

复旦大学

硕士学位论文

文本情感倾向分析若干问题研究

姓名：朱嫣岚

申请学位级别：硕士

专业：计算机应用技术

指导教师：吴立德

20070520

摘要

文本的情感倾向分析即对于人们关注的一个问题或者对象,通过自动化工具,对网络中的评论性文章进行自动分析,判断文章的情感倾向,获得有价值的信息,并且归纳、整理出不同的态度和立场。

文本情感倾向性分析是目前自然语言处理领域研究的热点,在商业产品用户反馈分析、政府舆情分析、垃圾邮件过滤、信息安全和自动文摘等领域都有广泛的应用。情感倾向分析研究的涉及面很广,本文结合了实际的系统开发,着重对其中几项关键技术进行了深入的探讨。

本文首先对文本情感倾向分析的研究背景进行了详细的介绍。在该研究领域,国外起步较早,因此相关文献和研究成果比较多,而中文的情感倾向研究目前仍处于摸索阶段,相关参考文献较少,文章对此都做了概括性的介绍。

其次,文章探讨了通用领域的词汇情感倾向性计算方法。其主要思想是分别利用 HowNet 的语义相似度和语义相关场计算功能,计算词汇与人工选择的褒贬基准词的相似度,从而获取词汇的倾向性值。通过多次对比实验、参数测试,最终在常用词集上得到 80% 以上的正确率。

文章接着通过一个针对汽车领域的情感倾向分析系统,讨论了实际系统开发过程中,评价词、评价对象的本体构建技术。我们根据系统的实际需要,设计了领域本体的存储结构,并使用了人工与机器相结合的方式对本体信息进行搜集、整理以及扩充。最终的本体共包括 93 家生产厂商、219 个品牌、504 种汽车型号、7482 个同义词以及 4235 个包含情感倾向的词语信息。

最后,阐述了作者在文本情感倾向研究过程中的心得体会,并对研究成果进行了总结,对该领域的发展前景进行了展望。

关键词: 情感倾向; 态度分类; 知网; 本体

中图分类号: TP3

Abstract

The task for semantic orientation analysis is to automatically analyze reviews on the web, extract authors' opinion towards specific issue and summarize the overall situation.

The semantic orientation analysis based on text is a hot topic in current natural language processing field. It's widely applied in many fields such as user feedback analysis of commercial products, public opinion survey, junk mail filtering, information security, text summarization, etc. The research of semantic orientation is involved in wide aspects. This article combined with practical system development performs a deep research on emphasizing on some key techniques of it.

First, a detailed introduction of the research background of text semantic orientation analysis is given. In this field, although we have a lot of associated documents and research achievements in English, it is only budding in Chinese, so there is merely few reference documentation.

Second, a simple method of semantic orientation computation of Chinese words is introduced. The main idea of the method is to calculate the similarity between only a few seed words and the input word by HowNet's Concept Relevance Calculator (CRC) and Concept Similarity Measure (CSM). In this way, the orientation value can be obtained. Finally the performance is rather satisfactory among frequently used words, with the frequency-weighted accuracy of above 80%.

In the next part, a domain-dependent semantic orientation analysis system is introduced. A domain Ontology storage structure has been created according to the actual requirements of the system. The Ontology is built and enlarged using the method of manual tagging combined with automatic methods. Finally we constructed Ontology containing the information about 93 manufacturers, 219 brands, 504 models, 7482 synonyms and 4235 opinion words.

At last, the research achievements are concluded and an affirmative outlook is presented for the research of semantic orientation.

Key words: Semantic Orientation; Sentiment Classification; HowNet; Ontology

CLC: TP3

论文独创性声明

本论文是我个人在导师指导下进行的研究工作及取得的研究成果。论文中除了特别加以标注和致谢的地方外，不包含其他人或其它机构已经发表或撰写过的研究成果。其他同志对本研究的启发和所做的贡献均已在论文中作了明确的声明并表示了谢意。

作者签名： 朱嫣岚 日期： 2007. 6. 13

论文使用授权声明

本人完全了解复旦大学有关保留、使用学位论文的规定，即：学校有权保留送交论文的复印件，允许论文被查阅和借阅；学校可以公布论文的全部或部分内容，可以采用影印、缩印或其它复制手段保存论文。保密的论文在解密后遵守此规定。

作者签名： 朱嫣岚 导师签名： 吴立德 日期： 2007. 6. 13

第一章 引言

1.1. 文本情感倾向分析的任务定义

WEB 时代, 大量信息通过电子文档的形式存储在网络中。目前发展最为迅速的自由论坛, 博客, 讨论区, 在网络中为人们提供了一个信息共享、交流的互动平台。人们可以自由地在博客、论坛中发表对于事件、产品的评论。而权威媒体或者专业网站上, 由专业人士撰写的大量评论、产品测试报告等, 也具有很强的参考价值。然而, 由于网络信息的急剧膨胀, 信息量太大, 使得用户必须浏览很多此类站点才能找到所需要的信息。

文本的情感倾向分析任务即对于人们关注的一个问题或者对象, 通过自动化工具, 对网络中的评论性文章进行自动分析, 判断文章的情感倾向, 获得有价值的信息, 并且归纳、整理出不同的态度和立场。

1.2. 研究意义和应用领域

随着互联网技术在全球范围内的飞速发展, 人们越来越习惯于通过网络发布信息 and 获取知识。文本是承载信息最主要的媒介, 书籍、杂志、报刊、信件等, 是人们在日常生活中获取信息最重要的途径。在互联网中, 由于文本形式的信息对存储空间的要求不高、易于传输, 电子文档已经成为网络时代信息的主要载体之一。

网络环境下的信息的主要来源有: 新闻评论、BBS、聊天室、博客、聚合新闻(RSS)等, 尽管人们可以方便地通过网络下载资料和数据, 但面对海量的文档数据, 如果不加以整理和归纳, 人们无法从中获取任何有价值的信息。不经过处理的数据只是一堆没有意义的符号, 只有通过分析提炼, 信息才能真正为人所用。如何通过有效的手段获取信息, 是目前计算机领域研究的热点。

文本的情感倾向研究正是在这一背景下应运而生的。人们根据各自所关注的问题在 Web 上进行搜索, 通过阅读大量相关文献、介绍、评论, 逐渐形成自己的观点。由于网络资源过于庞大, 搜集—阅读—评价的工作变得十分繁琐和耗时。因此迫切地需要一套能够自动搜索、评价、归纳的系统替人们完成这一任务。困难是显而易见的, 计算机虽然具有极快的处理速度, 但如何使其能够自动完成对文本的语义倾向的推断, 仍然是个需要研究的课题。

文本情感倾向研究具有极大的实用价值。在商业领域中,使用文本情感倾向分析技术的企业商业智能系统能够为企业进行市场分析、市场调查、顾客反馈提供更多有价值的信息;在管理领域,政府舆情分析系统能够帮助领导者更快地了解群众对各类政策措施的反馈意见。

在自然语言处理领域,文本的情感倾向分析也为文本过滤、自动文摘等研究工作提供了新的思路和新的手段。

以下是一些具体的应用:

对评论的分类

一情感倾向分析可用于对评论文章的分类。通过分析,可以鉴别作者对商品、事件或政策所持的态度。通过不同的情感倾向分为支持和反对两大类。

区别反义词和近义词

一在自然语言处理中,对于反义词和近义词的区分一直是个比较困难的问题。传统的方式,我们一般通过衡量词语的共现性寻找具有语义关联的词语对。但是反义词和近义词都具有共现可能性高的特性,很难区分。而情感倾向分析能够对词语的褒贬含义进行分析,若词语含有相反的倾向,即为反义词,而具有相同倾向的词语对,则为近义词。这样就很容易将二者分开了。

为搜索引擎提供新的搜索条件

一当用户使用搜索引擎时,可以通过输入关键词,以及情感倾向约束条件,这样能够更精确地获得用户所需要的信息,进一步提高搜索效率。

文本摘要

一对于评论文章,自动文摘的生成,可将具有强烈褒贬倾向的语句作为关键句摘出,从而更好地保留原作者的观点和意见:一篇褒义的评论文章,我们可以抽取其情感倾向值最高的句子,作为摘要输出;而贬义的评论,则可直接提取文章中负面倾向最强烈的语句作为摘要输出。

对新闻组和邮件的过滤

一我们可以对语义倾向度量值设定一个合适的阈值,对于倾向值低于或高于阈值,也就是态度倾向过于偏激的文章进行过滤操作,或者可将倾向值赋予一定的权值,作为文本过滤中需要考虑的一个因素。该方法如果应用在网络新闻组的自动过滤中,有利于维护讨论秩序,缓和讨论气氛。

话题跟踪趋势分析

一通过对不同时期网络评论的情感倾向跟踪分析,我们可以获得倾向性随时间变化的情况,从而分析某个主题在不同时间段内,人们的关注程度以及主流观点的变化,进行趋势分析。通过对事件跨时间、跨空间的综合分析,发现突发事件并对反常现象进行报警。

聊天机器人

一情感倾向分析能够根据文本信息分析对话人目前的心理状态,聊天机器人可根据不同的心理状态选择话题并使用不同的语气。从而使对话显得更真实、更人性化。

1.3. 本文的工作

首先,本文从情感倾向分析系统的角度出发,对所涉及的关键技术进行了详细介绍。并对目前国内外的研究情况进行了调查分析。

其次,作为情感倾向分析的通用技术,我们在词语级倾向分析研究中,提出了基于 HowNet 的两种词汇倾向性计算的方法:基于语义相似度的方法和基于语义相关场的方法。实验表明,本文的方法在汉语常用词中的效果较好,词频加权后的判别准确率可达 80% 以上,具有一定的实用价值。

最后,我们对情感倾向信息的抽取进行了研究。在领域相关的情感倾向分析系统中,信息抽取主要是针对领域相关的评价对象和评价词。我们人工构建领域 Ontology,并通过同义词归并、模糊匹配、义原扩展等方法,对 Ontology 进行扩展。最终的本体共包括 93 家生产厂商、219 个品牌、504 种汽车型号、7482 个同义词以及 4235 个包含情感倾向的词语信息。为进一步进行评论的情感倾向分析打下良好的基础。

1.4. 文本组织结构

本文的组织结构如下:

在第二章中,将给出文本情感倾向分析的任务细分,简要的介绍各子任务以及解决问题的相关技术,同时介绍国内外同行在该方面的主要研究方法和进展。

在第三章中,将详细介绍基于 HowNet 的语义相似度和语义相关场计算功

能进行的词汇情感倾向判断的研究。并通过一系列的实验比较了不同计算方法的优劣。

在第四章中，将详细介绍在领域相关的情感倾向分析系统中的信息抽取相关技术。并通过汽车品牌信誉度分析系统为例，介绍该项技术在系统中的应用。

最后，总结全文并探讨情感倾向分析进一步的研究方向。

第二章 研究背景及相关工作

2.1. 任务细分

从开发系统的角度来说，一个完整的情感倾向分析系统，一般包括评论采集、情感倾向分析、可视化结果显示三个主要模块。其中，情感倾向分析模块是决定整个系统性能优劣的关键。

情感倾向分析所涉及的主要问题有：

1. 主观/客观性判断
2. 褒贬情感倾向分析（词语级/句子级/篇章级）
3. 情感倾向强度计算
4. 评价词与评价对象的关联关系
5. 相同对象不同表达形式的同义词归类
6. 领域相关知识库构建

上述的问题有些是相互联系、相辅相成的，而有些是独立的。例如，词语级情感倾向判断是主客观分类、句子级倾向判断、强度计算的基础，因此，必须在研究加以重视。在对大量语料进行分析之后，我们也发现，篇章级的情感倾向判断实用性不强，主要原因在于评论文章一般包含针对不同对象的多个情感倾向，以篇章为单位进行分析与实际不符，因此，实际应用中，篇章情感倾向判断只适用于小部分评论对象确定、评论态度单一的篇章。在领域相关的情感倾向分析中，构建领域相关知识库，对不同领域的专有名词的识别和同义词归并可大大提升“评价词—评价对象”关联关系抽取的效果，而评价词与评价对象的关联关系作为不可或缺的一部分，其抽取性能好坏直接影响系统的表现。

2.1.1. 主观/客观性判断

主客观分析即区分文本中的主观性表述和客观性表述。

一般的评论性文章，作者经常会采用客观事实加主观理解的方式进行表述，因此，主客观表述通常是交织在一起的。直接对这样的文本进行情感倾向分析，由于客观表述较多而主观论点稀疏，有可能造成分析过程受到干扰，导致分析结果不够准确。

文本的情感倾向，是表达作者态度的主观性陈述。因此，为了剔除客观性表述对情感倾向分析结果的影响，可首先进行文本的主观/客观性判断。主客观判断的任务，就是通过一定的算法，将文本中客观性的表述过滤，只保留主观性表述，将经过处理的文本再进行情感倾向分析。

