，如情感词“好”，在七个属性类中，都存在关联实例，所以在指派过程中容易出现错误。另外从表 4.8 中我们还发现无法关联指派数有 2 个，主要是因为这 2 个情感词没有出现在 2 万条评论训练集中，所以也就没有对应的关联指派信息。针对接下去的情感计算，我们认为当出现没有指派信息的时候，默认添加到“整体”属性类。因为这样不会造成数据浪费，另外作为“整体”属性类，本身相对比较

宏观。

最后，情感汇总计算的结果可以通过图 4.15 来体现。

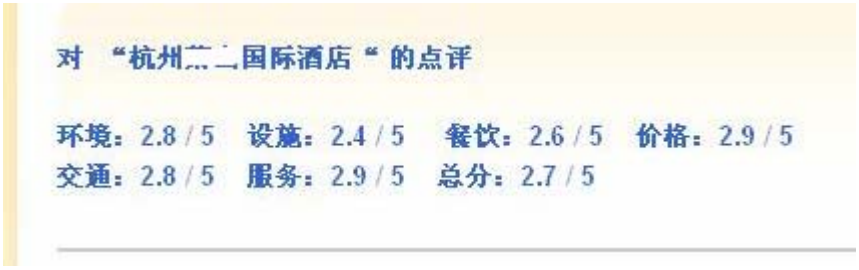

图 4.15 细粒度情感汇总结果

图 4.15 是对“杭州**国际酒店”的评论进行情感汇总计算结果显示，为了跟驴评网的评分标准一致，我们把情感倾向值的计算结果统一调整到 0 到 5 的分值区间，5 分最好，0 分最差。

4.5本章小结

本章首先通过有监督学习方法对属性分类进行相关研究，特征设计过程中主要考虑了属性描述及其上下文词汇信息，以及属性描述识别结果对应的词性标注信息及其上下文词性标注信息，通过实验对比，找出最为理想的特征组合。然后研究基于自举学习的属性分类任务，首先通过自举学习的分层种子选取策略，先将所有实例按某一属性进行分层，再分别从每层中按该层的实例数量占所有实例总数的比例来抽取样本。另外，我们也把这种分层思想应用到自举过程的每一步迭代中，同时也探讨了自举迭代的终止条件。最后针对评论中可能存在情感词但缺少对象属性描述的情况，我们首先构建 3 万条褒贬评论（1.5 万褒义评论，1.5 万贬义评论）作为训练集，然后利用前面的方法完成情感词与对象属性类的识别，再通过计算情感词与对象属性类的 PMI 值来确定两者之间的关联概率，实现对缺失对象属性的情感信息进行合理属性类的指派，使情感汇总计算更为合理有效。

第5章　基于细粒度情感分析方法的酒店评论意见挖掘系统

5.1引言

近年来，随着文本情感倾向性分析研究的深入，简单对商品评论进行文档级的情感分析已经不能满足应用的需要，不少研究者开始进行更细粒度的产品评论情感分析[9][56][57]。

从具体的应用来看，NEC 美国研究所 Dave 等人[49]研究并开发的 ReviewSeer 是世界上第一个情感分析工具和第一个针对给定产品评论区别其褒贬性的系统，通过对评论性文章的语义倾向分析，为商品的受欢迎程度进行打分评价，该评价结果是极具价值的商业信息。微软研究院的 Gamon 等人[120]研究利用聚类、半监督学习方法进行句子的语义分类，并开发了 Pulse 系统实现自动挖掘网上用户所上载的自由文本中有关汽车评价中的褒贬信息和强弱程度。美国伊利诺大学的 Liu 等人[121]研究并开发了 Opinion Observer 系统，实现网上顾客的在线商品评价处理，对评论中出现的产品各个属性（特征）的用户褒贬意见进行统计，给出友好的产品特征分类可视化界面展示，同时还提供了同类产品之间的评价比对功能，使各部分属性（特征）优劣一目了然，极大帮助了用户的购买决策。IBM 研究中心的 Yi 等人[54]研究并开发了一个面向在线评论的情感分析系统（Sentiment Analyzer），该系统利用自然语言处理技术建立情感词库和情感语言模式库，对在线评论进行特征术语抽取、观点提取以及观点和特征关系的关联性分析，最终实现在线评论的情感分析。美国匹兹堡大学的 Wilson 等人[122]研究并开发了 OpinionFinder 系统，它实现了主观性句子自动识别以及句子中各种与主观性有关的成分（例如，意见源、直接的主观性表达、说话事件（Speech Event）、情感等）挖掘。英国科波拉软件公司于 2005 年推出了一套舆情感情色彩分析软件，它主要是通过网络舆情过滤和分级技术实现的。该技术可自动分辨语法成分，例如名词、动词和形容词，并确定动词的主语和宾语，因此可以去除一些与文章主要内容无关的词语，从而判断文章的感情色彩是正面、负面还是中立的，以帮助政府和一些大

公司了解民意。另外美国国土安全部于 2006 年起利用能概述和分析新闻报道中公众意见的情感分析软件获取民众意愿，把握社情民意的走向。

从国内来看，文本情感倾向性分析技术更多是应用于网络舆情监控系统，如方正的智思系统、厦门美亚柏科、邦富软件和谷尼国际软件等。针对网络舆情中各类评论的情感分析，必然要用到文本倾向性分析技术，但由于上述软件更多的是基于篇章的粗粒度情感倾向性分析，从技术实现上相对比较简单和传统。另外，粗粒度情感分析学术研究在国内已经相对比较成熟，如清华大学的孟凡博等人[123]设计了一个基于关键词模板的电影评论褒贬倾向判定系统，从结果来看，集外测试的效果不够理想，主要缺乏对句子的语义理解。哈尔滨工业大学的徐军等人[59]使用机器学习方法实现了一个新闻情感自动分类系统，在一定实验环境下，最高达到了 90%的准确率，领先于其他基于篇章的情感倾向性分析方法。香港城市大学的 Tsou 等人[124]设计了一个面向报刊上关于政治人物具有褒贬性的报告的情感分类系统，通过利用统计分析方法得到最终的文本褒贬分类和强度。近几年，已有不少学者开始细粒度情感倾向性分析方法，如上海交通大学的姚天昉等人[125]研究开发了用于汉语汽车论坛的意见挖掘系统，可以实现在电子公告板、门户网站等各大论坛上的意见挖掘，对褒贬信息进行综合统计后给出可视化结果。

在前面章节中介绍的细粒度情感分析方法基础上，我们将进行功能集成并开发一个具体应用系统。本章主要介绍了基于细粒度情感倾向性分析的酒店评论意见挖掘系统架构及功能模块的设计与实现。首先详细分析了评论数据采集及预处理模块，实现了爬虫程序的设计和网络评论的预处理及格式化的数据保存，随后采用前面章节提出的细粒度情感分析方法，完成评论数据的对象属性及其情感表达元素的联合识别，并利用已有的情感词典及极性强度量化结果，完成以酒店为单位的基于属性类别的情感汇总计算，最后给出友好的可视化浏览及查询界面。系统还提供了根据不同区域以及用户关心的酒店属性类别进行排名推荐功能。为了方便外部应用的调用，我们还提供了相应的接口以帮助实现在线评论的实时细粒度情感分析需求。

