

**工程硕士学位论文**

**基于依存句法分析的中文评价对象抽取和情感倾向性分析（摘要）**

**专业 名称**：计算机技术

**研究生姓名**：贾治中

**导师 姓名**：高志强 教授

* **背景**

观点或情感对于人类的一切活动和行为起着决定性的作用。人类的信仰、对现实的认知以及个体所做出的任何决定，从某种程度上，都取决于自身以及他人如何看待和评价所在的世界。正因为如此，每当人们需要做决定的时候，都会习惯性地先参考别人的观点。这不仅对于个体如此，对于大多数群体和组织而言更是如此。在现实生活中，个体消费者在购买某产品前往往想知道使用者对该产品的评价；在选举中，选民也希望在投票前知道其他人对于候选者的观点；商业组织必须清楚地了解顾客或公众对于他们的产品或服务的看法，才能更好地做出商业计划和营销决策。

随着互联网的飞速发展，人们一方面在浏览别人发表的评论，一方面也在不停地分享自己对于某些人或物的观点和情感。尤其是微博、论坛、购物平台、新闻网站等公众媒体的快速发展，评论性信息迅速膨胀，仅靠人工的方法难以处理海量信息的收集、分类和分析。与此同时，仅仅获取个体的评价信息是没有意义的，群体对于特定对象的褒贬能更好地为决策者提供更好的决策依据。因此，迫切需要计算机帮助用户快速获取和整理这些相关评价信息，情感分析（Sentiment Analysis）技术应运而生。观点以及与之相关联的概念包括情感、评价、态度和情绪等，都是情感分析和观点挖掘的主要研究对象。自从2000年开始，情感分析已经逐步成为自然语言处理领域中最为活跃的研究领域之一，并且被广泛应用在数据挖掘、网络挖掘、文本挖掘等领域。由于情感分析对于商业和社会的重要性，它已经从计算机科学领域扩展到管理科学和社会科学领域。

情感分析，又叫做观点挖掘，主要研究人们对于实体（如产品、服务、组织、个体、议题、事件、主题和它们的属性）的观点、情感、评价、评估、态度和情绪。根据任务的侧重点不同，情感分析还有其他不同的名称，例如观点抽取、情感抽取、主观性分析、影响分析、情绪分析、回顾分析等。

按照处理文本的粒度不同，情感分析主要分为文档级、句子级、实体和特征级。按照处理文本的类别不同，可分为基于新闻评论的情感分析和基于产品评论的情感分析。

目前情感分析有六个比较关键的任务，分别是实体抽取和分类、方面（特征）抽取和分类、观点持有者抽取和分类、时间抽取和标准化、实体方面情感分类、观点元组生成。基于这个框架的情感分析常常被称为基于方面或者基于特征的情感分析。

基于方面的情感分析又可分为两个子任务，分别是评价对象抽取和情感倾向性分析。文档级和句子级的情感分析在现实使用中往往不足以满足需求，因为它们无法识别出每个实体或属性及其对应的情感倾向性。即使假设每个文档评估一个单独的实体，一个含有关于某个实体正面评价的文档并不意味着作者对该实体的所有方面或属性都持有正面评价。对于更加完整的情感分析，需要抽取出所有的评价对象并判断对于该对象的评价是正面还是负面的。

评价对象抽取（Opinion Target Extraction）是文本情感分析中的重要研究任务之一，评价对象是指一段文本中所讨论或者描述的主题、实体或者实体的某些方面和属性等，具体表现为评论文本中修饰词语或评论词语所修饰的对象，如新闻评论中的某个事件、话题或者商品评论中的某个产品的组件（如“手机电池”）等。本文研究主要针对中文短文本中显式的评价对象，而非隐藏在文本背后的评价对象。

倾向性分析（Orientation Analysis），又称作情感分类，指的是抽取出文本中描述评价对象的情感词、极性词，并对其做情感倾向性分析（正面评价或负面评价），这里的情感词特指的是具有情感倾向性的词语。对于有些具体场景，比如购物网站上的商品评论，其评价对象可能就已经给定了，此时就不必再做评价对象的抽取工作，只需要判断该评价对象的情感倾向性。