2.1.2. 褒贬情感倾向分析

在包含作者的主观情感倾向的文章中，作者的情感状态可以是赞赏、批评、支持、反对；高兴、沮丧等。对于各类情感状态，我们可以笼统地将它们分为正面态度和负面态度两大类。褒贬性分析的任务就是区分文本中正面和负面的语言描述。

2.1.2.1. 词语级情感倾向判断

我们将具有情感倾向性的词语称为情感词或者评价词。这类词语对于分析作者观点具有非常关键的作用。例如，“这车太棒了！”，句中“太棒了”表达了作者对车的褒义评价，如果我们能够对于每个词语，给出褒贬性判断，那么也就很容易分析出作者的情感倾向。情感词不仅可以是形容词，动词、名词、副词都可以成为情感词。例如“他是国家的栋梁之材。”中“栋梁之材”就是具有褒义倾向的名词。

情感倾向分析中，情感词的褒贬判断就是对于一个给定的词语，使用算法进行褒贬倾向判断。这个过程类似于词性标注，而这里我们是对词语的褒贬性进行标注。

2.1.2.2. 句子级情感倾向判断

词语的情感倾向判断有其局限性，由于语言表述丰富多彩，表达形式多种多样，相同的词语所表现的褒贬倾向在不同的上下文环境下会产生变化。另外，随着时代的变化，词语含义的变化、新词的产生都将使得只针对词语的倾向判断工作的难度加大。

在文章中，句子是能够完整传达作者观点的基本单位。我们通过分析词，可以为句子的情感倾向判断提供帮助，但句型、句法分析、上下文分析有其不可替代的作用。例如，否定句式、反问句式等，尽管句子中褒义词出现较多，但通过否定词或者反问句表达，则整句的情感倾向将与其包含的倾向词相反，只分析词无法得到正确的结果。因此只有通过整句的分析，我们所获得的情感倾向判断结果才能更准确。

2.1.2.3. 篇章级情感倾向判断

篇章级情感倾向判断一般应用于主题单一、倾向一致且非常明确的文章中。该任务类似于文本分类问题，对于所有文章，分为褒、贬两大类，通过各种分类算法将文本归入其中某个类别。我们可以利用词语、句子级的倾向判断结果，通过统计的方法进行篇章级倾向分析。

但是由于篇章级情感倾向判断对于文章的题材、风格要求较为严格，所以应用领域受到极大限制。网络上的评论大多为自由文体，表述可能既有正面描述，又有负面描述，简单地对篇章进行褒贬判断并不符合实际情况。

2.1.3. 情感倾向强度计算

情感倾向强度就是在区分褒贬倾向的基础上，进一步细分，用量化的标准将情感倾向非常明显、比较明显和不太明显的情况表示出来。也就是说，对于某个对象的态度，希望能够通过打分的形式给出其强度信息。对于态度倾向不是简单的判断其褒贬态度，而是给出强度信息，由此作为情感倾向分析结果输出的排序依据。对于文本摘要或过滤系统来说，通过对强度设置阈值，进行更精确更实用的操作。

但如何使用较为有效且为公众所认可的方式进行强度计算，仍是一个值得研究的问题。

2.1.4. 评价词与评价对象的关联关系抽取

该任务即从文本中抽取“评价词—评价对象”关联关系。这里的评价词就是具有褒贬情感倾向的情感词。

很多的评论性文本，在一篇文章中、甚至一个句子中，会对不同的对象进行比较和评价。例如：“这款车的性能不错，但价格过高”，在例句中，有两个评价对象：“性能”和“价格”，针对“性能”有正面评价，针对“价格”则包含了负面的评价。因此，即使我们可以精确地抽取出具有情感倾向的评价词，并正确判断其情感倾向，在没有确定具体评价对象之前，我们仍然无法获得有用的信息。因为评价词是与评价对象相关联的，我们需要知道每个评价词所针对的不同对象，才能完整地理解文本传达的信息。评价词并不是孤立的，它只有与它所描述的评价对象关联在一起之后，才真正有其意义。

因此，我们可以通过上下文分析的方式，找出评价词所描述的对象。例如通过词语的距离信息进行简单分析，更深一步，可以通过对词语的语言特征进行归纳，确定目标词。

另外，如果能够确定评论的发表者，对于实际文本的分析将会有更大的帮助。例如，对于文章中的引用部分，通过分析，判断出该引用代表了第三方的观点而非直接代表作者的态度，则可在分析情感倾向时降低第三方言论这一干扰因素对作者观点提取造成的影响。

评价词和评价对象的关联关系抽取，是情感倾向分析中非常重要的一环。

2.1.5. 相同评价对象的不同表达形式同义词归并

对于相同的评价对象，每个作者的表述不尽相同，因此应做好评价对象的同义词归并工作。例如，“电脑”、“计算机”、“PC 机”，所描述的是同一事物，但不同作者可能有不同的表述。而面对网络中的海量文本，这样的情况对情感倾向分析带来的困难显而易见。抽取出大量的评价对象，而不进行归类整理，所得到的数据仍然不能为人所用。

对于描述同一对象的情感倾向分析应作分类汇总处理，由此提高最终倾向分析结果的实用性。

2.1.6. 领域相关知识库构建

不同的领域中，相同的表达方式所传递的情感信息却不尽相同。因此，情

感倾向分析系统目前主要应用在预定义的领域中。

领域相关知识对于情感倾向分析有着非常重要的作用，例如，评价对象的抽取。以商用系统开发为例，一般这类系统所关注的评价对象为商品，而商品名通常为专用名词，很难在现有的词典中找到，需要通过领域相关知识库的构建得到所需的信息。同样，在领域相关的评论中，评价词也经常使用专业术语，单单使用通用词典远远不够。为此，我们必须构建尽可能完善的领域相关知识库，以满足任务的需要。

2.2. 研究现状

从上世纪九十年代开始，文本情感倾向性研究在逐步受到人们的关注并迅速发展起来。传统的文本情感倾向研究较多地关注于文本中的情感倾向关键词，也就是说，人们认为特殊的包含情感倾向的词语能够最直接有效地表达作者的观点和态度。因此研究人员使用各种方法来利用情感词进行情感倾向的分析。

对于词语级的情感倾向研究，国外已有了一些研究成果。

Hatzivassiloglou 和 McKeown 早在 1997 年就通过分析大型语料库中形容词之间的关联词（and, or, but, either...or, neither...nor 等）的信息，以关联词为约束条件，利用对数线性回归模型来预测相关联的形容词的情感倾向，训练生成词汇间的同义或反义倾向连接图。在只考虑单个关联词信息的情况下，对形容词的情感倾向判断准确率达到 82%。当综合考虑训练语料中所有关联词、形容词信息后，使用聚类算法对测试集中的形容词按情感倾向进行聚类，准确率可达到 90%以上[1]。

Turney 2003 年在其论文[2]中提出了利用统计信息，衡量单词与预先定义的褒贬词集的关联关系，从而对单词进行情感倾向判断的新方法。其准确率在包含形容词、副词、名词、动词的完整测试集上达到 82.8%。

2005 年，Hiroya Takamura 使用物理学中的电极螺旋模型（spin model）进行词语情感倾向判别。将所有单词看作电子的集合，而将词语情感的正负极看作电子的正负极性，通过平均场（mean field）近似计算每个电子的平均极性。由此对词语的情感极进行抽取。该方法在 500 个单词的测试集上获得 85% 的准确率[3]。

Subasic 和 Huettner 通过人工为每个单词标注情感倾向级别的方式在词汇集上进行情感词分类[4]。

但是,所有基于情感关键词的分析方式都存在以下的一些问题:1)词的情感倾向的在不同上下文中的不确定性。2)在没有情感词的句子中如何识别情感倾向。3)无法利用语义和句法分析。

由于以上问题,除了利用情感关键词进行情感倾向分析,一些研究人员开始尝试使用其他文本信息,如作者的写作意图、句型、段落结构、上下文等,在该研究领域进行新的探索。

Boucoulalas 和 Zhe 将语法分析器应用到识别评价词—评价对象关联关系的任务中去[5][6]。Tao 和 Tan 使用情感功能词(emotional function words)代替情感关键词(emotional keywords)来进行情感倾向分析[7]。Litman 和 Forbes 将对话中的各类信息,包括语言特征、对话方式、对话人发言顺序等整合在一起,作为对话系统情感倾向分析的依据[8]。

Chung-Hsien Wu 则从语义的角度对句子级情感倾向分析进行了新的探索和尝试, Wu 在其论文[9]中,介绍了基于语言学规则的情感倾向识别方法。作者认为,通常使用情感倾向关键词进行情感倾向分析的方法,忽略了人类语言多样性的问题,同样的含义,用不同的方式进行表述,表达相同的情感倾向,但所用的词汇会有很大的区别,表述中甚至不出现表示情感倾向的关键词。作者提出了新的方法,该方法不使用情感倾向关键词,而是从语义角度出发,由研究人员人工定义了一些推理规则,将文本中的句子与这些推理规则使用 SMM 模型进行比较分析,最终得到该句子的情感倾向。

篇章的语义倾向判别也可被看作一个褒贬的分类问题,因此,文本分类中的方法同样被应用到了语义倾向判别研究中。2002 年,Turney 在其论文[10]中介绍了基于语义倾向的非监督文本分类方法。根据褒贬含义的倾向信息对评论性文章进行分类,其分类结果更符合人们对评论性文章分类的实际需求。在 Epinions 上的 410 篇评论性文章中,利用 Turney 的算法对褒贬倾向进行分类,正确率达到 74%。

同年, Pang 等人使用标准的机器学习技术做了同样文本分类工作[11]。比较了简单贝叶斯、最大熵、SVM 方法在语义倾向文本分类上的效果。SVM 的分类准确率最高达到约 80%,为几种方法中分类效果最好的。之后, Pang 提出了对语义倾向分类结果进行后续处理,进一步将分类结果按照语义倾向强度进行细分的方法,并通过一系列的实验证明了其方法的可行性[12]。

在主客观性判别方面, **Wiebe** 使用学习算法从语料库中提取主观性形容词[14]。**Riloff** 等则实验了新的方法提取主观性名词[16]。**Theresa Wilson** 在其论文[17]中, 利用了统计的方法, 使用大量特征进行训练, 完成主客观分析并对评论按情感倾向强度进行分类。

语义关联关系抽取方面, **Peter D. Turney** 也做出了重要贡献。他在论文[18]中将自然语言研究中的相似性研究分为两大类: 词语间的属性相似 (**Attributional Similarity**) 和词语对之间的关系相似 (**Relational Similarity**)。属性相似即通常我们所说的近义词。关系相似即存在相近关联关系的词语对。对于具有一定关联关系的词语对, 通过潜在关联分析 (**Latent Relational Analysis**), 对词语进行近义词扩充、共现频率过滤、模板抽取、关系矩阵生成、奇异值分解等处理, 最终从文本中提取出具有相似关联关系的词语对。实验结果表明, **Turney** 将该方法运用在 374 个类比问题集上, 用以抽取正确的类比关系, 其 **F** 值达到 56.5%, 接近人类真实答题水平 (**F** 值为 57%)。

在对情感状态分类问题上, **Deville et al.** 通过计算情感词和情感状态的条件概率, 找到了合理的情感状态分类法则[19][20]。

近年来, 陆续出现了一些利用语义倾向分析技术开发的商业智能系统, 例如, **NEC** 公司的 **Kusha** 等人所开发的 **ReviewSeer**, 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[21]。

又如, **Bing Liu** 在其论文[22]中介绍了商用产品信息反馈系统 **Opinion Observer**, 利用网络上丰富的顾客评论资源, 针对电子产品领域的评论, 通过对文本信息的细致挖掘, 得到了产品不同属性的评价信息, 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。与一般的情感倾向分析方式不同的是, 作为适用于电子产品领域的情感倾向分析系统, **Opinion Observer** 更有针对性地考虑了数码产品领域的专用表达方式以及该类产品的属性特征, 从而使得用户不仅从简单的产品评论的比较中获得信息, 更可获得针对不同属性特征的评论分析结果。这对于生产商、用户来说都是更具有实际应用价值的工作。

另外, 如站点: **opinmind.com**, 从博客站点搜集关于各种目标的评论 (例如人物, 地点, 事件, 产品等)。用户输入感兴趣的目标, 系统提供对该目标的总体态度, 返回相关评论, 并按态度 (肯定或者否定) 分组。为网络用户提供更方便有效的评论信息检索方式[23]。

情感倾向分析在英文领域的应用已初见端倪。而基于中文的情感倾向分析研究目前仍然处于起步阶段。

北京大学的王治敏、朱学锋、俞士汶等，利用现代汉语语法信息词典，从计算机实现角度对词语所表现出来的情感倾向进行了简单分类，而且依据对大规模语料的统计分析，发现词语的上下文环境对词语的情感判定起到关键作用[24]。