5.2 系统架构及功能模块

本系统主要把前面几部分研究内容与系统实际运行的功能模块进行有机结合，系

统由下往上主要可以分为评论数据采集模块、数据处理模块、数据分析模块、信息展示模块等四部分。其中评论数据采集模块，我们主要针对目标网站设计爬虫软件进行评论数据的采集和存储，在本文中，根据我们的需要主要针对酒店评论网站——驴评网进行信息爬虫，数据存储之前先对网页进行过滤和格式化信息抽取，只保存每条评论的发表时间、发表人、评论标题以及评论内容。数据处理模块主要针对评论数据进行相应的处理，其中包括应用自然语言处理技术进行评论数据的语义特征提取，如分词、词性标注、语义角色标注等，另外利用机器学习方法实现对抽取后的各种特征建立相应的学习模型，然后对新的评论信息进行预测。数据分析模块主要针对数据处理模块处理后的信息进行情感分析，利用对象属性与情感词之间的关联信息以及情感词与修饰词之间的关系进行细粒度情感强度量化统计和计算。信息展示模块主要针对处理和分析后的评论信息进行友好的可视化展示，以及提供相应的查询接口，帮助用户根据评论信息中各属性的情感值进行酒店推荐。系统的总体框架如图 5.1 所示。

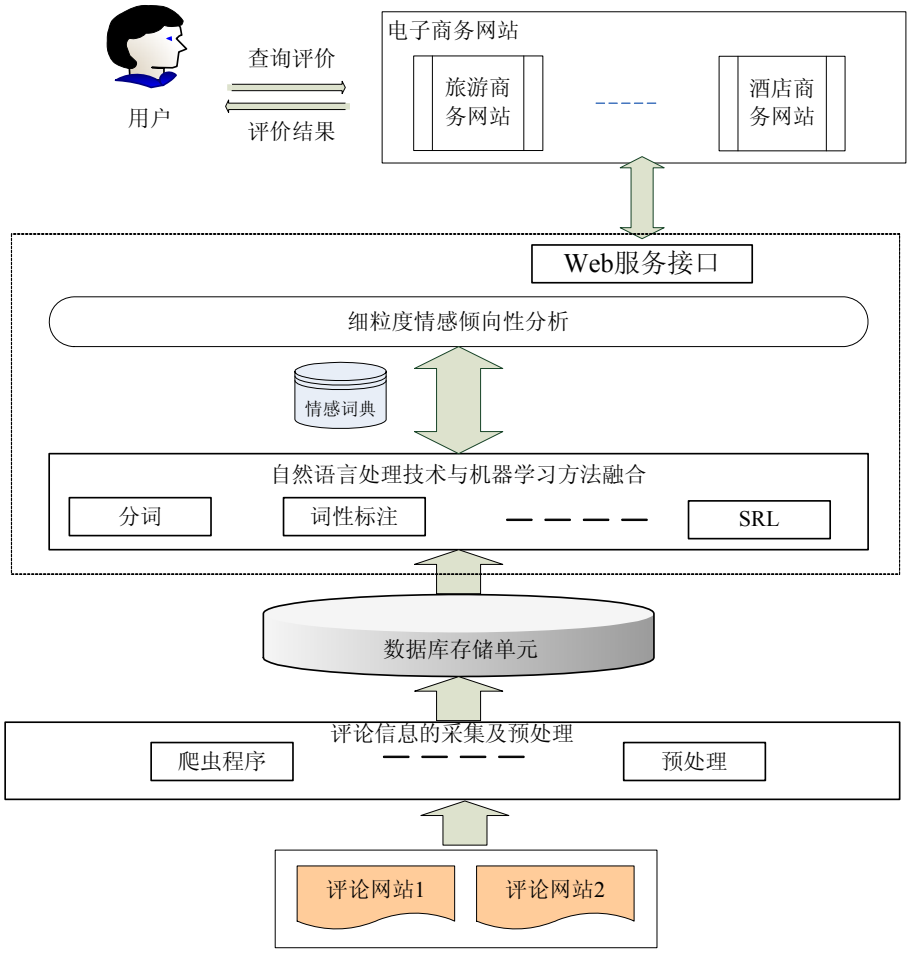

图5.1 系统的总体框架图

本系统的信息展示模块是通过自己构建一个网站来实现，用户可以在我们的网站上查阅酒店信息、相关评论及其情感倾向值，也提供相关检索界面，方便用户进行目标酒店的搜索。为了方便统一调用，我们对情感倾向性分析模块进行服务接口封装，这样也方便了第三方用户对本系统模块的调用。

5.2.1 评论数据采集及预处理模块

评论数据采集及预处理模块主要实现了目标评论网站的网页爬虫和网页中评论内容的格式化抽取。

随着网络的迅速发展，万维网成为大量信息的载体，如何有效地提取并利用这些信息成为一个巨大的挑战。网络爬虫是一种按照一定规则，自动抓取万维网信息的程序或者脚本。系统的主要目标是获取评论网站上的评论信息，所以我们仅锁定网站上评论相关区域的页面进行爬虫。以往所使用的技术为人工分析页面结构、编写标识符来定位目标信息，而本系统使用了 XPath 以及 Python 的扩展库 lxml，极大的提高了编写抓取程序的效率和程序运行速度、可读性。XPath 是一门在 XML 文档中查找信息的语言，可用来在 XML 文档中对元素和属性进行遍历，它基于 XML 的树状结构，提供在数据结构树中找寻节点的能力。而 lxml 库则可以快速正确地分析 XML 文档。具体到实现中，我们可以将 HTML 页面数据看成是 XML 数据的特殊形式，所以可以使用 XPath 来表示一个评论在此 HTML 文档中的具体位置。此外 XPath 的使用也非常方便，可以使用工具自动生成。另外，结合 lxml 所提供的方法，我们可以高效的实现目标信息的提取，如评论内容、用户名、评论发表日期、用户打分等，从而实现评论信息的格式化抽取，并存入预先设计好的相关数据库。

由于评论是用户生成的数据，所以往往存在书写格式不规范的情况，为了降低对后面文本分析和自然语言理解的影响，我们首先做了一些预处理，如去除空行、去除多余的空格、去除重复标点符号等，然后把预处理后的评论信息保存到对应的原始评论记录中的相应字段。

5.2.2 数据处理与分析模块

评论数据的处理与分析模块是本系统的核心部分，因为它直接关系到系统的处理

性能。我们主要集成应用了前几章中所提到的相关技术，实现评论数据的语义分析。概括的讲，主要分为三块内容：

（1） 利用自然语言处理技术实现评论语句的语义特征提取

我们主要利用分词技术、词性标注（POS Tagging）、语义角色标注（SRL）等自然语言处理技术，实现对评论句子的语义分析和处理，并抽取和转化为相应的特征表示，为后面的机器学习奠定了基础。

（2） 机器学习方法进行评价对象属性与情感要素的联合识别

主要利用已获取的语义特征信息，通过构建相应的学习模型实现评价对象属性及其情感要素的联合识别。在模型的生成过程中我们通过反复试验，调整特征模板和充分利用上下文信息，努力提高了模型的性能。通过前面章节的分析，我们利用性能最好的模型对所有评论进行识别，并对所有评价对象属性描述进行属性分类。这为后面的情感量化计算奠定基础。

（3） 基于分类属性的情感量化计算方法设计

主要利用已抽取的“属性-情感-修饰词”词对、属性分类信息以及上下文语义信息，找出情感词与相关修饰词之间的各种关系，设计出不同的情感计算方法，进一步提高计算精确度。具体内容包括：基于上述的实验数据和结果，研究相应的语言学规律，总结不同的计算方法实现最终的基于分类属性的情感量化汇总。通过以上工作，我们把这些研究成果应用到目前所有评论中进行情感计算，最后以酒店为单位进行基于分类属性的情感汇总。