* **中文评价对象抽取和情感倾向性分析的难点**

由于中文词语语义的多样性，在不同语境下可能会有不同的情感表述效果，加上网络用语的随意性，每个人表述情感的方式个性化，中文评价对象抽取和倾向性分析在实际应用中的效果很不理想，尤其是应用于微博平台上的情感分析。目前，中文评价对象抽取和倾向性分析的难点主要有以下几点：

1. 同一句中评价对象与观点词的距离较远，同时，二者之间存在其他干扰性的评价对象和观点词，导致二者无法被正确地关联和抽取出来。
2. 情感词存在，但是句中无明显的评价对象或只有相应的代词存在时，即评价对象缺失，如何根据上下文正确找到情感词关联的评价对象。
3. 句子含有情感词，但是并不包含情感倾向性，可能只是表达某种心理活动或非评价性的想法、愿望等。如“我希望小明能取得好成绩！”。
4. 句子中无明显的情感词存在，但依旧表达情感倾向性。如“他的店铺很有人气！”。
5. 转折词、否定词、情态动词等表达方式及其位置给倾向性分析带来不确定性。
6. 一些与情感词关联的名词，干扰了真正评价对象的抽取，例如“查韦斯病逝，又一个伟大的英雄离开了。”，该句中可能抽取出来的评价对象和观点词是[英雄，伟大]。
7. 情感分析基础工作的错误带来的干扰，例如中文分词、句法分析、语义分析过程中的标注错误可能给后续的分析带来影响，尤其是后两者的效果还不理想。
8. 英文情感词典已经有非常完备的SentiWordNet，而中文情感词典资源却质量不高，不够细致，要得到准确的效果不容易。
9. 大多数语料库都是基于商品评论的，如电脑、酒店等，标注细致的新闻评论、微博等语料库太少，而这些正是中文评论语境最为丰富的领域。
10. 新词不断产生，基本常识、上下文知识和各式各样的实体词，以及中文句式类型的复杂性都对中文评价对象的抽取带来很大的影响。

* **本文主要工作**

使用依存句法分析能够很好地揭示出评论语句中各成分之间的修饰关系或依赖关系。通常观点词都是和评价对象相互关联的，而多数情况下观点词可以根据词典或分词的标注结果而事先得知，所以利用依存关系分析可以通过观点词抽取出一些不频繁出现的评价对象。同时，通过依存句法关系分析，可以获得观点词的上下文特征（如转折关系、递进关系等），进而有利于分析出观点词的情感倾向。所以，基于依存句法分析的方法能够很好地对评价对象和观点词进行抽取和分析。然而，目前对于中文文本进行评价对象抽取和倾向性分析的精度还不是很好，因此，本文希望通过丰富情感词典、添加语义知识、改进评价对象搜索方法等手段来改善目前基于依存句法分析的中文评价对象抽取和倾向性分析的性能。

目前很多使用依存句法分析抽取评价对象的方法往往难以较为精准地抽取出观点词真正关联的评价对象，尤其是评价对象与观点词不在同一子句中的时候。针对该问题，本文在利用中文评论句子中词汇间依存关系进行情感分析的基础上，通过语义角色标注、添加抽取规则和搜索算法，以提高情感分析的性能。论文的主要工作包括：

1. 在现有词典的基础上，构建用于情感分析的情感词典，包括：正面情绪词典、负面情绪词典、正面评价词典、负面评价词典、观点引述词典、虚拟语气词典、转折词典、名词性情感词典等。这些词典主要用于处理评价句中无用成分或只是表达想法、意愿的非评价句对情感分析的干扰，提供语义规则和倾向性分析需要的词库支持。
2. 本文在依存句法分析的基础上，利用语义角色标注，添加了一系列的抽取规则进行情感分析。同时使用了定中短语替换通常的名词短语抽取出候选评价对象，用以提高评价对象和观点词的抽取精确度。这些规则主要考虑了中文语义知识、常用句式等对情感分析的影响。通过对比实验发现，这些规则和方法确实显著提高的情感要素抽取的性能。
3. 提出一种评价对象搜索方法，用于改善在只抽取出代词或句法关系中无评价对象的情况下，搜索上下文中真正的评价对象的精确度。该方法主要结合了词义和词语相似度计算算法，缩小了在上下文中潜在评价对象的搜索范围。通过实验验证，本文的评价对象搜索方法确实能够改善远距离的评价对象和评价词的抽取效果。

导师签字：

日期：