香港城市大学的邹嘉彦从语言学角度对大量真实语料进行了分析研究[25]，进一步强调了情感倾向分析的重要性。基于文本的情感倾向分析研究有其广阔的应用前景，它作为一种有效的信息获取手段，已成为当前计算机领域研究的热点。

第三章 词汇倾向分析

词汇倾向分析是文本情感倾向分析中的基础工作。识别出文本中的情感词，对于帮助理解文本的态度倾向起到至关重要的作用。

3.1. 任务背景等介绍

所谓词汇的情感倾向，即对于词汇的褒贬程度计算出一个度量值。为了便于统计和比较，目前比较常用的做法是将度量值规定为位于 ± 1 之间的实数。当度量值高于某阈值时，判别为褒义倾向；反之，则判为贬义倾向。此外，我们可以通过对篇章中词汇的情感倾向值求平均的方式，获得篇章的情感倾向。因此，对词汇的情感倾向计算是此类研究中的关键工作。

单词的情感倾向判别是篇章情感倾向研究的基础。早在 1997 年，Hatzivassiloglou 和 McKeown 就尝试使用监督学习的方法对词语进行情感倾向判别，通过对训练语料的学习进行情感倾向判别，准确率约 82%，在加入篇章中形容词之间的关联词信息后，准确率提升到 90%左右[1]。在 2003 年，Turney 在其论文[2]中提出了利用统计信息对单词进行情感倾向判断的新方法。其准确率在包含形容词、副词、名词、动词的完整测试集上达到 82.8%。研究人员尝试使用统计、语义分析等各种方法对词汇的情感倾向进行分析，在英文领域已经积累了不少的经验，但在中文领域，这方面的研究才刚刚起步。

3.2. HowNet 介绍

本章的研究工作侧重于词汇的情感倾向性判别。利用 HowNet[27]提供的语义相似度和语义相关场计算功能，尝试了一些词语情感倾向判别的方法。主要做了以下工作：通过计算词语间的相似程度，对词语的褒贬倾向按照一定的计算法则进行赋值。根据所得的情感倾向度量值判别其褒贬倾向。

知网（英文名称为 HowNet）是一个以汉语和英语的词语所代表的概念为描述对象，以揭示概念与概念之间以及概念所具有的属性之间的关系为基本内容的常识知识库。

由于汉语中“词”的含义非常复杂，往往一个词在不同的语境中会表达不同的语义。因此，在 HowNet 中，使用两个更基本的单位：“概念”与“义原”

来描述汉语中的词。

“概念”是对词汇语义的一种描述。每一个词可以表达为几个概念。“概念”是用一种“知识表示语言”来描述的，这种“知识表示语言”所用的“词汇”叫做“义原”。“义原”，可以被认为是汉语中最基本的、不易于再分隔的最小语义单位。**HowNet** 试图用一系列的“义原”来对每一个“概念”进行描述。

简单地说，即一个词语包含多个概念，在 **HowNet** 系统中每个概念对应一个义项，而每个义项则通过一些义原来描述。词语与概念是一对多关系，概念与义项是一一对应关系。

例如：在 **HowNet** 系统中，输入词语“洗衣”，可获得多个义项，即该词语的多个概念，选择其中一个义项，可获得该义项的描述： $DEF=\{wash|洗涤:patient=\{clothing|衣物\}\}$ ，这里冒号前面的义原“wash|洗涤”，是对该义项的描述，冒号后面的 $patient=\{clothing|衣物\}$ 是对冒号前面的“wash|洗涤”的补充描述。

本章所介绍的算法和实验分别使用了 **HowNet** 提供的语义相似度和语义相关场的计算功能。以下将简单介绍 **HowNet** 语义相似度和语义相关场计算的原理。

3.2.1. HowNet 语义相似度计算

词语的相似度是一个相当主观的概念，对于不同的应用，其含义也会有所区别，因此我们必须基于不同的应用来定义相似度的概念。在信息检索的背景下，相似度主要反映文本或者用户查询在意义上的符合程度。

而 **HowNet** 中定义的词语相似度，则是基于机器翻译中的应用，反映两个词语在不同上下文中可以互相替换使用而不改变文本的句法语义结构的程度。如果在不同上下文中可以互相替换且不改变文本句法语义结构的可能性越大，二者的相似度就越高，否则相似度就越低。一个词语与其本身的相似度为 1，如果两个词语在任何上下文中都不能替换，那么它们的相似度应为 0。

HowNet 的相似度计算方法，是一种基于世界知识（Ontology）的方法。

我们已经知道，**HowNet** 中，一个义项代表一个完整的概念。而义原则是描述义项的基本单位。因此词语的相似度计算首要考虑的是义原之间的语义相似度。

HowNet 义原与义原之间存在着非常复杂的关系。在 HowNet 中，一共描述了义原之间的八种关系：上下位关系、同义关系、反义关系、对义关系、属性-宿主关系、部件-整体关系、材料-成品关系、事件-角色关系。可以看出，义原之间组成的是一个复杂的网状结构，而不是一个单纯的树状结构。不过，义原关系中最重要还是的上下位关系。根据义原的上下位关系，所有的“基本义原”组成了一个义原层次体系（如图 1）。这个义原层次体系是一个树状结构，

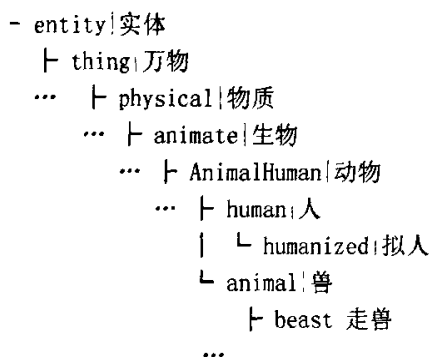


图 1 树状的义原层次结构

这也是 HowNet 进行语义相似度计算的基础。

由于所有的义原根据上下位关系构成了一个树状的义原层次体系，这里采用简单的通过语义距离计算相似度的办法。假设两个义原在这个层次体系中的路径距离为 d ，根据公式(1)，我们可以得到这两个义原之间的语义距离：

$$S(p_1, p_2) = \frac{\alpha}{d + \alpha} \quad \dots\dots (1)$$

其中 p_1 和 p_2 表示两个义原（primitive）， d 是 p_1 和 p_2 在义原层次体系中的路径长度，是一个正整数。 α 是一个可调节的参数。 α 的含义是：当相似度为 0.5 时的词语距离值。用这种方法计算义原相似度的时候，只利用了义原的上下位关系。

HowNet 的语义相似度计算是基于概念（义项）的，利用了描述概念的义原之间的语义关系。具体的计算步骤如下：

首先，对于给定的两个义项，如果它们的义原描述中具有同义关系，则相似度赋值为 1.000000，即 $S(c_1, c_2) = 1.000000$ 。

第二步，如果给定义项，在义原描述中属于同类，则相似度直接赋值为 0.950000，即 $S(c_1, c_2) = 0.950000$ 。

第三步，如果两个义项的义原描述具有反义或者对义关系，则将相似度赋值

为 0.000000, 即 $S(c1, c2) = 0.000000$ 。

第四步, 对于语义动态角色为“对应之事件”(CoEvent)或“之材料”(MaterialOf)的义原, 只抽取 CoEvent / MaterialOf 中的那部分义原进行相似度计算。例如, 词语“慈善事业”的义项为 DEF={affairs|事务:CoEvent={TakeCare|照料:manner={voluntary|自愿}}{help|帮助:manner={voluntary|自愿}}}, 当需要计算“慈善事业”与其他词相似度时, 只考虑 {TakeCare|照料:manner={voluntary|自愿}}{help|帮助:manner={voluntary|自愿}} 这部分义原。

第五步, 对于不能满足 1~4 的词语。根据具体情况, 使用公式 $S(p_1, p_2) = \frac{\alpha}{d + \alpha}$ 进行义原间的相似度计算, 综合结果并得到义项的相似度。具体过程在这里不做详细的介绍。

HowNet 语义相似度的计算功能是根据刘群的论文[29]中的原理编写的词汇语义相似度计算程序, 实现了义项之间语义相似度的计算。

图一为 HowNet 语义相似度功能界面。通过输入两个词语并分别选取确切的义项, 在结果显示框中即可得到相似度的数值。举个简单的例子, 输入词语“好”、“差”, 并分别选择相应的义项为“{HighQuality | 优质}”、“{unqualified | 不合格}”, 得到的输出结果为 0.021053, 即这两个词语在相应义项下的相似度。语义相似度主要反映的是词语含义的相似程度。HowNet 相似度值被定义为 0 到 1 之间的一个实数。至于词与词之间的语义相似度, 我们使用了词语两两义项相似度的最大值, 希望最大限度的保留词语之间相似性信息。

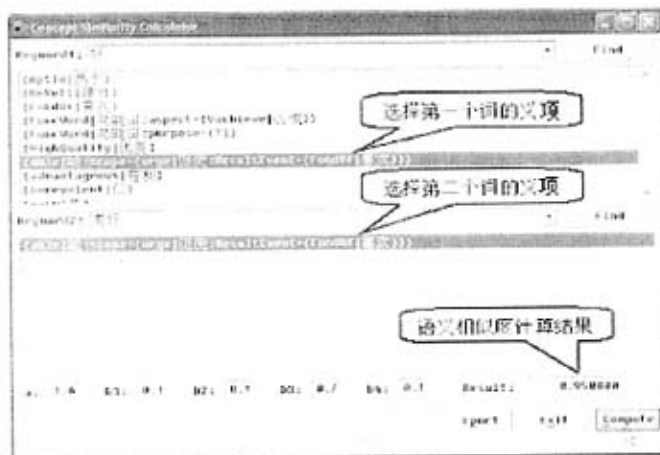


图 2. HowNet 语义相似度计算功能界面

3.2.2. HowNet 语义相关场计算

度量两个词语关系的另一个重要指标是词语的相关性。

与词语相似性一样，我们可以将词语相关性量化为一个 $[0,1]$ 之间的实数。

所谓语义相关场，不同于语义相似度，反映的是两个词语互相关联的程度。可以用这两个词语在同一个语境中共现的可能性来衡量。因此相关的词语语义上可能相近，也可能相反。

这里必须强调的是，词语的相似性和词语相关性虽然有着非常密切的联系，但它们是两个不同的概念。

如果两个词语非常相似，那么这两个词语与其他词语的相关性也会非常接近。反之，如果两个词语与其他词语的相关性特点很接近，那么这两个词一般相似程度也很高。

但是，需要注意的是，根据词语相关性定义，当两个词语的相似性很低时，由于它们经常在同一语境中出现，它们的相关性可能会很高。例如：“医生”和“疾病”这两个词语，从语义角度考虑，它们的含义不同，根据相似性定义，它们在不同上下文中几乎无法相互替代，相似性很低。但根据相关性定义，这两个词语在同一语境中出现频率非常高，因此相关性非常高。可以这么认为，词语相似性反映的是词语之间的聚合特点，而词语相关性反映的是词语之间的组合特点。

HowNet 中语义相关场的获取仍是以概念（义项）为单位的，具体步骤如下：

首先，根据输入词语的指定义项，在所有词语的义项描述中进行模糊搜索。如果某个词语的义项描述中，有部分义原与已知的义项中的部分义原匹配，那么就把它加入相关词集合中。这一步主要获得输入词的同义词、同类词、上位词等。例如：对于输入词“编程”，我们根据它的义项 $DEF=\{compile|编辑;ContentProduct=\{software|软件\}\}$ ，通过模糊匹配的方式进行搜索，获得具有相似义项描述的词语有：程序、ALGOL 语言、FORTRAN 语言、JAVA、LISP 语言、PROLOG 语言、编程语言、程序设计、程序员等。

第二步，根据指定义项的义原描述，获得输入词的反义词、对义词。例如：“贫民”，根据它的义项 $DEF=\{human|人;modifier=\{poor|穷\}\}$ ，我们通过寻找具有反义或者对义关系的义原描述，从而获得一组反义词或者对义词，如：亿

万富翁、财主、百万富翁、财阀等。

第三步，通过 domain、RelateTo 标签，获得输入词属于同一领域的词语和 相关词。例如：“跳球”的义项为 DEF={phenomena| 现象:domain={basketball|篮球}}, 根据其中的领域信息 domain={basketball|篮球}, 可以获得同领域的一组词: 篮球、篮坛、职业篮球、持球等。而根据词语“主页”的义项 DEF={readings|读物:RelateTo={internet|因特网}}中的“相关”(RelateTo) 属性, 可以获得一组相关词, 如: 超文本传输协议、上传、上载、通信协议等。

第四步, 对以上三步获得的词集, 根据实体、事件、属性、属性值分成四个大类。

图 3 为 HowNet 语义相关场计算功能界面。通过输入一个词语并选取词语的某一义项, 可以获得与该义项语义相关的一组词, 例如, 对于“好”这个词, 若选择其义项为“{ HighQuality | 优质}”, 即可获得一组语义相关的词语, 包括“优良”、“优质”、“良好”、“佳境”、“完好”等。同时语义相关场的范围(即语义相关词语的数量)可通过参数调整(分三级, 从小到大分别是 Rank1~Rank3)。