5.2.3 信息展示模块

信息展示模块主要面向用户，而前面两个模块主要是系统内部处理为主，跟用户几乎没有互动。本模块主要实现图形化界面，以便捷友好的展示方式进行商品查询和推荐。

（1） 详细查看商品

向用户展示详细、全面的酒店信息列表，并提供分页显示功能。用户可以查看酒店的名称、图片、地址、网址以及已评论的人数之外，还可以直观地看到酒店各项分类属性的情感得分。

（2） 评论个性化展示

评论内容如果比较长的话，在传统的网页显示条件下，用户往往不愿意仔细浏览，这样容易导致信息的丢失。而我们设计的页面充分利用前面的评价对象属性及其情感表达元素识别结果，通过不同颜色标记出评论中的评价对象属性、情感描述及其修饰词，同时还在相应的位置显示该评论的情感计算结果，这样使得用户浏览商品评论时更加直观和方便。

（3） 人性化检索方式

用户可以对酒店进行关键字检索和情感检索。由于在评论爬虫时，我们保存了酒店与评论之间的完整关系，所以本系统给用户直接提供了地区选择查询，并提供了评价对象属性类别选择，支持以基于细粒度分类属性的情感得分作为排序条件，从而为用户提供更好的排名推荐。

（4） 快速酒店预订导航

用户通过本系统检索到感兴趣的酒店时，可以直接点击订购选项，系统将自动跳转到酒店预订导航页面，通过给出多家酒店预订网站的价格对比，使用户有了更多的选择余地。

5.2.4 情感分析服务化封装

本文所提出的情感计算方法可应用于其他各酒店预订网站的评论分析，而这些网站系统往往是分布、异构的。为了方便本研究成果的推广使用，我们提出一种基于 Web 服务框架的服务化封装方法，将文本情感分析方法封装为 Web 服务资源，以屏蔽方法自身的复杂性，对外呈现统一的调用接口，实现在网络环境中的共享，以适应不同的应用需求。同时本方法的设计为今后多领域的情感分析服务化封装奠定了基础。

Web 服务是一种基于 XML 和 SOAP 的技术，它完全屏蔽不同软件平台之间的差异。Web 服务本质就是一个应用程序，向用户提供了一个能够通过 Web 进行调用的 API，用户能够用编程的方法通过 Web 来调用这个应用程序。因此，将本文研究的情感分析方法封装成 Web 服务后，其他的系统就可以按标准的 Web 形式进行访问，而不需要考虑分析方法的具体实现，简化用户的程序设计。情感分析 Web 服务框架如

图 5.2 所示。

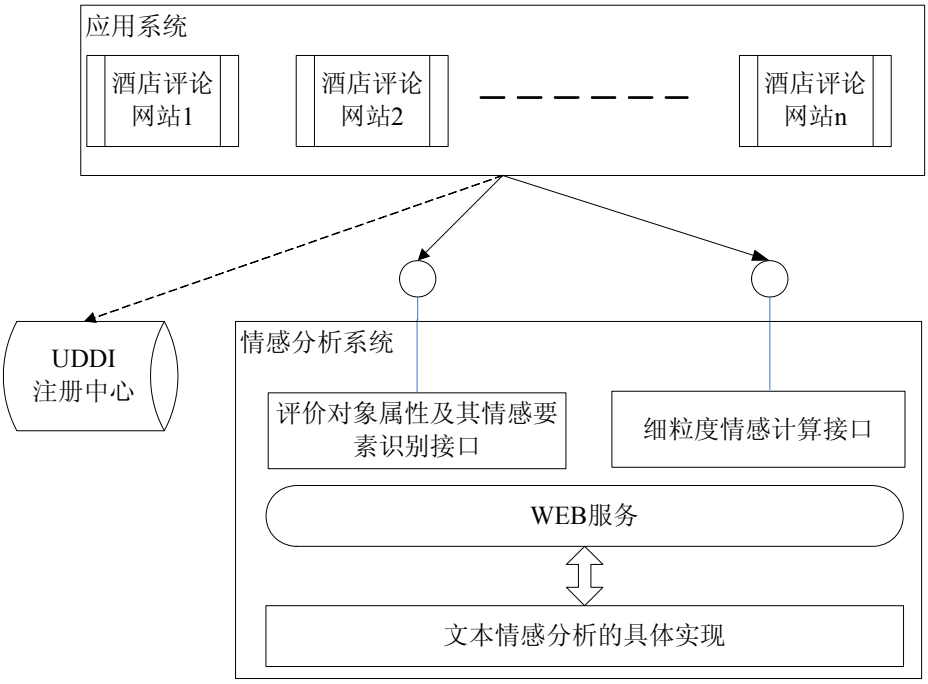

图5.2 基于Web服务的文本情感分析框架

在这个框架中，将本文研究的情感分析研究成果发布为 Web 服务，并注册到 UDDI 注册中心。各需要进行文本情感分析的应用系统可通过 UDDI 注册中心查询并调用相应的 Web 服务得到对应的情感分析结果。

5.3 系统实现

打开网站主页，我们可以看到如图 5.3 所示的主页界面，通过主页上的信息提示，用户可以方便进入相关城市酒店的查询，或者在文本框中输入相关条件进行酒店检索。当我们点击进入某个具体酒店时，就可以查看到酒店评论的所有信息，以及分类属性的情感得分。

图5.3 网站主页界面

5.3.1 酒店评论处理

点击主页中任一家酒店的超链接均可进入该酒店评论的具体信息页面。图 5.4 显示了其中 2 条评论的信息。

图 5.4 的顶部区域显示该酒店评论总数为 82 条，通过利用前面章节介绍的方法最后对 82 条评论进行情感汇总，最后向用户展示了酒店各项属性特征的细粒度评分：“环境：2.8”，“设施：2.4”，“餐饮：2.6”，“价格：2.9”，“交通：2.8”，“服务：2.9”，“总分：2.7”。另外，从图 5.4 的下半部分我们可以看到了一条评论的具体处理结果，系统利用评价对象属性及其情感表达元素的识别结果，对它们进行了颜色标注，绿色代表着评论者的情感描述词，即评论者最直接的情感体现，例如“干净”、“合适”、“不错”等；红色代表评论者的评论对象属性，例如“价格”、“床”、“设施”等；蓝色代表评论的程度词，例如“很”、“稍微有点”。这样的显示效果，使评论浏览者更加一目了然，方便用户做出正确的判断。每条评论上方都有一个分数，这是对该评论进行情感分析计算后的得分显示，是通过求解该评论中出现的所有情感的平均值来获

得。

图5.4 酒店评论的信息页面

通过点击每条评论下方的“评论出处及预订”链接可以直接进入酒店的预订界面导航，如图 5.5 所示。

图5.5 酒店预订页面

用户可以通过预订页面中给出的不同酒店预订网站的价格对比，选择自己最满意的预订网站进行酒店预订。

5.3.2 基于评论对象属性的酒店检索

本系统还提供可以满足用户个性化需求的 3 种检索方式：关键字检索、情感检索以及关键字情感组合检索功能，其中情感检索可以就用户所关心的酒店属性进行推荐排序，从而满足用户的个性化需求，而组合检索结合了关键字检索和情感检索。