图 3. HowNet 语义相关场功能界面

3.3. 实验原理

我们希望为每个单词赋予一个情感倾向的度量值。其大小由这个单词与基准词的语义关联的紧密程度有关。基准词，在这里指褒贬态度非常明显、强烈，具有代表性的词语。与褒义基准词联系越紧密，则词语的褒义倾向越强烈。与贬义基准词联系越紧密，则词语贬义倾向越明显。

基于这样的设想，设计如下实验：

假设共有 k 对基准词，每对基准词包括一个褒义词和一个贬义词。褒义基准词表示为 key_p ，贬义基准词表示为 key_n ，单词 w 的情感倾向值用 $Orientation(w)$ 表示，以 0 作为默认阈值，最终倾向值大于阈值为褒义，小于阈值为贬义。 $Orientation(w)$ 数值大小代表单词 w 褒贬强烈程度。

单词 w 的情感倾向值计算公式如下

$$Orientation(w) = \sum_{i=1}^k Similarity(key_p_i, w) - \sum_{j=1}^k Similarity(key_n_j, w) \quad (2)$$

我们的试验中，公式（1）中的 $Similarity(key, w)$ 的计算分别采用了 HowNet 语义相似度和语义相关场两种方法。

1. 利用 HowNet 语义相似度计算单词的情感倾向值

$$\text{公式（2）中的 } Similarity(key, w) = Sim(key, w) \quad (3)$$

利用 HowNet 计算两个单词义项之间的相似度，取两个单词各个义项之间相似度的最大值作为 $Sim(key, w)$ 。

2. 利用 HowNet 语义相关场计算单词的情感倾向值

$$\text{公式（2）中的 } Similarity(key, w) = \frac{|Relevance(w) \cap Relevance(key)|}{|Relevance(w) \cup Relevance(key)|} \quad (4)$$

HowNet 的语义相关场的获得是以义项为单位，即给出一个义项，可获得属于其相关场的单词集合。这里所说的 $Relevance(w)$ 表示单词 w 所有义项下的语义相关场词语的并集。式（4）的分子和分母分别表示两个词语相关场交集和并集的元素个数。

3.4. 实验

3.4.1. 测试词表

度量情感倾向判别效果的方法有两种，一是直接判断其算法对单词的褒贬倾向判断的准确率，二是通过情感倾向判别为某个实际应用带来的性能提高来衡量。这里使用前一种较直观的方式对实验效果进行评价。

情感倾向实验所使用的测试集共有 3 组。测试集 1~测试集 3。

测试集编号	褒义词数	贬义词数	总词数
1	3146	3299	6445
2	1370	1425	2795
3	69	211	280

表 1. 测试集

测试集 1 使用了 HowNet 第一版中文词表中标注“良”(褒义),“莠”(贬义)属性的词汇。总共选用 6445 词。其中褒义词 3146 个,贬义词 3299 个。

在实验过程中,我们发现,测试集 1 中包含这样一些词语,其义原集合中某一义项虽然被标注为褒义或贬义,但该义项作为该词的一个特殊义项并不常用;或者该词语在不同语境下,或为褒义,或为贬义,并不能简单地将其判断为褒义词或贬义词。例如:词语“好看”,在描述事物时,可作褒义,如“这花真好看”,而在“要你好看”这样的语句中,显然带有强烈的贬义。为了使实验结果更能说明问题,更具有说服力,我们从测试集 1 中人工挑选了 2795 个褒贬含义较明确的词语作为测试集 2。

为了测试我们的方法在常用词中的效果,我们把测试集 2 中的词语按 Google 搜索返回 Hits 数(截止至 2004 年 8 月 31 日),即它们在 Web 上的词频降序排列,选取词频最高的 280 个词语,作为测试集 3。

表 1 中列出了各个测试集合的具体信息。

3.4.2. 基准词选择

在我们的方法中,计算情感倾向度必须首先选择一组褒贬基准词对,根据被测试词与这组基准词的语义关系紧密程度,通过计算得到情感倾向值。显然,所选择的基准词必须是具有强烈褒贬倾向,并且具有代表性的词语。基于以上

原因，我们选择了测试集 2 中的词语作为基准词对的候选集合，根据词频对词语进行降序排列，分别从褒贬词中选取出现频率最高的作为褒贬基准词。总共选择了褒贬词对 3 组，分别进行实验比较。基准词对 1 由词频最高的前 10 个褒义词和前 10 个贬义词组成，基准词对 2 选取了词频最高的前 20 对褒贬基准词，基准词对 3 选取了词频最高的前 40 对褒贬基准词。表 2 给出了基准词对 3 的全部词汇，而基准词对 1 和基准词对 2 显然都是它的子集。本章中所有实验结果都是从测试集中扣除了这 40 组基准词后得到的。

褒义基准词：

健康	安全	天下第一	美丽	超级	保险	卫生	天使	英雄	精选
快乐	权威	稳定	优秀	高级	精英	最好	最佳	幸福	容易
高手	文明	积极	著名	漂亮	完美	简单	和平	开通	真实
先进	便宜	优质	欢乐	美好	良好	不错	出色	成熟	完善

贬义基准词：

不合作	黑客	疯狂	错误	事故	非法	失败	背后	麻烦	不良
病人	恶意	色情	暴力	黄色	浪费	落后	漏洞	有害	讨厌
白负	不安	魔鬼	花样	野蛮	陷阱	不当	腐败	无情	失误
浮秽	流氓	虚假	残酷	变态	脆弱	不合格	愚人	恶劣	恶魔

表 2. 40 组褒贬基准词

3.4.3. 实验概述

实验的目的是为每个测试集中的单词赋予情感倾向值，通过倾向值来度量其褒贬程度。在以下实验中，默认使用 0 为阈值，即倾向值大于 0 则判断为褒义，小于 0 则判断为贬义。情感倾向判别准确率=判别正确的词数/测试集总词数，我们以此来衡量算法效果。

情感倾向值的计算使用了两种方法，分别利用了 HowNet 的语义相似度和语义相关场的计算功能。具体计算公式在 2.2 中已有详细介绍。

首先，对于这两种计算方法的效果进行了实验，在此，我们使用随机赋值的方法作为基准参与比较。为了避免避免获得的相关场集合较小而导致 $Relevance(w) \cap Relevance(key)$ 交集为空的情况，因此，我们选用可调参数 Rank=3，获得单词的最大的语义相关场。

实验使用基准词对 3 (40 对褒贬基准词), 实验结果如下表:

测试集编号	1	2	3
HowNet 相似度计算	68.04%	68.32%	85.50%
HowNet 相关场计算	62.03%	63.43%	75.00%
随机赋值 (baseline)	49.74%	51.16%	45.00%

表 3. Similarity(w) 在不同计算方法下的情感倾向判别准确率 (使用 40 对基准词)

表 3 显示了在 3 个测试集上, 两种不同的计算方法对情感倾向判别中的效果。我们发现, 在同一测试集上, 使用相同的基准词对, 基于 HowNet 语义相似度的计算方法比基于语义相关场的计算方法准确率高。其原因主要是由于, 语义相关场的定义中, 所谓相关, 既包括了褒贬倾向相近的词语, 同时又包括了褒贬倾向相反的一些词语, 例如: “好” 这个词, 选择其义原为 “优质”, 所获得的相关场词集中, 既包括了 “优良”、“良好” 这类语义相近的褒义词, 又包含了 “坏”、“劣质” 这些语义相反的贬义词。由于这个原因, 导致了语义相关场并不能将词语间褒贬的相近程度很好地反映出来。而语义相似度恰恰能在一定程度上反映词语间的语义倾向的一致性, 因此实验效果较好。

为了更好地反映本章所提出的情感倾向算法在实际应用中的效果, 我们把词语使用频率作为衡量算法性能的一个指标。我们对测试集中的单词按 Google 搜索返回 Hits 数进行了统计。由此作为该单词使用频率的度量指标。

测试结果的准确率根据单词使用频率加权计算。准确率 = $\frac{\sum result(rw_j)}{\sum result(w_j)}$, 公式中

result(word) 表示词语 word 的词频, rw 代表褒贬倾向判断正确的单词, w 表示测试集中的任意单词。

下表将经过词频加权的实验结果与初始实验结果在测试集 2 上进行了比较。情感倾向值的计算使用基于 HowNet 语义相似度的计算方法。

基准词对数	10	20	40
未经加权的结果	55.73%	59.48%	68.32%
经过词频加权后的结果	60.77%	64.22%	78.58%

表 4. 对准确率进行词频加权后的实验结果 (测试集 2, 使用 HowNet 语义相似度计算)

由表4可以很明显地看出,根据单词词频加权的准确率较初始结果好得多。在不同基准词对下,其准确率高出 5%~10%。由此说明,我们的算法在常用词中的判别效果较低频词好,因此该方法更具有实际应用的价值。

另外,基准词对数目的大小对实验效果的影响也是我们所关心的一个问题。在测试集2中,我们挑选了使用频率最高的一部分词,作为测试集3,而基准词仍然根据词频选取前1对、4对、5对、10对、15对、20对、30对、40对褒贬词进行测试。词语相似度的计算仍使用基于 HowNet 语义相似度的方法。实验结果如下:

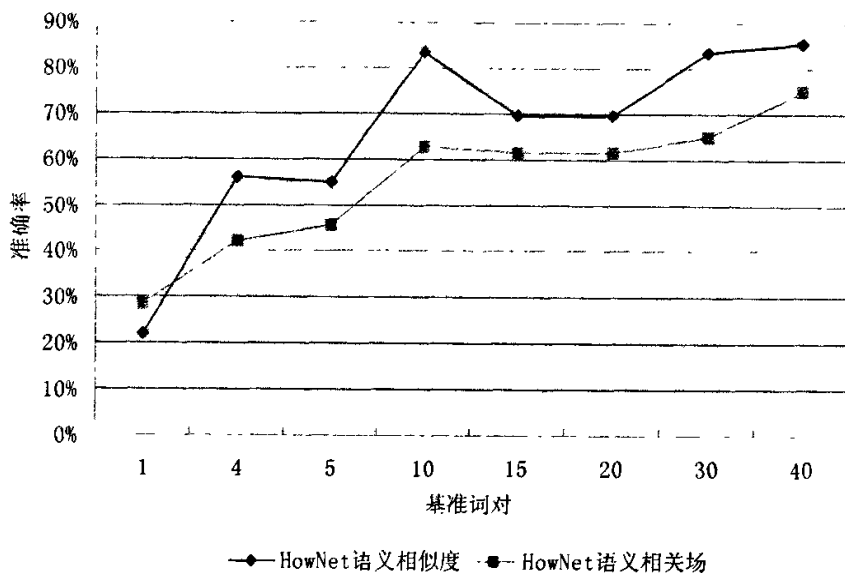


图 4. 基准词数量对实验效果的影响 (测试集 3)

从上图可以看到,情感倾向判别的准确率大致随基准词的数量增加而上升,也就是说,给出的褒贬“范例”一基准词越多,算法的效果就越好。同时,与表3列出的结果相一致的,在测试集3上仍能得到相似度计算方法较相关场计算方法更优的结论,但随着基准词的数量增加,两种方法的差异逐渐减小,

以上的实验都将0作为默认阈值,计算所得的倾向值大于0的判为褒义词,小于0的则判为贬义词。考虑到基准词选择中,可能出现的偏向,我们尝试调整阈值,对实验结果数据进行更深入分析。

阈值调整的范围规定在-0.05~0.05之间,调整步长为0.0005。我们使用枚举的方法,在预定义的范围内,尝试所有可能的临界值,希望找到一个更合

理的临界值，它应该在各个测试集下都表现稳定，并且对于词语的情感倾向判别准确率应高于使用默认值时的情况。

图 5 是对测试集 1 和测试集 2 的实验结果数据进行阈值调整后取得的效果。这里所使用的实验结果数据都是利用 HowNet 语义相似度方法，使用 40 对基准词进行计算所得到的结果。

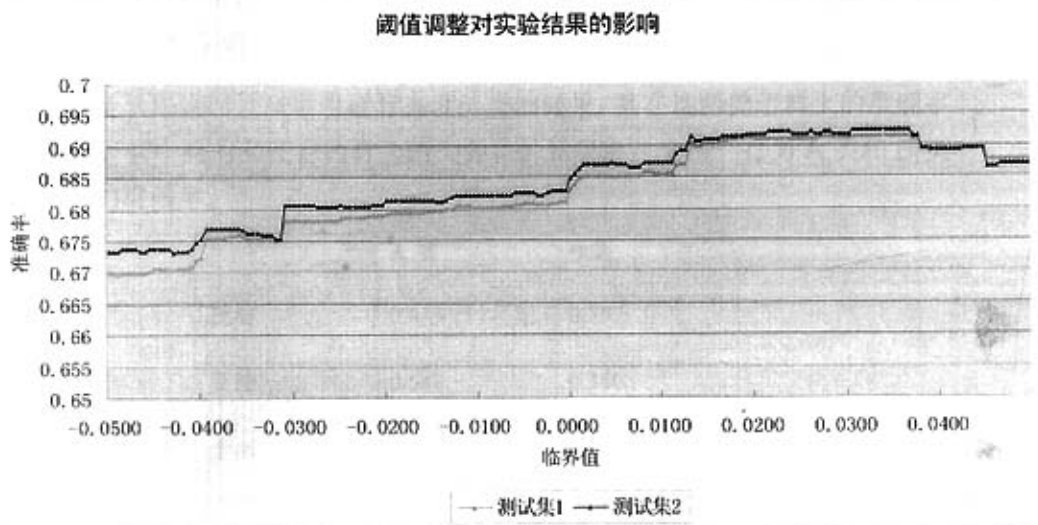


图 5. 阈值调整对实验效果的影响（测试集 3）

根据图 5 我们发现，判别褒贬倾向的临界值虽然较原始临界值 0 有偏大的倾向，但即使在最好的情况下，阈值调整仅仅能使判别准确率比临界值为 0 时的情况提高不到 1%，对总体性能的改善并没有起到显著的作用。而且在各个测试集上，产生最好结果的阈值也并不稳定。因此我们认为，阈值的调整对准确率的提高并不具有显著的作用。

另外，利用在实验中得到的词语情感倾向值列表，我们在中文语料上进行了简单的篇章情感倾向判断实验，该实验所使用的情感倾向值列表共包括测试集 1 中的 6445 个词语，倾向值计算方法基于 HowNet 语义相似度方法，使用基准词的数目为 40 对。

在 3.1 节中已经提到，如果我们仅仅利用倾向词表中词的褒贬倾向，而不涉及具体倾向值，就可以用比较简单的方法计算篇章情感倾向。有了文档中每个词

的情感倾向，一篇文档的情感倾向可通过统计文档中所有肯定倾向词的出现频率减去所有表示否定倾向词的出现频率来进行判定。

在这里，我们利用倾向词表中的情感倾向值，通过统计这些词在文章中出现的频率，相应的乘以该词的倾向性值，相加所有词的值最后得到整篇文章的倾向性值。

为了测试该方法的性能，我们从因特网上收集了两类语料：**A 类语料**(这里确切主题的名称略去，下同)，选取其中支持 500 篇，反对的 500 篇；**B 类语料**，选取其中支持 150 篇 反对 150 篇。

下图是采用词的倾向性值进行态度分类的结果：其在这两类语料上的准确率如 0 所示。其中反对和支持两行为两类各自的精确率，全部一行为 **A 类** 和 **B 类** 各自整体的准确率。

	A 类	B 类
反对	0.644	0.777
支持	0.547	0.380
全部	0.595	0.577

表 5. 态度倾向性词表的无监督分类结果

从表 5 可以看出分类的效果不甚理想，这里主要原因在于汉语的表达极其丰富，已有的倾向性词表在待判断的语料中出现频率不高，因此性能的提高需要对篇章进行更深入的分析，仅仅利用词汇的情感倾向信息直接对篇章进行倾向性判断是远远不够的，应考虑结合更多的语义分析方法，共同完成这一任务。使用无监督分类的好处在于应用到不同的语料的时候，不需要人工标注的工作，并且可以对任何主题的文章进行态度分类，但准确率目前还比较低。因此，对无监督态度篇章分类算法性能的改进显得尤为必要。

3.5. 小结

本章所提出的方法是作者在汉语情感倾向判别研究中的一个初步尝试。基于 **HowNet** 的情感倾向判别，只需利用 **HowNet** 的本地资源和少量的基准词，

比较容易实现且不受外界条件（如网络环境）的干扰。从实验结果来看，基准词的增加对判别的准确性提高有明显作用，但即使是极少量基准词，在常用词集中同样可以达到 80% 以上的准确率。

我们所使用的算法较简单，仍有许多值得商榷和改进的余地。例如，在基准词的选择方面，选择一些具有代表性且倾向强烈的基准词，对实验效果的提高将会有很大的帮助，作者的实验所使用的基准词只是简单的根据词频进行挑选，难免有一些偏颇和失当；在使用的算法方面，对于情感倾向值的计算方式比较直观，如果改用其他更加科学的计算方式，相信准确率会有更大的提高。

词语级情感倾向的判断直接影响情感倾向系统的分析结果。虽然可以通过现有的词表解决一部分判别问题，但是在真实语料中，根据上下文的不同，相同词语所表现的倾向性差异很大。因此，如何利用上下文信息，判别动态语境下的词语级情感倾向将是一个极具研究价值的问题。利用本章中所提的方法我们目前所能得到的还只是一个静态的词语情感倾向值列表，如果考虑词语在具体语境下倾向性的动态变化，实际应用价值将会更高。

作为篇章情感倾向判断的关键问题，我们在词语的褒贬倾向判别的方法上进行了一些探讨，期望对相关研究工作有所帮助。

第四章 文本情感倾向分析中的信息抽取

与通用倾向分析技术的研究相比,开发一个实际应用的系统,所要考虑的问题更为具体、更为复杂。在领域无关的情感倾向分析中,我们主要考虑的是情感词即评价词的识别。而在领域相关的系统开发过程中,我们还需要考虑评价对象的识别、领域知识库构建等领域相关的问题,只有将情感倾向分析所需的所有相关信息识别并抽取出来,才能进行有效的情感倾向分析。

本章以一个商用情感倾向分析系统为例,介绍了针对领域相关的情感倾向分析系统中的信息抽取任务所做的一系列工作。

此项目由项目小组多位成员合力完成,作者在其中主要负责评论分析模块的信息抽取以及领域本体的构建扩展等情感倾向分析相关工作。

4.1. 任务定义

由于信息抽取任务是基于领域的,针对不同的应用背景将使用不同的解决方案。因此我们无法撇开具体应用来研究这项技术。本节首先介绍一个基于汽车领域的商用品牌信誉度分析系统。然后对基于该领域的信息抽取做了一个明确的任务定义。

4.1.1. 项目背景

在 WEB 时代之前,用户和用户之间,用户和厂商之间的交流互动,必须通过传统媒体或者口头信息作为媒介。无论是用户获取信息和表达意见,还是厂商发布信息和搜集反馈,都必须通过这些方式。WEB2.0 的出现,颠覆了这一传统。在电子商务或者权威媒体网站,编辑一般会对新产品进行测试,发表评论,用户也可以对特定的产品发表意见。网上还有很多自由论坛,博客,讨论区,可供人们发表关于产品的评论。人工搜集网络信息并进行评论分析,耗时耗力,而找到的信息也未必是全面的。

面对这种情况,如果有一个商业智能系统能够为厂商和用户从 Web 中挖掘产品(包括品牌)评论,提供整体评分和详细摘要,无疑是一件深受欢迎的工作,用户就不用费时费力地寻找信息,厂商也可以容易地进行用户调查,获取产品的反馈意见。

开发品牌信誉度分析系统的目的是，对产品和品牌的评论进行更加细致的挖掘，不仅得到评论的总体态度，也就是我们所说的情感倾向，还试图发现评论所针对的具体对象和属性，并提供可视化的结果。

这里我们选择了汽车领域作为我们的研究对象。

这样的—个系统，主要针对两大类用户。第一类用户是正在关注汽车产品的潜在顾客，第二类是汽车生产厂商。从用户的角度出发，第一类用户最希望通过系统，获取相关的产品信息，特别是网络上其他顾客的对该产品的评价，并以此作为他们进行购买决策前的参考；而第二类用户，非常希望了解他们的产品在顾客心目中的地位、口碑和信誉度，对网络上丰富的评论资源进行分析，能够帮助他们掌握市场动态、及时获得顾客反馈，树立良好的品牌形象。

毫无疑问，对网络评论的情感倾向分析是该系统最为关键的部分。

4.1.2. 系统设计

我们所开发的系统是一个基于 Web 的品牌信誉度分析系统。由用户选择感兴趣的—品牌，系统对关于它的评论进行分析，以可视化的形式给出分析结果。

下图是系统框架的设计图：

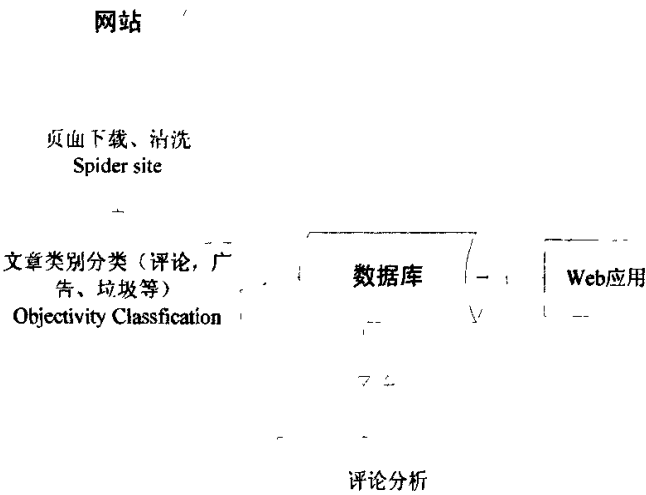


图 7

首先对权威站点、BBS、blog 的网页，调用“页面下载和清洗模块”，进行初步筛选。

对于经过清洗的文章，调用“文章分类”模块，区分评论、广告、垃圾信息，并将评论和广告分别存储到数据库中；

对于数据库中的评论文章，调用“评论分析”模块，对文章进行预处理后，从文章中抽取评论对象，评价词，检索评论对象相关信息，并进行归类整理，分析情感倾向，分析结果记录到数据库中。

最后调用“Web 应用”模块，将经过情感倾向分析的结果数据显式输出。

4.1.3. 系统实现的功能

该系统主要实现了以下功能：

1. 自动下载

根据预定义的评论网站列表，定期自动下载评论并进行过滤，获取汽车相关的评论，导入数据库。

2. 近期热点排名

显示当天（或一段时间）内评论最多的十大汽车品牌或型号，按由高到低顺序排列；罗列出评论最多的汽车被评到的评论对象及相对应的评论词。

3. 具体车型的详细信息

根据不同的汽车品牌、型号进行查询，获得更加详细的评论分析结果。包括对该车型的具体部件或性能，如车身、车尾、油耗、安全性、车内装潢、舒适度等的评论分析数据。针对不同评价对象的正面、负面评价的柱状图；相关评论的评论对象分布情况的统计饼状图等。

4. 不同车型的相关评论比较

用户可选择两个具体车型进行比较。通过具体车型性能、部件等各个方面的相关评论情感倾向分析，并通过图标的形式反映给用户。使得用户能够更直观地了解不同品牌的汽车，甚至是汽车的某项性能指标在大众心目中的口碑或信誉。

下图是品牌信誉度分析系统的比较界面，柱状图中用不同的颜色分别代表被比较的两种汽车品牌型号。柱状条代表的是针对某一评价对象的褒义评价词数占总数的百分比，柱状条越长，褒义评论词的百分比越高。不同柱条后方框内的文

字代表其对应的评价对象，数字代表了对应的评论词的个数。



图 8. 品牌信誉度分析系统的比较界面

4.1.4. 信息抽取任务

系统的主要工作是搜集、整理网络上与汽车相关的评论，并对相关评论文章进行情感倾向分析。系统的核心模块是评论分析模块，它的主要任务就是从评论文章中，分析出这篇评论描述的对象是什么，并且判断出针对该评论对象的情感倾向，通过整理归纳将结果提交给用户。因此，在信息抽取过程中，对于每篇评论，不但要能够得到情感倾向，即抽取评价词，同时要抽取出其相应的评价对象。

对于评价词，我们可以利用第三章所提到的词汇情感倾向值列表，因为列表中的情感词是领域无关的。但是由于该词表规模较小，应该根据实际情况进行扩充，构建适用于该系统的评价词 Ontology；而对于评价对象，由于汽车领

域所涉及的评价对象主要是生产商、品牌、型号以及汽车的部件、性能等，这些名词都是专有名词，无法在通用词库里找到，所以必须通过构建评价对象 **Ontology**，才能顺利的完成信息抽取的工作。

在这个系统中，领域 **Ontology** 的质量直接决定了信息抽取效果的好坏。

4.2. 本体（**Ontology**）的构建

为了准确地抽取评价词及评价对象，获得更可靠的文本情感倾向分析结果，需要建立领域相关 **Ontology**。

针对汽车领域的 **Ontology** 构建，我们首先要考虑的是通过怎样的结构来存储领域相关信息，更有利于系统对评论进行信息抽取。**Ontology** 中必须能够包含两部分内容：一部分是评价对象的信息，即汽车领域的生产商、品牌、型号、部件等专用名词以及评价对象之间存在的从属关系；一部分是评价词信息，包括词语的情感倾向、倾向强度等。