（1） 关键字检索

网站首页就给出了检索界面，如图 5.3 所示。以酒店预订为例，关键字检索可以按照酒店所在的地区或者酒店的名字进行检索。

搜索关键字：“快捷”，下拉框中选择“酒店名”进行搜索，结果如图 5.6 所示。搜索结果界面向用户显示了包含该关键字的所有酒店，并显示搜索到的酒店数量，以及各个酒店的基本信息和各项属性评分信息。

图5.6 关键字检索结果

（2） 情感检索

情感检索可以按照用户选择的酒店城市和关心的酒店属性类别进行检索，并将搜索结果根据用户检索的属性类别分数进行排序推荐。如图 5.7，5.8 所示。

搜素界面：选择城市为“杭州”，选择属性类别为“设施”，进行检索。

搜索结果：向用户显示搜索到的符合要求的酒店数量，并且所有酒店以“设施”属性类别的情感得分降序排列。

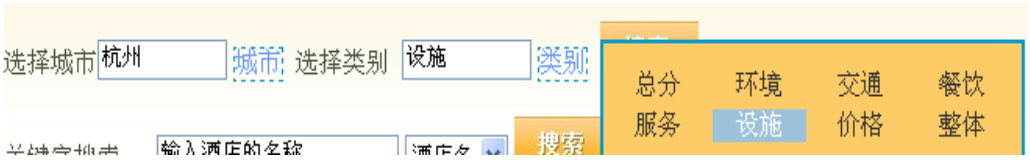

图5.7 情感检索界面

图5.8 情感检索结果

通过这个功能，用户可以方便的按照自己所关心的酒店某几个属性类别进行检索，由于酒店属性类别分数是基于用户主观性评论的情感量化计算结果，能够真实反

映用户在这一方面的现实体验，这样的检索功能更符合用户的实际需求。假设用户预定酒店，他只关心“服务”和“餐饮”这两方面的用户评价，那么本系统所提供的功能就能满足其需求，帮助他预订到满意的酒店。

另外，用户还可以在关键字检索的同时，选择是否需要按照酒店某一情感属性类别进行排序推荐。在用户输入关键字的同时，可以选择相应的酒店属性类别，并在情感排序处打钩。结果将把符合关键字的酒店按照相应属性类别评分由高到底推荐给用户。

5.4本章小结

本章首先设计相关的爬虫软件实现对目标评论网站数据的收集，然后利用相关技术实现对网络评论的预处理及格式化的数据保存，随后采用前面章节提出的细粒度情感分析方法，完成评论数据的对象属性和情感表达元素联合识别，并利用已有的情感词典及极性强度量化结果，完成以酒店为单位的基于属性类别的情感汇总计算，最后给出友好的可视化浏览及查询界面。系统还提供了根据不同区域以及用户关心的酒店属性类别进行排名推荐功能。为了方便外部应用的调用，我们还提供了相应的接口以帮助实现在线评论的实时细粒度情感分析需求。

本章内容主要是面向用户，但从系统的核心模块来讲，主要还是建立在前面几章的研究基础之上，所以本章主要解决了从系统内部核心功能模块的集成到结果的用户展示这一整套流程，使我们的研究成果更好地与应用相结合，实现从技术研究到具体应用的集成。

第6章　总结与展望

本章简要总结了本文所研究的主要内容及相关贡献，并对下一步研究工作进行了展望。

6.1本文研究总结

本文针对细粒度情感分析任务进行了详细研究。首先研究了现有的情感词极性强度量化方法，通过分析提出了相应的改进方法，实验结果表明，改进后的方法较大地提高了计算性能。其次研究了评价对象属性及其情感表达元素的联合识别，分析了基本特征和语义特征的相关知识和抽取方法，特别针对语义特征的抽取进行了技术分析和算法设计，并通过实验对比证明了语义特征及特征模板设计的重要性。然后研究了细粒度属性分类及情感计算，通过设计不同的实验，证明了半监督学习方法在属性分类中的有效性，同时通过设计合理的情感计算方法实现基于属性类的情感汇总。最后设计了一个基于细粒度情感分析方法的酒店评论意见挖掘系统，有效地实现系统内部核心功能的封装，并最终实现友好的用户界面展示。

具体来说，本文取得的成果表现在以下几个方面：

1. 研究了情感词极性强度量化方法。首先分析了情感极性强度模糊性的特点以及极性强度量化计算对情感分析工作的重要性；在现有算法的基础上，通过分析和研究，提出了改进方案，对情感词的情感极性强度量化进行分类处理，充分利用了字词之间的关系以及语言学知识，并设计出相应的规则和方法。实验结果表明，改进后的方法较大地提高了计算性能。

2. 研究了评价对象属性及其情感表达元素的联合识别。通过对任务的深入分析，结合条件随机场理论，介绍了基于序列化结构的联合识别模型，有效利用评价对象属性及其情感表达元素之间的类别关系；分析了基本特征和语义特征的相关知识和抽取方法，特别针对语义特征的抽取进行了技术分析和算法设计；最后通过实验对比证明了语义特征及特征模板设计的重要性。

3. 研究了半监督学习理论，重点介绍了自举学习在属性分类中的应用，通过自

举学习的分层种子选取策略，先将所有实例按某一属性进行分层，再分别从每层中按该层的实例数量占所有实例总数的比例来抽取样本；另外，我们也把这种分层思想应用到自举过程的每一步迭代中，同时也探讨了自举迭代的终止条件；最后针对评论中可能存在情感词缺少对象属性的情况，我们通过计算PMI值来确定评价对象属性类与情感词之间的关联概率，实现对缺失对象属性的情感信息进行合理属性类的指派，使情感汇总计算更为合理有效。

6.2下一步工作设想

本文的研究虽然在酒店评论领域实现了整个系统的集成，也体现出一定的应用效果，但总结整个过程，还有许多有待继续研究的地方。另外，面对互联网日新月异的变化，不断涌现出新的应用环境，情感分析的领域也在不断的扩大。归纳起来，有以下几个方面可以作为下一步的研究工作：

1. 跨领域情感分析。现在已有不少针对跨领域情感分析的研究，但离真正应用还有不小的距离。主要原因是当领域差别过大时，分析性能下降的非常厉害，这比领域内的情感分析效果差了许多，因此需要我们对跨领域情感分析过程中的学习算法和相关问题做更深入的研究。

2. 面向社交网络的情感分析。社交网络在近几年飞速发展，面向微博的情感分析任务已经引起大家的注意，特别针对微博中出现的社会事件讨论，如何有效的进行事件的倾向性分析、情绪状态识别和情绪程度识别等，随着时间的推移，甚至还可以研究一个社会事件在传播过程中网民情感倾向的变化过程。这些问题的研究一定程度上跟评论情感分析任务有相似之处，但具有更多的新特点，所以需要我们针对新的应用环境和任务进行深入研究。

3. 跨语言的情感分析。由于评论中有时会出现多种语言的情况，比如中文评论中，会出现某些英文单词，这对局限在单语言环境下的情感分析任务来说需要改进。跨语言的情感分析任务在一定程度上要求相当高，除了需要解决情感分析任务中遇到的问题，往往还需要实现机器翻译、多语言文本处理等工作，这对情感分析提出了更大的挑战。

参考文献

[1] CNNIC, 2013. 第 31 次中国互联网络发展状况统计报告[EB]. http://cnnic.net.