根据情感分析系统的需要，我们设计了三个领域相关 **Ontology**。由于 **XML** 形式的结构化文档既方便存储又易于检索，所以我们使用 **xml** 格式存储 **Ontology** 信息并通过 **xml schema** 定义 **Ontology** 结构。以下将分别对它们进行介绍。

4.2.1. 评价对象关系 **Ontology**

对评论文章进行分析时，不同文章，其评论对象也不尽相同，可能是针对某生产商，也可能是针对某品牌下的一个具体车型。在系统的产品信誉度结果反馈时，必须将关于评论对象的分析结果进行归并，将同品牌/生产商下的各个具体车型的评论综合起来，从而更好地反映某品牌/生产商的信誉度。因此在分析评论对象的同时，我们需要关于该评论对象的相关车型、品牌、生产商信息。

因为生产商、品牌、车型这三者之间具有层次结构，每个生产商旗下包括很多汽车品牌，而每个品牌下又有不同的车型，所以这里我们选择使用树型结构存储信息，

下图为汽车生产商、品牌、型号的 **Ontology** 树型存储结构（以 **XML Schema** 图形化结构表示），在其相对应的 **XML** 文件中，**Manufacturer** 节点代表一个生产商，**MName** 节点存储该生产商名称，**Brand** 节点表示其下属汽车品牌（允许有多个品牌），**BName** 节点存储品牌名称，**Model** 节点存储某品牌

下各车型的信息（允许多个车型）。

所有的信息无法通过自动的方式获取，都是人工搜集并整理的。根据我们收集的资料，目前该 Ontology 总共包含 93 家生产厂商、219 个品牌、504 种型号的信息。

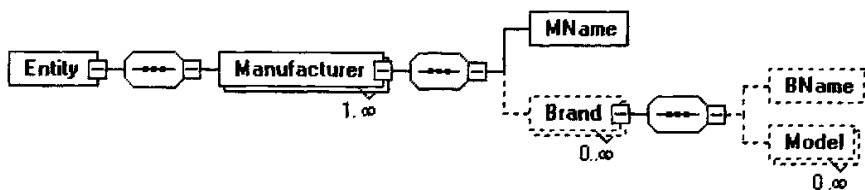


图 9

以下是评价对象关系 Ontology 的一个例子，在这个简单的 XML 文件中，一共存储了 2 家生产商：“BenZ”、“广州本田”的信息。每家生产商下包含若干品牌及型号。这里共包含了 3 个品牌和 5 个汽车型号。

```
<?xml version="1.0" encoding="UTF-8"?>
```

```
<Entity xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance" xsi:
noNamespaceSchemaLocation="C:\Ontology\ Entity_Relation.xsd">
```

```
<Manufactor Name="BenZ">
```

```
<Brand Name="Vito">
```

```
<Model> 奔驰 Vito L （7 座）</Model>
```

```
</Brand>
```

```
<Brand Name="Viano">
```

```
<Model>奔驰 Viano 马可波罗 3.0</Model>
```

```
</Brand>
```

```
</Manufactor>
```

```
<Manufactor Name="广州本田">
```

```

<Brand Name="雅阁">

    <Model> 06 款新雅阁 2.0 手动普通版</Model>

    <Model> 06 款新雅阁 2.4 自动豪华版(NAVI)</Model>

    <Model> 06 款新雅阁 3.0V6 自动豪华版(NAVI)</Model>

</Brand>

</Manufactor>

</Entity>
    
```

4.2.2. 评价对象同义词 Ontology

由于不同评论者对同一评论对象的称呼不尽相同,因此必须通过模糊匹配算法以及同义词 Ontology 的构建对评论对象进行归类整理。

在上一小节的评价对象关系的 Ontology 中,并不包括汽车的部件、属性,原因是这一类评价对象与生产商、品牌、型号之间没有逐级包含的关系,例如“油耗”,人们不会脱离某个具体的品牌或车型来谈论“油耗”,但是“油耗”是各类车型都具有的特征,并不是某一品牌或型号专属的特征。在同义词 Ontology 中,这些评价对象都将被包含进去。这里我们不考虑评价对象间的关系,只关心可能成为评价对象的实体以及它在评论中可能出现的形式。同义词信息的搜集整理过程是一个半自动化的过程,采用了人工标注和自动扩展相结合的方式,我们将在 4.3 节对此做详细介绍。

下图为同义词 Ontology 结构(以 XML Schema 图形结构表示)。图中 Entity 节点代表一组同义词,FullName 为该组同义词的全称,Type 代表该组同义词所表示的对象类别,用来标示这个评论对象是生产商还是品牌、型号或是汽车部件、属性等,Synonym 节点存储对同一对象的多种表达方式。

目前收集并整理了 8755 组共 12712 个同义词的信息。

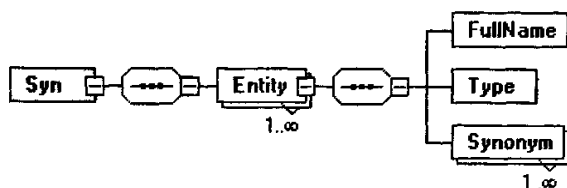


图 10

以下是 Ontology 的 XML 实例：

```
<?xml version="1.0" encoding="UTF-8"?>
```

```
<Syn xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance" xsi:
noNamespaceSchemaLocation="C:\Ontology\Syn.xsd">
```

```
<Entity Name="北京吉普" Type="Manufactor">
```

```
<Synonymy>北京吉普</Synonymy>
```

```
<Synonymy>北吉</Synonymy>
```

```
</Entity >
```

```
<Entity Name="切诺基" Type="Brand">
```

```
<Synonymy>切诺基</Synonymy>
```

```
</Entity >
```

```
<Entity Name="大切诺基 Jeep 4700" Type="Model">
```

```
<Synonymy>大切诺基 Jeep 4700</Synonymy>
```

```
<Synonymy>大切 4700</Synonymy>
```

```
</Entity >
```

```
</Syn>
```

4.2.3. 评价词 Ontology

在第三章中，介绍了词语的情感倾向计算方法，由此我们获得了评价词列表，其中包含了词语及其情感强度的信息。对于一篇评论文章，只要抽取出正确的评价对象及相关评价词，就可以对评论的情感倾向进行分析。在本系统中，我们以原有的评价词列表为基础，对评价词信息进行了扩充，建立了评价词 Ontology，包括了词语、情感倾向、倾向强度、词语的应用模板等信息，这里的倾向强度我们没有像第三章那样用一个 $[-1,1]$ 的实数精确表示，而是使用 1、2、3 简化地表示其强度级别。评价词信息的搜集整理过程是一个半自动化的过程，采用了人工标注和自动扩展相结合的方式，我们将在 4.3 节对此做详细介绍。

下图为评论词 Ontology 结构（以 XML Schema 图形结构表示），每个 OpinionWord 节点表示一个评价词对象，其子节点 Word 存储该评价词本身，Orientation 节点存储其褒贬倾向，Level 对应贬褒强度，Pattern 对应该评价词的词性和使用模板。目前共包含 4460 个评价词信息。

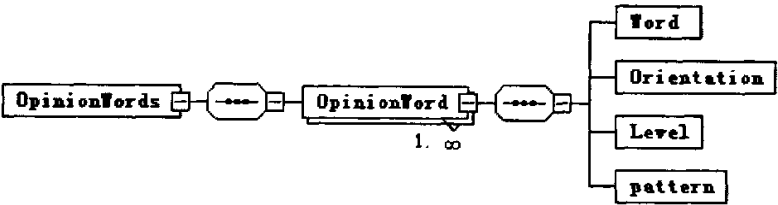


图 11

以下是该领域 Ontology 的 xml 实例，包含 2 个评价词的信息，其中 Pattern 标签中“a”表示词性为形容词，Orientation 标签中的“A”代表 Ambiguous 即中性词；“P”代表 Positive，即褒义词；若值为“N”则表示 Negative 贬义词：

```
<?xml version="1.0" encoding="GB2312"?>

<OpinionWords
xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
xsi:noNamespaceSchemaLocation="C:\Ontology\OpinionWords.xsd">

  <OpinionWord>
```

```
<Pattern>a</Pattern>
```

```
<Level>1</Level>
```

```
<Word>可兼容</Word>
```

```
<Orientation>P</Orientation>
```

```
</OpinionWord>
```

```
<OpinionWord>
```

```
<Pattern>a</Pattern>
```

```
<Level>2</Level>
```

```
<Word>可靠</Word>
```

```
<Orientation>P</Orientation>
```

```
</OpinionWord>
```

```
</OpinionWords>
```

4.2.4. Ontology 编辑工具

人工整理编辑 Ontology 工作量较大,且手工编辑 XML 文件非常麻烦。虽然目前有很多 XML 编辑工具,例如 Xmlspy, Xml Writer 等,但是由于是通用工具,对特定格式的 XML 文件支持不够好,仍然需要人工完成许多重复性的工作,如添加、删除或修改节点等。

为了方便高效地进行 Ontology 的编辑,我们开发了一个编辑工具,该工具能够自动加载 Ontology,通过较简单的点击操作,完成添加,修改,删除节点的功能,减少了大量鼠标操作,提高了工作效率。但由于该工具是针对项目中所需的已知结构的 Ontology 设计的,所以不能作为通用的工具使用。

以下是使用工具编辑同义词 Ontology 的示意图:

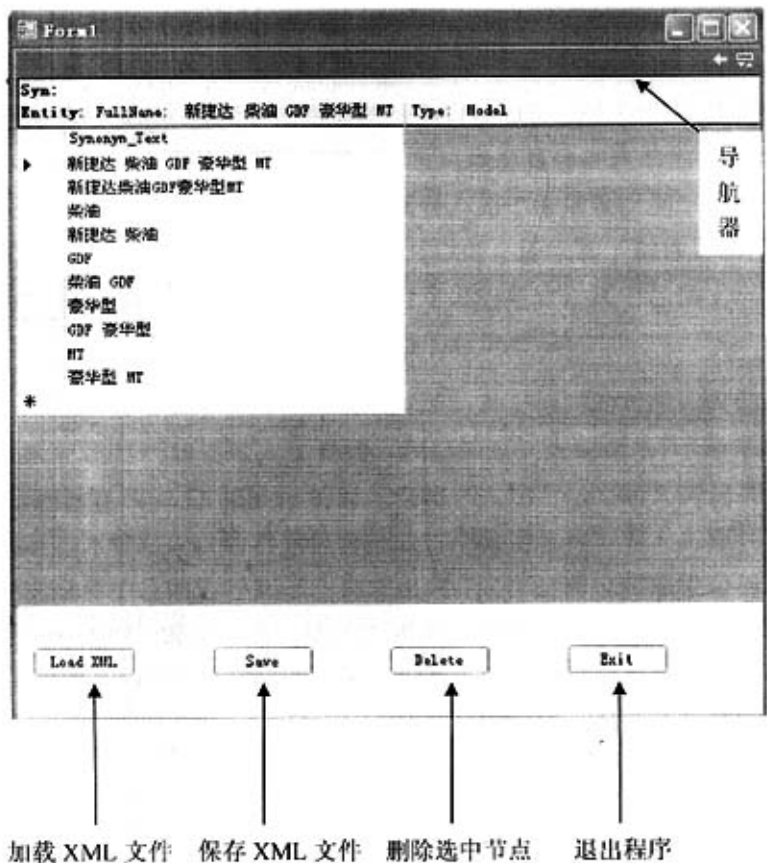


图 12. 使用工具编辑同义词 Ontology 的示意图

4.3. Ontology 扩展

在领域相关的系统中，Ontology 中包含的信息大都需要人工收集整理，但是人力毕竟有限，通过人工的方式获得的信息，相对于网络上的海量文本，仍然是少之又少，其信息涵盖面无法满足实际应用的要求，很多评论中出现的评价对象名称、评价词在人工构建的信息中找不到，也就无法被抽取出来，这对情感倾向分析系统的性能将会造成严重的影响。

因此，我们设计了一些自动或半自动的方法，对 Ontology 进行扩展，期望扩展后的 Ontology 能够匹配更多的评论内容。

4.3.1. 评价对象 Ontology 扩展

同义词 Ontology 包括了所有可能成为评价对象的实体, 以及它们在网络评论中可能出现的不同表达形式。评论对象可能会具体到汽车的某部件, 如“安全气囊”、“导航系统”等, 或者是某些属性, 如“排量”、“刹车性能”等, 因此, 在同义词 Ontology 中, 除了品牌型号以外, 还添加了部件、属性实体名。通过人工收集整理, 我们获得了一个评价对象的集合, 其中包括了汽车的生产商、品牌、型号、零部件、属性五种类型的评价对象。