[2] Ku L.W., Liang Y.T., Chen H.H. Opinion Extraction, Summarization and Tracking in News and Blog Corpora[C]. In: Proceedings of AAAI'2006, 2006: 100-107.

[3] Baccianella S., Esuli A., Sebastiani F. Multi-facet Rating of Product Reviews[C]. In: Proceedings of ECIR'2009, 2009: 461-472.

[4] Gyamfi Y., Wiebe J., Mihalcea R., Akkaya C. Integrating Knowledge for Subjectivity Sense Labeling[C]. In: Proceedings of NAACL-HLT'2009, 2009: 10-18.

[5] Devitt A, Ahmad K. Sentiment Polarity Identification in Financial News: A Cohesion-based Approach[C]. In: Proceedings of ACL'2007, 2007: 984-991.

[6] Esuli A., Sebastiani F. PageRanking WordNet Synsets: An Application to Opinion Mining[C]. In: Proceedings of ACL'2007, 2007: 424-431.

[7] Turney P.D. Thumbs up or Thumbs down? Sentiment Orientation Applied to Unsupervised Classification of Reviews[C]. In: Proceedings of ACL'2002, 2002: 417-424.

[8] Weichselbraun A., Gindl S., Scharl A. Using Games with a Purpose and Bootstrapping to Create Domain-Specific Sentiment Lexicons[C]. In: Proceedings of CIKM'2011, 2011: 1053-1060.

[9] Hu M., Liu B. Mining and Summarizing Customer Reviews [C]. In: Proceedings of KDD'2004, 2004: 168-177.

[10] Bollegala D., Weir D., Carroll J. Using Multiple Sources to Construct a Sentiment Sensitive Thesaurus for Cross-Domain Sentiment Classification [C]. In: Proceedings of ACL'2011, 2011: 132-141.

[11] Kim S.M., Hovy E. Determining the Sentiment of Opinions[C]. In: Proceedings of COLING'2004, 2004: 1367-1373.

[12] Yu H., Hatzivassiloglou V. Towards Answering Opinion Questions: Separating Facts from Opinions and Identifying the Polarity of Opinion Sentences[C]. In: Proceedings of EMNLP'2003, 2003: 129-136.

[13] Zhuang L., Jing F., Zhu X.Y. Movie Review Mining and Summarization[C]. In: Proceedings of CIKM'2006, 2006: 43-50.

[14] Su Q., Xu X.Y., Guo H.L., Guo Z.L., Wu X., Zhang X.X., Swen B., Su Z. Hidden Sentiment Association in Chinese Web Opinion Mining[C]. In: Proceedings of WWW'2008, 2008: 959-968.

[15] Hatzivassiloglou V., McKeown K. Predicting the Semantic Orientation of Adjectives [C]. In: Proceedings of ACL'1997, 1997: 174–181.

[16] Kobayashi N., Inui T., Inui K. Dictionary-based Acquisition of the Lexical Knowledge for P/n

Analysis (in Japanese) [C]. In: Proceedings of Japanese society for Artificial Intelligence, SLUD-33. 2001: 45–50.

[17] 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.

[18] 路斌, 万小军, 杨建武, 等. 基于同义词词林的词汇褒贬计算[C]. 中国计算技术与语言问题研究-第七届中文信息处理国际会议论文集. 2007: 17-23.

[19] Kamps J., Marx M., Mokken R.J., et al. Using Wordnet to Measure Semantic Orientation of Adjectives [C]. In: Proceedings of LREC'2004, 2004: 1115-1118.

[20] Quan C.Q., Ren F.J. Construction of a Blog Emotion Corpus for Chinese Emotional Expression Analysis [C]. In: Proceedings of EMNLP'2009, 2009: 1446–1454.

[21] Dong Z.D. HowNet Knowledge Database[EB]. http://www.keenage.com. 2010.

[22] 柳位平, 朱艳辉, 栗春亮, 向华政, 文志强. 中文基础情感词词典构建方法研究[J]. 计算机应用. 2009, 29(10): 2875-2877.

[23] 徐琳宏, 林鸿飞, 潘宇, 任惠, 陈建美. 情感词汇本体的构造[J]. 情报学报, 2008, 27(2): 180-185.

[24] Wiebe J., Breck E., Buckley C., et al. NRRC summer workshop on multi-perspective question answering [R]. 2002.

[25] Yang Hongwu, Meng Helen M., Wu Zhiyong and Cai Lianhong. Modeling the Global Acoustic Correlates of Expressivity for Chinese Text-to-Speech Synthesis[C]. In: Proceedings of IEEE/ACL 2006 Workshop on Spoken Language Technology. 2006. 10-13.

[26] 宋鸿彦, 刘军, 姚天昉, 刘全升, 黄高辉. 汉语意见型主观性文本标注语料库的构建[J]. 中文信息学报, 2009, 23(2): 123-128.

[27] 徐琳宏, 林鸿飞, 赵晶. 情感语料库的构建和分析[J] .中文信息学报, 2008, 22(1):116-122.

[28] Ku L.W., Wu T.H., Lee L. Y., Chen H.H. Construction of an Evaluation Corpus for Opinion Extraction. In: Proceedings of NTCIR-5 Workshop Meeting, Tokyo, Japan, 2005.

[29] 姚天昉, 程希文, 徐飞玉等. 文本意见挖掘综述[J]. 中文信息学报, 2008, 22(3): 71-80.

[30] Spertus E. Smokey: Automatic Recognition of Hostile Messages [C]. In: Proceedings of Innovative Applications of Artificial Intelligence, 1997: 1058-1065.

[31] Aone C., Ramos-Santacruz M., Niehaus W.J. Assentor: An NLP-Based Solution to E-mail Monitoring [C]. In: Proceedings of AAAI'2000, 2000: 945-950.

[32] Barzilay R. Collins M., Hirschberg J., Whittaker S. The Rules Behind Roles: Identifying Speaker Role in Radio Broadcasts [C]. In: Proceedings of AAAI'2000, 2000: 679–684.

[33] Hovy E. Combining and Standardizing Large-Scale, Practical Ontologies for Machine Translation and Other Uses [C]. In: Proceedings of LREC'1998, 1998: 535-542.

[34] Riloff E., Wiebe J., Phillips W. Exploiting Subjectivity Classification to Improve Information Extraction [C]. In: Proceedings of AAAI'2005, 2005: 1106-1111

[35] Toprak C., Gurevych I. Document Level Subjectivity Classification Experiments in DEFT'09

Challenge [C]. In: Proceedings of the DEFT'09 Text Mining Challenge, 2009: 89-97

[36] Remus R. Improving Sentence-level Subjectivity Classification through Readability Measurement [C]. In: Proceedings of NODALIDA'2011, pp. 168–174.

[37] Wang B., Spencer B., Ling C.X., Zhang H. Semi-supervised Self-training for Sentence Subjectivity Classification[C]. In: Proceedings of Canadian Conference on AI, 2008: 344-355.

[38] Lambov D., Dias G., Graca J.V. Multi-view Learning for Text Subjectivity Classification [C]. In: Proceedings of 19th European Conference on Artificial Intelligence Workshop on Computational Approaches to Subjectivity and Sentiment Analysis (ECAI'2010), 2010.

[39] Finn A., Kushmerick N., Smyth B. Genre Classification and Domain Transfer for Information Filtering[C]. In: Proceedings of the 24th BCS-IRSG European Colloquium on Information Retrieval Research: Advances in Information Retrieval. 2002: 353-362.

[40] Pang B., Lee L. A Sentimental Education: Sentiment Analysis using Subjectivity Summarization based on Minimum Cuts[C]. In: Proceedings of ACL'2004, 2004: 271-278.

[41] 姚天昉, 彭思崴. 汉语主客观文本分类方法的研究[C]. 第 3 届全国信息检索与内容安全学术会议论文集. 2007. 117−123.

[42] 叶强, 张紫琼, 罗振雄. 面向互联网评论情感分析的中文主观性自动判别方法研究[J]. 信息系统学报, 2007, 1(1): 79-91.

[43] Hearst M. Direction-based Text Interpretation as an Information Access Refinement[J]. Text-Based Intelligent Systems. 1992: 257-274.

[44] Sack W. On the Computation of Point of View[C]. In: Proceedings of AAAI'1994, Student abstract, 1994.

[45] Huettner A., Subasic P. Fuzzy Typing for Document Management[C]. In: Proceedings of ACL'2000, 2000:26-27.

[46] Das S., Chen M. Yahoo! for Amazon: Extracting Market Sentiment from Stock Message Boards[C]. In: Proceedings of the 8th Asia Pacific Finance Association Annual Conference, 2001.

[47] Tong R.M. An Operational System for Detecting and Tracking Opinions in on-line Discussion. In: Proceedings of SIGIR 2001 Workshop on Operational Text Classification. 2001.

[48] Dini L., Mazzini G. Opinion Classification through Information Extraction[C]. In: Proceedings of the Conference on Data Mining Methods and Databases for Engineering, Finance and Other Fields, 2002: 299-310.

[49] Dave K., Lawrence S., Pennock D.M. Mining the Peanut Gallery: Opinion Extraction and. Semantic Classification of Product Reviews[C]. In: Proceedings of WWW'2003, 2003: 519-528.

[50] Pang B, Lee L, Vaithyanathan S. Thumbs up?: Sentiment Classification using Machine Learning Techniques[C]. In: Proceedings of EMNLP'2002, 2002. 79-86.

[51] Chaovalit P., Zhou L. Movie Review Mining: A Comparison between Supervised and Unsupervised Classification Approaches[C]. In: Proceedings of the 38th Hawaii International

Conference on System Sciences. 2005: 1-9.

[52] Mullen T., Collier N. Sentiment Analysis using Support Vector Machines with Diverse Information Sources[C]. In: Proceedings of EMNLP'2004, 2004: 412-418.

[53] Whitelaw C., Garg N., Argamon S. Using Appraisal Groups for Sentiment Analysis. In: Proceedings of CIKM'2005, 2005: 625-631.

[54] Yi J., Nasukawa T., Bunescu R., Niblack W. Sentiment Analyzer: Extracting Sentiments about a Given Topic using Natural Language Processing Techniques[C]. In: Proceedings of ICDM'2003, 2003: 427-434.

[55] Fu G., Wang X. Chinese Sentence-Level Sentiment Classification Based on Fuzzy Sets[C]. In: Proceedings of Coling'2010, 2010: 312-319.

[56] Johansson R., Moschitti A. Extracting Opinion Expressions and Their Polarities - Exploration of Pipelines and Joint Model[C]. In: Proceedings of ACL'2011, 2011:101-106.

[57] Choi Y., Cardie C. Hierarchical Sequential Learning for Extracting Opinions and Their Attributes. In: Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics, 2010: 269-274.

[58] 徐琳宏, 林鸿飞, 杨志豪. 基于语义理解的文本倾向性识别机制[J]. 中文信息学报, 2007, 21(1): 96-100.

[59] 徐军, 丁宇新, 王晓龙. 使用机器学习方法进行新闻的情感自动分类[J]. 中文信息学报, 2007, 21(6): 95-100.

[60] 唐慧丰, 谭松波, 程学旗. 基于监督学习的中文情感分类技术比较研究[J]. 中文信息学报, 2007, 21(6): 88-94,108.

[61] 张伟, 李培峰, 朱巧明. 基于树核函数的英文句子情感分类研究[J]. 计算机应用与软件, 2011, 28(4): 30-32,39.

[62] Kaji N., Kitsuregawa M. Building Lexicon for Sentiment Analysis from Massive Collection of HTML Documents[C]. In: Proceedings of the Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL), 2007: 1075–1083.

[63] Qiu G., Liu B., Bu J., et al. Expanding Domain Sentiment Lexicon through Double Propagation[C]. In: Proceedings of the 21st international jont conference on Artifical intelligence, 2009: 1199–1204.

[64] Kanayama H., Nasukawa T. Fully Automatic Lexicon Expansion for Domain-oriented Sentiment Analysis[C]. In: Proceedings of EMNLP'2006, 2006: 355–363.

[65] 庄丽. 评论性信息挖掘研究[D]. 北京: 清华大学, 2007.

[66] Popescu A.M., Etzioni O. Extracting Product Features and Opinions from Reviews[C]. In: Proceedings of HLT/ACL'2005, 2005: 339–346.

[67] Kamal A., Abulaish M., Anwar T. Mining Feature-Opinion Pairs and Their Reliability Scores from Web Opinion Sources[C]. In: Proceedings of WIMS'2012, 2012.

[68] Fang L., Huang M. Fine Granular Aspect Analysis using Latent Structural Models[C]. In: Proceedings of ACL'2012, 2012:333-337.

[69] Filho P.B., Brun C., Rondeau G. A Graphical User Interface for Feature-Based Opinion Mining. In: Proceedings of NAACL'2012, 2012: 5-8.

[70] Ait-Mokthar S., Chanod J.P. Robustness beyond Shallowness: Incremental Dependency Parsing[J]. Special Issue of NLE Journal, 2002.

[71] Liu K., Xu L.H., Zhao J. Opinion Target Extraction Using Word-Based Translation Model[C]. In: Proceedings of EMNLP-CoNLL'2012, 2012: 1346-1356

[72] Brown P.F., Della Pietra S.A., Della Pietra V.J., Mercer R.L. The Mathematics of Statistical Machine Translation: Parameter Estimation[J]. Computational Linguistics. 1993, 19(2): 263-311.

[73] Miao Q., Li Q., Dai R. AMAZING: A sentiment mining and retrieval system[J]. Expert Systems with Applications. 2009, 36(3): 7192-7198.

[74] Qiu G., Liu B., Bu J., et al. Opinion Word Expansion and Target Extraction through Double Propagation[J]. Computational Linguistics. 2009, 37:9–27.

[75] Jin W., Ho H.H., Srihari R.K. OpinionMiner: a Novel Machine Learning System for Web Opinion Mining and Extraction[C]. In: Proceedings of KDD'2009, 2009: 1195–1204.

[76] Wilson T., Wiebe J., Hoffmann P. Recognizing contextual polarity in phrase-level sentiment analysis[C]. In: Proceedings of the Conference on Human Language Technology and Empirical Methods in Natural Language Processing, Vancouver, British Columbia, Canada: ACL, 2005: 347–354.

[77] Wilson T., Wiebe J., Hoffmann P. Recognizing Contextual Polarity: An Exploration of Features for Phrase-level Sentiment Analysis[J]. Comput. Linguist., 2009, 35:399–433.

[78] Kim S.M., Hovy E. Extracting Opinions, Opinion Holders, and Topics Expressed in Online News Media Text[C]. In: Proceedings of the Workshop on Sentiment and Subjectivity in Text at the joint COLING-ACL'2006, 2006.