但是当处理真实文本的时候, 我们发现, 对于相同的对象, 人们会使用不同的方式来表述。例如通过人工查找, 我们知道“大切诺基”, 经常被人称为“大切”, 因此在编辑 Ontology 的时候会将“大切”作为同义词记录下来。但是人工整理的评价对象, 通常只包含一或二个常用名称, 靠人工查找归纳, 很难将网络中形形色色的称呼都一一列举出来。这样的评价对象集合与真实语料中的匹配程度较低, 在评论中, 人们经常使用简称、绰号等各种各样的方式来表示同一个对象, 因此我们需要尝试更加有效的手段对已有的信息进行扩充。

同义词 Ontology 的扩展主要就是希望通过我们已知的评价对象名称, 根据一些合理的规则进行扩展, 获得更多可能的表达方式, 即同义词, 希望在真实语料中识别出更多的评价对象, 以此提高系统的分析性能, 使获得的分析数据更具参考价值。

通过对大量网络评论的分析, 我们发现, 人们在评论中, 特别是汽车相关的评论中, 很少使用汽车的全称。汽车型号的全称一般比较长, 从实际的评论文章中, 我们观察到, 评论者总是使用比较简略的称呼来表示车型, 例如: “威驰 GLX-S 导航版”, 可简略成“GLX-S”或者“GLX-S 导航版”等, 而原始本体中只包含部分厂商或者型号的全称, 因此, 我们使用了 2-gram 的方式进行全称的拆分组合。对车型、品牌以及部分部件进行了同义词扩充。

同义词扩充的过程如下:

第一步, 将全称进行人工切分, 获得一些有意义的片断: 例如: 将“威驰 GLX-S 导航版”切分为“威驰 | GLX-S | 导航版”。

第二步, 对切分好的片断进行组合。使用上一步的切分片断, 我们可以组合成“威驰 GLX-S”, “威驰导航版”等。

第三步, 计算所有组合的 PMI 值[30], 根据选定的阈值删除不常出现的组

合。

PMI 计算公式如下：
$$\frac{\text{hits}(\text{片断1}+\text{片断2})}{\text{hits}(\text{片断1})\text{hits}(\text{片断2})}$$

在第三步之所以使用 PMI 值作为标准对新生成的候选同义词进行筛选，是希望通过搜索引擎进一步核实各个候选同义词在大量文本中共同出现的情况，尽量避免自动扩展的过程中生成过多无意义的短语。

经扩充，同义词总数增加到 3854 组共 7482 个。

除此以外，我们在对网络文本的分析中发现，由于网络评论很大一部分源自论坛、博客、新闻组等处，通常讨论氛围比较轻松，评论的形式也比较随意，作者发言以表达观点为主要目的，文字上的小错误若不影响理解，就不太在意。在这种情况下，很多评论文本中，错别字现象比较常见。

另外，由于许多品牌名称是由英文直接音译成中文，所以人们在进行评论的时候，并不关心中文名称的对错，只要读音准确，则有可能使用不同的汉字来表示。

针对这样的现象，我们在评价对象识别中还使用了汉语拼音模糊匹配的方法。只要评论文章中某个词语的汉语拼音与 Ontology 中已有的实体名相同，则将它作为该实体的同义词抽取出来。利用汉语拼音的模糊匹配方法，进一步扩展了同义词 Ontology 的覆盖面。

4.3.2. 评价词 Ontology 扩展

评价词 Ontology 的扩展主要通过人工标注和机器自动扩展相结合的方式完成。主要工作是对原有的情感倾向词表进行重新标注词语倾向性，并扩充新词。扩展分为两步，第一步是扩展所有的形容词和副词，第二步是根据义原进行扩展。

原始词表包含 1011 个词语，包含词语倾向性及模版，但是不包含强度。

第一步，形容词和副词扩展。我们首先根据 HowNet 增加了所有的形容词和副词，除去重复后共有词语 1409 个。对这些词语，由两个标注人员分别标注词语的倾向性和强度等信息。对词语的倾向标注分褒义、贬义、不确定三种，其中褒贬倾向强度再细分为四个等级。强度等级使用整数进行量化表示。对于褒贬倾向性确定的词，使用 0~3 分级标注强度，数字越大表示情感倾向强度越大。这里需要注意的是，即便是情感倾向为不确定或者中性的词，根据其自

身的特点，它的语义强度也可能不为 0。标注结果如下表：

类 别	个 数	百分比
倾向性和强度都相同	431	30.59%
倾向性不同	339	24.06%
倾向性相同，强度不同	639	45.35%

表 6

标注过程由于是标注人员的主观判断，因此会出现不一致的现象。标注结果中，标注的倾向性相同但是强度不同的词语共有 639 个，其中有 544 个（占总数的 85%）的情感倾向强度差别为 1，只有 95 个词语（占总数的 15%）的强度差别为 2。这些强度差别，主要是由个人主观判断的细微差别所造成的，对于这样的不一致，由标注人员共同协商确定最终的强度值。对于倾向性不同的部分，由第三人根据原始词典的结果和相同词语的标注结果综合考虑给予最终判断。例如：

标注的褒贬一致，强度不同，如“爱戴”，A 标的强度为 3，B 标的强度为 2，则最终值由标注人员 A 和 B 协商确定。

标注的褒贬不一致，如“白衣苍狗”，A 标了 P (Positive)，B 标了 A (Ambiguous)，则由 C 根据原始词典结果最终标注为 A(Ambiguous)。

第二步，根据义原对评价词 Ontology 进行扩展。HowNet 使用 2193 个类似“wise|智”的“义原”对全部 8 万个词语进行意义上的标注，我们根据这个特点，并结合“如果义原有强烈的情感倾向，那么它所描述的词语也应该有情感倾向”的假设对 HowNet 中的词语进行标注。过程如下：

- 1) 对所有义原的情感强度进行标定 (0, 1, 2, 3)
- 2) 抽取出所有包含强情感倾向的词语
- 3) 对抽取出的词语的情感倾向和强度进行标注

使用第二步的义原扩展方法我们总共得到 2943 个词语。

通过以上两个步骤，我们完成了对评价词 Ontology 的扩展。将第一步和第二步的评价词合并，并去掉重复的词语，总共得到 4235 个包含情感倾向的词语。

4.4. Ontology 在系统中的应用

4.2.1 节中，我们介绍了汽车品牌信誉度分析系统的总体设计，我们所构建的三个领域 Ontology 主要被用在系统的评论分析模块中，帮助完成对网络评论的情感倾向分析任务。三个领域 Ontology 中，两个是针对评价对象的 Ontology，一个是针对评价词的 Ontology。

评论分析模块的流程图如下图所示：

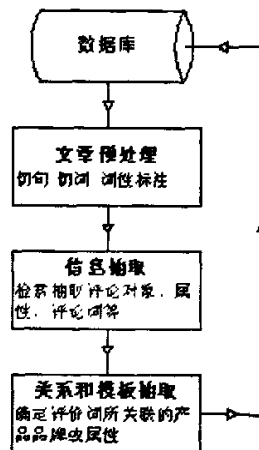


图 13

评论分析模块，分四个步骤对评论进行处理：

1. 文本预处理：对中文的评论进行切句、切词、词性标注。
2. 信息抽取：根据“评价词 Ontology”和“评价对象同义词 Ontology”抽取出评论文本中与评论的情感倾向分析有关的词语和片断。
3. 关系和模板抽取：确定评价词所对应的产品品牌或属性。通过评价词与其评价对象的关联线索分析，确定评价词及其评价对象的关联关系，判断情感倾向。
4. 分析结果汇总：根据“评价对象关系 Ontology”，将情感倾向分析结果根据评价对象进行归类，将分析结果记入数据库。

评论分析模块的信息抽取和结果汇总工作都依靠这三个领域 Ontology 完成。

在信息抽取过程中，对于一篇待处理的评论文本，我们需要借助“评价词

Ontology”，识别出文本中的评价词。由于 Ontology 中包含了评价词、倾向、强度、模板等信息，因此只需通过 Ontology 的检索，系统就可以判断出评论中哪些是评价词，以及评价词的情感倾向。评论中的评价词，只有和它对应的评价对象联系在一起才有意义。因此，系统还需利用“评价对象同义词 Ontology”，通过对该 Ontology 的检索，将评论中的所有可能的评价对象都识别出来。

关系和模板抽取的过程，就是将评价词和评价对象联系在一起的过程。这里，我们运用了最大熵模型通过训练，来预测评价词与评价对象的关系。使用的特征包括评价词与评价对象间的距离、前后位置、词性等。

通过以上步骤，系统已经从网络上下载的评论中抽取了评价对象、评价词，并作出了情感倾向判断。一般每篇评论文章中，系统都能提取出针对不同评价对象的观点若干条，所以对于从网络上下载的数量极其庞大的评论文本，系统获得的大量情感倾向分析结果必须通过归纳整理之后才能提供给用户。也就是说，需要对情感倾向分析结果根据评价对象进行归纳整理。

结果汇总将涉及两个评价对象相关的 Ontology，即“评价对象同义词 Ontology”和“评价对象关系 Ontology”。

对于一条“评价对象—评价词—情感倾向”信息，系统先根据“评价对象同义词 Ontology”获得该评价对象的一系列同义词，这样，很多针对同一对象的情感倾向分析结果将被合并在一起。

另外，从用户角度考虑，在做汽车的品牌信誉度的归类统计时，用户最关心的是网络评论对某个品牌汽车所持的态度。而真实的网络评论中，评论作者一般会直接讨论某个具体的车型而不提及其品牌或生产商，对此，系统可以根据已构建的“评价对象关系 Ontology”，将用户的评价对象的相关信息检索出来，例如，给出一个车型，可以得到相应的品牌—生产商信息。通过评价对象的相关信息检索，系统可将情感倾向结果数据进一步根据生产商、品牌进行分类、归并，按不同生产商，不同品牌进行数据分析，并将分类整理后的结果提供给用户。

4.5. 小结

本章以一个实际的汽车品牌信誉度分析系统为例，介绍了在领域相关的文本情感倾向分析系统中，评价词和评价对象相关的 Ontology 构建和在情感倾向分析中的应用。

领域相关 Ontology 的构建首先要考虑情感倾向分析的要求。Ontology 中

信息的搜集和整理通常都由人来完成,但是完全人工的方式耗时耗力,所以在本章中我们使用了人工搜集和机器自动扩展相结合的方式。对于评价对象,我们使用名称的切分组合以及汉语拼音的模糊匹配完成了评价对象的同义词扩展。对于评价词,我们利用了经情感强度标注的 **HowNet** 义原进行扩展,使得 **Ontology** 与真实语料的匹配程度更高,也使情感倾向分析结果更具参考价值。

我们所使用的扩展算法较简单,仍有许多有待改进的方面。例如,我们的 **Ontology** 只考虑了静态的词汇倾向性,而在许多情况下,情感倾向性是与上下文相关的,因此如果能在 **Ontology** 中引入上下文信息,对于情感倾向判断准确性的提高将会有很大的帮助。另外,领域相关的评价对象主要是一些专用名词,一般字典中无法查到。这对信息抽取的准确性有很大的影响,这里我们使用的方法主要基于人工搜集的评价对象,通过简单的切分组合进行扩展,若能引入实体名识别的技术,相信评价对象的识别效果也将有进一步的提高。

领域 **Ontology** 的构建,对于领域相关的文本情感倾向分析系统而言,是个无法回避的关键问题,**Ontology** 质量的好坏直接影响情感倾向分析结果的好坏。我们在构建的方法和应用上进行了一些探讨,期望对相关研究工作有所帮助。

第五章 总结和展望

本文综合了作者在攻读硕士学位期间所参与的研究工作和系统开发项目,对文本情感倾向分析中的若干问题,进行了深入的探讨。

文本的情感倾向研究是近年来自然语言处理领域的研究热点。国外已经有了一系列的研究成果。但是在中文领域,这方面的研究才刚刚起步。本文所涉及的所有研究工作都是在中文领域进行,希望我们的工作能够为推动该项研究课题的进一步发展起到抛砖引玉的作用。

文章首先从文本情感倾向研究的通用技术入手,对其中的关键技术—词汇的情感倾向计算方法进行了一系列的探讨。对词的倾向性分析和态度分类的研究,到目前为止仍然是个崭新的课题,本文提出了利用中文 **HowNet** 汉语知识库的语义相似度和语义相关场计算功能,对词汇进行情感倾向性判别的方法,并通过多次实验,对该方法的有效性进行验证。从实验结果来看,基准词的增加对判别的准确性提高有明显作用,但即使是极少量基准词,在常用词集上仍然获得了令人满意的效果,达到了 80% 以上的准确率。同时,该算法只利用了 **HowNet** 的本地知识库资源,无需考虑网络等外界条件的影响,算法容易实现且运行效率高。

但是由于作者在自然语言领域的研究经验尚浅,所提出的方法比较简单、直观,不可避免的有一些有失偏颇或考虑不当之处。比如在基准词的选择方面,只是单纯地使用词频作为选择依据,而没有考虑词汇的情感强度等语义信息,若能选择一些具有代表性且倾向强烈的基准词,对实验效果的提高可能会有帮助;在使用的算法方面,对于情感倾向值的计算方式比较直观,如果改用其他更加科学的计算方式,相信准确率会有更大的提高;另外,在词汇的倾向性判断中,若能考虑语境对倾向性的影响,其计算结果在真实语料中的实际应用价值也将大大增加。

本文在第四章中进一步对领域相关的文本情感倾向分析相关技术进行了探讨。基于对该项课题更深入细致的研究,我们开发了一个面向汽车领域企业和顾客的网络评论情感倾向性分析系统—汽车品牌信誉度分析系统。在进行系统开发的过程中,我们发现,情感倾向关键信息的抽取是系统不可或缺的重要组成部分。换句话说,如果无法识别出文本中大部分的评价词和评价对象,那么系统的最终分析结果就失去了意义。因此,我们在利用第三章词汇情感倾向计算结果的基础上,进一步对评价词进行了扩充,建立评价词 **Ontology**。同时,对于领域相关

的评价对象,使用人工搜集整理加上机器自动扩充的方式,构建了评价对象相关的两个 **Ontology**。在实际系统中,这些 **Ontology** 被运用在评论分析核心模块,对于整个系统的性能好坏起到了至关重要的作用。但是,在构建 **Ontology** 的过程中,虽然使用了部分自动扩展的方法,可手工标注的工作量仍然很大,如果简化人的工作,在不降低准确度的情况下增加 **Ontology** 构建的自动化程度,是我们下一步需要考虑的问题。

文本情感倾向分析是一项涉及面较广的领域,还有很多问题需要深化研究。希望本文的工作能为中文的词汇、文本的倾向性分析、倾向分析系统开发等方面的研究发展贡献一点微薄的力量。

最后,对信息内容进行处理,已是时代大势所趋,也面临着诸多挑战。特别是由于自然语言的复杂性,使得自然语言处理的一些领域困难重重,很难在短时间内有大的突破,因此需要长期不懈的努力。

总之,在这个研究领域中,我们任重而道远。

参考文献

- [1]. Vasileios Hatzivassiloglou, Kathleen R. McKeown. Predicting the semantic orientation of adjectives. Proceedings of the 35th Annual Meeting of the Association for Computational Linguistics and the 8th Conference of the European Chapter of the ACL, 1997:174–181.
- [2]. Turney, Peter, Littman Michael. Measuring praise and criticism: Inference of semantic orientation from association. ACM Transactions on Information Systems, 2003, 21(4): 315-346
- [3]. Hiroya Takamura, Takashi Inui and Manabu Okumura. Extracting Semantic Orientation of Words using Spin Model. ACL 2005, pages: 133 – 140
- [4]. Pero Subasic, Alison Huettner. Affect Analysis of Text Using Fuzzy Semantic Typing. IEEE-FS, 2001, Volume 9: 483-496
- [5]. Boucouvalas, A.C., and Zhe X. 2002. Text-to-emotion engine for real time internet communication. In Proceedings of the International Symposium on CSNDSP 2002. 164-168
- [6]. Boucouvalas, A.C. 2002. Real time text-to-emotion engine for expressive Internet communications. In Emerging Communication: Studies on New Technologies and Practices in Communication. G. Riva et al. eds. IOS Press. 305-318
- [7]. Tao, J. and Tan, T. 2004. Emotional Chinese talking head system. In Proceedings of the 6th International Conference on Multimodal Interface. (Oct. 13-15). 273-280
- [8]. Litman, D. and Forbes, K. 2003. Recognizing emotions from student speech in tutoring dialogues. In Proceedings of the IEEE Automatic Speech Recognition and Understanding Workshop (Virgin Islands, Dec.). 39-52.
- [9]. Chung-Hsien Wu, Ze-Jing Chuang and Yu-Chung Lin. Emotion recognition from text using semantic labels and separable mixture models. ACM Transactions on Asian Language Information Processing (TALIP) archive, Volume 5, Issue 2 (June 2006): 165 - 183
- [10]. Turney Peter. Thumbs Up or Thumbs Down? Semantic Orientation Applied to Unsupervised Classification of Reviews, Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics, 2002: 417-424

- [22]. Bing Liu, Minqing Hu, Junsheng Cheng. Opinion observer: analyzing and comparing opinions on the Web, WWW2005, pages 324-351
- [23]. I-Heng Mei, Hongcheng Mi and Julius Quiaot. Sentiment Mining and Indexing in Opinmind. ICWSM2007 Boulder, Colorado, USA
- [24]. 王治敏、朱学锋、俞士汶. 基于现代汉语语法信息词典的词语情感评价研究[J]. 中文计算语言学期刊, 2005(4)581-591.
- [25]. Tsou Benjamin KY, Tsoi WF, Lai TomBY, Kwong OY and Wong WL. Polarity Classification of Celebrity Coverage in the Chinese Press. International Conference on Interlligence Analysis, 2005
- [26]. Yuen Raymond WM, Chan Terence YW, Lai Tom BY, Kwong OY, Tsou Benjamin KY. Morpheme-based derivation of bipolar semantic orientation of Chinese words. COLING-2004
- [27]. HowNet's Home Page. <http://www.keenage.com>
- [28]. Zhendong Dong and Qiang Dong (2006). HowNet and the computation of meaning. Chinese Academy of Sciences, China. ISBN 978-981-256-491-7
- [29]. 刘群, 李素建. 基于《知网》的词汇语义相似度的计算, 第三届汉语词汇语义学研讨会, 台北, 2002
- [30]. Church, K.W. and Hanks, P (1989). Word association norms, mutual information and lexicography. In Proceedings of the 27th Annual Conference of the Association of Computational Linguistics. Association for Computational Linguistics, New Brunswick, N.J., 76-83.
- [31]. Minqing Hu and Bing Liu. Mining and summarizing customer reviews. Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery & Data Mining (KDD-2004, full paper), Seattle, Washington, USA, Aug 22-25, 2004
- [32]. Minqing Hu and Bing Liu. Mining Opinion Features in Customer Reviews. Proceedings of Nineteenth National Conference on Artificial Intellgence (AAAI-2004), San Jose, USA, July 2004.
- [33]. Riloff, E., Wiebe, J., and Phillips, W. (2005). Exploiting subjectivity classification to improve information extraction. Proc. 20th National Conference on Artificial Intelligence (AAAI-2005)
- [34]. J Wiebe, E Riloff, Creating Subjective and Objective Sentence Classifiers from Unannotated Texts, Proceedings of CICLing, 2005
- [35]. Hatzivassiloglou V and Wiebe JM. Effects of Adjective Orientation and

- [11]. Bo Pang, Lillian Lee, Shivakumar Vaithyanathan. Thumbs up? Sentiment classification using machine learning techniques. In Proceedings of the 2002 Conference on Empirical Methods in Natural Language Processing. 2002:79-86.
- [12]. Bo Pang, Lillian Lee. Seeing Stars: Exploiting Class Relationships for Sentiment Categorization with respect to Rating Scales, ACL2005, pages 115-124
- [13]. Janyce Wiebe. 1990. Recognizing Subjective Sentences: A Computational Investigation of Narrative Text. Ph.D. thesis, State University of New York at Buffalo.
- [14]. Janyce Wiebe. 2000. Learning subjective adjectives from corpora. In Proceedings of the Seventeenth National Conference on Artificial Intelligence (AAAI-2000), pages 735-740
- [15]. Ellen Riloff, Janyce Wiebe. 2003. Learning extraction patterns for subjective expressions. In Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP-2003), pages 105-112
- [16]. Ellen Riloff, Janyce Wiebe and Theresa Wilson. 2003. Learning subjective nouns using extraction pattern bootstrapping. In Proceedings of the 7th Conference on Natural Language Learning (CoNLL-2003), pages 25-32
- [17]. Theresa Wilson and Janyce Wiebe. 2003. Annotating opinions in the world press. In Proceedings of the 4th ACL SIGdial Workshop on Discourse and Dialogue (SIGdial-03), pages 13-22
- [18]. Turney, Peter D. (2006) Similarity of Semantic Relations. Computational Linguistics 32(3): 379-416.
- [19]. Devillers L., Vasilescu I., and Laniel L. 2002. Annotation and detection of emotion in a task-oriented human-human dialog corpus. In Proceedings of the ISLE Workshop on Dialogue Tagging for Multi-Modal Human-Compute Interaction (Dec. 15-17)
- [20]. Devillers L., Laniel L., and Vasilescu I. 2003. Emotion detection in task-oriented spoken dialogues. In Proceedings of the International Conference on Multimedia and Expo (Baltimore, MD, July 6-9). 549-552.
- [21]. K Dave, S Lawrence, DM Pennock. Mining the peanut gallery: opinion extraction and semantic classification of product reviews, WWW2003, pages 519-528

- Grad-ability on Sentence Subjectivity. COLING 2000
- [36]. Peter D. Turney. Thumbs Up or Thumbs Down? Semantic Orientation Applied to Unsupervised Classification of Reviews. Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics, 2002.
- [37]. Wilson Theresa & Wiebe Janyce & Hwa Rebecca. Just how mad are you? Finding strong and weak opinion clauses. AAAI-2004
- [38]. P.D.Turney, M.L.Littman. Unsupervised Learning of Semantic Orientation from a Hundred-Billion-Word Corpus. NRC/ERB-1094.May 15,2002.9 pages. NRC44929
- [39]. H. Yu and V. Hatzivassiloglou. Towards answering opinion questions: Separating facts from opinions and identifying the polarity of opinion sentences. In Proceedings of EMNLP-03
- [40]. Bo Pang, Lillian Lee. A Sentimental Education: Sentiment Analysis Using Subjectivity Summarization Based on Minimum Cuts. ACL2004
- [41]. Ana-Maria Popescu and Oren Etzioni. Extracting Product Features and Opinions from Reviews. EMNLP2005

论文发表情况

发表论文：

朱嫣岚，闵锦，周雅倩，黄萱菁，吴立德，“基于 HowNet 的词汇语义倾向计算”，中文信息学报，2006 年第 20 期，第 1 卷。

参加项目：

上海市科学技术委员会重点项目“基于语义的特定领域过滤工具的研究与实现”（项目编号035115028）。

富士通公司合作项目“基于Web的汽车品牌信誉度分析”

致 谢

值此论文即将完成之际，回首走过的道路，一路结伴的可敬的老师、同窗们，一时感慨万千。

能够顺利的完成论文，首先要感谢吴老师和黄老师的悉心指导。三年前加入实验室至今，两位老师对我悉心的教导和无微不至的关怀令我受益匪浅。吴老师严肃的科学态度，严谨的治学精神和渊博的学识令我钦佩之余，深深地感染和激励着我。他经常提出一些高屋建瓴的想法，给予我无尽的启迪；而黄老师则不仅在学业上给我以精心指导，同时还在思想、生活上给我以无微不至的关怀。两位老师引导我进入了自然语言这个研究领域，不仅仅教授给我丰富的专业知识和经验，更身体力行地教给我做研究的正确方法和态度，帮助我打下了坚实的基础，并在我的研究成长过程中不断纠正我的错误，细心的看护着我走上正确的道路。感谢两位老师在繁忙的工作之中对我始终如一的指导，没有他们两位良师的无私帮助，就不会有我现在取得的这些成果。

实验室是个大家庭，在此，我还要感谢在一起愉快的度过研究生生活的实验室的老师和同学们。周雅倩老师对我的研究工作提出许多宝贵的意见，并且不厌其烦地帮助我修改完善实验细节。在初次接触情感倾向研究时，闵锦、武晓春同学和我一起完成过滤项目，在过程中热心解答我的问题，并推荐了许多参考资料，在项目中帮助我解决技术难题，对我的帮助非常大。张奇同学在情感倾向分析项目中，花费大量时间和我讨论解决方案，并对我的研究工作提出许多有价值的建议，帮助我一步步完成自己的工作。章建锋、王秉清、孙磊等同学分别负责情感倾向分析系统的其余各模块，我们配合默契，是大家共同的努力，才能顺利完成项目，我的研究工作才得以顺利进行。同时还要感谢钱线、吴中勤、刘菲、计锋、袁晓丰、邱锡鹏等同学，和大家一起上课，一起参加讨论班，一起做项目，一同出游，在一起的日子虽然短暂却充满了欢乐。感谢大家对我的支持和帮助，和大家在一起的日子着实令人难忘。身处这个积极向上的集体，我无时无刻不被大家的热情所感染，以最佳的状态投入研究工作中。

最后，还要感谢始终关心照顾着我的家人，无论我身处顺境还是逆境，你们都一直在我身边默默支持我，你们是我无尽的动力来源，在这里祝福你们永远健康快乐。

文本情感倾向分析若干问题研究

作者：[朱嫣岚](#)
学位授予单位：[复旦大学](#)

本文链接：http://d.g.wanfangdata.com.cn/Thesis_Y1168307.aspx

授权使用：武汉大学(whdx)，授权号：523e61ea-9f15-43c3-8b38-9e0200dbe4a8

下载时间：2010年10月1日