[79] Kobayashi N., Iida R., Inui K., Matsumoto Y. Opinion Extraction Using a Learning-Based Anaphora Resolution[C]. In: Proceedings of ACL'2005.

[80] Stoyanov V., Cardie C. Toward Opinion Summarization: Linking the Sources[C]. In: Proceedings of ACL'2006, 2006: 9-14.

[81] Stoyanov V., Cardie C. Partially Supervised Coreference Resolution for Opinion Summarization through Structured Rule Learning[C]. In: Proceedings of EMNLP'2006.

[82] 刘永丹, 曾海泉, 李荣陆, 胡运发. 基于语义分析的倾向性文本过滤[J]. 通信学报, 2004, 25(7): 78-85.

[83] 姚天昉, 娄德成. 汉语语句主题语义倾向分析方法的研究[J]. 中文信息学报, 2007, 21(5): 73-79.

[84] Narayanan R., Liu B., Choudhary A. Sentiment Analysis of Conditional Sentences[C]. In: Proceedings of EMNLP'2009, 2009: 180-189.

[85] Andreevskaia A., Bergler S. Mining WordNet for a Fuzzy Sentiment: Sentiment Tag Extraction from WordNet Glosses[C]. In: Proceedings of EACL'2006, 2006.

[86] Wilson T., Wiebe J. Annotating Opinions in the World Press[C]. In: Proceedings of the 4th ACL SIGdial Workshop on Discourse and Dialogue (SIGdial'2003), 2003:13-22.

[87] 韩琳. 黄侃字词关系研究学术史价值考察[J]. 湖北民族学院学报(哲学社会科学版), 2007, 25(6): 92-96.

[88] 魏慧萍. 关于汉语字词关系的再思考[J]. 南京师大学报(社会科学版), 2004, (1): 135-140.

[89] Lafferty J.D., McCallum A., Pereira F.C.N. Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data[C]. In: Proceedings of ICML'2001, 2001: 282-289.

[90] McDonald R., Hannan K., Neylon T., et al. Structured Models for Fine-to-Coarse Sentiment Analysis[C]. In: Proceedings of ACL'2007, 2007: 432–439.

[91] Breck E., Choi Y., Cardie C. Identifying Expressions of Opinion in Context[C]. In: Proceedings of the 20th International Joint Conference on Artifical Intelligence, 2007: 2683–2688.

[92] Choi Y., Cardie C., Riloff E., et al. Identifying Sources of Opinions with Conditional Random Fields and Extraction Patterns[C]. In: Proceedings HLT-EMNLP'2005, 2005.

[93] Jakob N., Gurevych I. Extracting Opinion Targets in a Single and Cross-Domain Setting with Conditional Random Fields[C]. In: Proceedings of EMNLP'2010, 2010: 1035-1045.

[94] 黄昌宁, 赵海. 中文分词十年回顾[J]. 中文信息学报, 2007, 12(3): 8-19.

[95] Xia F. The Part-Of-Speech Tagging Guidelines for the Penn Chinese Treebank (3.0)[R], Institute for Research in Cognitive Science, 2000.

[96] Baker C.F., Fillmore C.J., Lowe J.B. The Berkeley FrameNet Project [C]. In: Proceedings of the ACL-COLING'1998, 1998: 86-90.

[97] Palmer M., Gildea D., Kingsbury P. The Proposition Bank: An Annotated Corpus of Semantic Roles[J]. Comput. Linguist. 2005, 31(1): 71-106.

[98] Meyers A., Reeves R., Macleod C., et al. The Nombank Project: An Interim Report[C]. In: Proceedings of HLT-NAACL'2004 Workshop: Fronties in Corpus Annotation. Boston, Massachusetts, USA, 2004:24-31.

[99] Xue N., Xia F., Chiou F.D., Palmer M. The Penn Chinese Treebank: Phrase Structure Annotation of a Large Corpus [J]. Nat. Lang. Eng. 2005, 11(2): 207-238.

[100] Shen D., Lapata M. Using Semantic Roles to Improve Question Answering [C]. In: Proceedings of EMNLP-CoNLL'2007, 2007:12-21.

[101] Narayanan S., Harabagiu S. Question Answering based on Semantic Structures [C]. In: Proceedings of the COLING'2004, 2004: 184-191

[102] Surdeanu M., Harabagiu S., Williams J., Aarseth P. Using predicate-argument structures for Information Extraction [C]. In: Proceedings of ACL 2003.

[103] Melli G., Wang Y., Liu Y., Kashani M.M., Shi Z., Gu B., Sarkar A., Popowich F. Description of SQUASH, the SFU Question Answering Summary Handler for the DUC-2005 Summarization

Task [C]. In: Proceedings of DUC 2005.

[104] Fu G., Lukev K.K. Chinese Named Entity Recognition using Lexicalized HMMs[C]. In: Proceedings of KDD'2005. 2005. 19-25.

[105] Guo H., Zhu H., Guo Z., Zhang X., Su Z. Product Feature Categorization with Multilevel Latent Semantic Association[C]. In: Proceedings of CIKM'2009, 2009: 1087-1096.

[106] Zhai Z., Liu B., Xu H., Jia P. Grouping Product Features Using Semi-Supervised Learning with Soft-Constraints[C]. In: Proceeding of COLING'2010, 2010: 1272-1280.

[107] Pietra S.D., Pietra V.D., Mercer R.L., Roukos S. Adaptive Language Modeling using Minimum Discriminant Estimation [C]. In: Proceedings of the 1992 IEEE International Conference on Acoustics, Speech and Signal Processing, 1992: 633-636.

[108] Zhu X. J. Semi-supervised Learning Literature Survey [R]. Technique Report 1530. Computer Sciences, University of Wisconsin-Madison, 2005.6

[109] Miller D.J., Uyar H.S. A Mixture of Experts Classifier with Learning based on Both Labeled and Unlabeled Data [C]. In: Proceedings of Advances in Neural Information Processing Systems, 1997: 571-577.

[110] Nigam K., McCallum A., Thrun S., Mitchell T. Text Classification from Labeled and Unlabeled Documents using EM[J]. Machine Learning, 2000, 39(2-3): 103-134.

[111] Blum A., Mitchell T. Combining Labeled and Unlabeled Data with Co-training [C]. In: Proceedings of the Workshop on Computational Learning Theory, 1998.

[112] Blum A., Chawla S. Learning from Labeled and Unlabeled Data using Graph Mincuts [C]. In: Proceedings of the 18th International Conference on Machine Learning (ICML'2001), 2001: 19-26.

[113] Joachims T. Transductive Inference for Text Classification using Support Vector Machines [C]. In: Proceedings of the 16th International Conference on Machine Learning (ICML'1999), 1999: 200-209.

[114] Bennett K., Demiriz A. Semi-supervised Support Vector Machines [J]. Advances in Neural Information Processing Systems, 1999, 11: 368-374.

[115] Belkin M., Niyogi P. Semi-supervised Learning on Riemannian Manifolds [J]. Machine Learning, 2004, 56(1-3): 209-239.

[116] Abney S. Bootstrapping [C]. In: Proceedings of ACL'2002, 2002: 221-229.

[117] Neyman J. On the Two Different Aspects of the Representative Method: The Method of Stratified Sampling and the Method of Purposive Selection[J]. Journal of the Royal Statistical Society, 1934, 97(4): 558-625.

[118] Zhang Z. Weakly Supervised Relation Classification for Information Extraction[C]. In: Proceedings of ACM 13th conference on Information and Knowledge Management (CIKM'2004), 2004: 581-588.

[119] Thomas M.C., Thomas J.A. Elements of Information Theory, Second Edition[M]. John Wiley &

Sons, ISBN: 0-471-24195-4, 2006.

[120] Gamon M., Aue A., Corston-Oliver S., Ringger E. Pulse: Mining Customer Opinions from Free Text[C]. In: Proceedings of the 6th International Symposium on Intelligent Data Analysis. Lecture Notes in Computer Science. 2005: 121-132.

[121] Liu B., Hu M., Cheng J. Opinion Observer: Analyzing and Comparing Opinions on the Web[C]. In: Proceedings of WWW'2005. 2005: 342-351.

[122] Wilson T., Hoffmann P., Somasundaran S., Kessler J., Wiebe J., Choi Y., Cardie C., Riloff E., Patwardhan S. OpinionFinder: A System for Subjectivity Analysis[C]. In: Proceedings of HLT-EMNLP'2005, 2005: 34-35.

[123] 孟凡博, 蔡莲红, 陈斌, 吴鹏. 文本褒贬倾向判定系统的研究[J]. 小型微型计算机系统, 2009, 30(7): 1458-1462.

[124] Tsou B., Kwong O., Wong W.L., Tom L. Sentiment and Content Analysis of Chinese News Coverage[J]. International Journal of Computer Processing of Oriental Languages, 2005, 18(2): 171-183

[125] 姚天昉, 聂青阳, 李建超, 李林琳等. 一个用于汉语汽车评论的意见挖掘系统[A]. 中国中文信息学会二十五周年学术会议论文集[C]. 北京:清华大学出版社, 2006: 260-281.

作者在攻读博士学位期间完成的论文及科研工作

1. **施寒潇**, 周国栋, 钱培德. 基于分层抽样的情感对象属性自举分类研究. *计算机研究与发展*, 2013. (已投稿)
2. **Hanxiao Shi**, Guodong Zhou, Peide Qian, Guanglan Zhou. Research on Sentiment Analysis of Reviews Using Supervised Learning. *Intelligent Automation And Soft Computing*, 2013. (SCI 期刊, 已投稿)
3. **Hanxiao Shi**, Guodong Zhou, Peide Qian, Xiaojun Li. An Unsupervised Fine-grained Sentiment Analysis Model for Chinese Online Reviews. *Information-an International Interdisciplinary Journal*, 2012. 15(10): 4277-4294. (SCI 收录)
4. **Hanxiao Shi**, Xiaojun Li. A Sentiment Analysis Model for Hotel Reviews based on Supervised Learning, *Proceedings of 2011 International Conference on Machine Learning and Cybernetics (ICMLC'2011)*, 2011.7, pp. 950-954. [EI: 20114514487101]
5. 厉小军, 戴霖, **施寒潇**, 黄琦. 文本倾向性分析综述. *浙江大学学报(工学版)*, 2011, 45(7): 1167-1174. [EI: 20113414261070]
6. **施寒潇**，厉小军. 主观性句子情感倾向性分析方法的研究. *情报学报*, 2011, 30(5): 522-529.
7. Ming Wang, **Hanxiao Shi**. Research on Sentiment Analysis Technology and Polarity Computation of Sentiment Words, *Proceedings of the 2010 1st IEEE International Conference on Progress in Informatics and Computing (PIC'2010)*, 2010, pp. 331-334. [EI: 20110713666495]
8. **Hanxiao Shi**, Guodong Zhou, Peide Qian. An Attribute-based Sentiment Analysis System. *Information Technology Journal*, 2010, 9(8): 1607-1614. [EI: 20110313589868]
9. Xiaojun Li, Lin Dai, **Hanxiao Shi**. Opinion Mining of Camera Reviews Based on Sematic Role Labeling, *Proceedings of Seventh International Conference on Fuzzy Systems and Knowledge Discovery (FSKD'2010)*, 2010, pp. 2372-2375. [EI: 20104813422794]
10. **Hanxiao Shi**, Guodong Zhou, Peide Qian, Xiaojun Li. Semantic Role Labeling based on Dependency Tree with Multi-features, *Proceedings of The International Joint Conference on Bioinformatics, Systems Biology and Intelligent Computing (IJCBS'2009)*, 2009, pp. 584-587. [EI: 20094812511893]

攻读博士学位期间参与的科研项目：

1. 基于机器学习的高性能自适应信息抽取关键技术研究, 国家自然科学基金项目(#60673041).
2. 中文句法分析和语义角色标注的联合学习机制研究, 国家自然科学基金项目(#60970056)
3. 主观性文本的情感倾向性分析方法研究，浙江省自然科学基金重点项目(#Z1110551)
4. 文本倾向性分析研究及其在商品评价中的应用, 浙江省科技厅公益性技术应用研究计划项目(#2011C23075)

参加 NLP&CC’2012 中文微博情感分析评测：

1. 观点句识别评测结果, 第 3 名
2. 情感倾向性判断评测结果, 第 2 名

致　谢

值此论文完成之际，我衷心感谢所有关心与帮助过我的老师、同学、朋友和家人。

首先衷心感谢我的导师钱培德教授，多年来在学习上、工作上对我的关怀和指导。钱老师博大精深的学识、精益求精的风格、谦虚平易的作风以及诲人不倦的教育风范，一直感染并激励着我，并将使我终身受益。

衷心感谢周国栋教授在论文完成期间对我的指导和关心。无论在论文的选题、构思和资料的收集方面，还是在论文的研究方法以及成文定稿方面，我都得到了周老师的悉心教诲和无私帮助，特别是他那渊博的学术知识、敏锐的学术眼光以及严谨的治学精神将成为我今后努力的方向。

本文的顺利完成还与朱巧明教授的帮助分不开的。虽然平时朱老师的行政工作十分繁忙，但对我的学习和工作情况一直非常关心。朱老师在科研工作中敏锐的洞察力、独到的见解，以及富有创新精神的科学态度，是我永远学习的榜样。在此衷心感谢朱老师给予我的帮助。

感谢杨季文教授和吕强教授在我博士论文开题过程中提出问题，给出建议，使我受益匪浅。他们严谨的科研态度和渊博的学术知识也是我学习的榜样。

在论文撰写过程中，还得到了许多其他老师学长的帮助，特别感些钱龙华、孔芳、王红玲、李军辉，和他们的交流和探讨中，我学到了许多新方法，获得了许多新思路、新想法。在此我要衷心感谢他们。

另外，还要感谢协助我完成本文的几位硕士研究生、本科生，他们是贺姗姗、高田田、陈威等。是他们为我的研究实现了部分代码，帮助我验证了相关算法，并提供部分实验数据。

衷心感谢父母的养育之恩，祝您们健康长寿。特别感谢我的妻子金慧兰，没有她的支持，没有她全心全意的照顾家庭，我不可能完成今天的论文，非常非常感谢她的真心付出。另外还要感谢我的两位小朋友施天酝和施哲航，你们是我生活的乐趣，工作的动力，生命的意义，祝愿你们健康成长。

谨向所有给予关心和帮助我的老师、同学、朋友和家人再次表示衷心感谢。

最后，要感谢评阅、评议论文和出席论文答辨会的各位专家，感谢他们给予的意见和指导。

施寒潇

2012 年 9 月 9 日于信息楼

苏州大学

博士学位论文

苏州大学研究生部统一印制