山 西 大 学
2008 届硕士学位论文

基于汽车评论的文本情感分类特征选择方法研究

作者姓名 魏英杰
指导教师 王素格 副教授
学科专业 模式识别与智能系统
研究方向 文本挖掘与信息检索
培养单位 山西大学数学科学学院
学习年限 2005 年 9 月－2008 年 7 月

二〇〇八 年 六 月

Shan xi University
2008th Dissertation of Master Degree

Research on Feature Selection Method for Text Sentiment Classification Based on Automobile Reviews

Name	Wei Yingjie
Supervisor	Prof. Wang Suge
Major	Pattern Recognition and Intelligence System
Field of Research	Text Mining and Information Retrieval
Department	School of Mathematics Science
Research Duration	Sep.2005——Jun.2008

June 2008

目 录

CONTENTS

中 文 摘 要

近年来随着信息技术的迅猛发展，互联网迎来前所未有的新局面。以网络为传播媒介的文本评论信息越来越受到企事业单位和个人的关注。传统的主题分类已经不能满足人们的需求，用户希望得到更多的主观性信息，如：公共事件的社会反映、焦点新闻的追踪报道、产品的用户反馈及民意调查信息等。然而，网上每天都有大量的新评论出现，对于这些评论，仅靠人工进行跟踪和分析显然是行不通的，人们开始关注并研究评论文本的主观性情感倾向分析。

本文针对文本情感分类中的特征选择问题进行了研究，主要内容如下：

（1）建立了以汽车产品评论为主的中文文本情感分类语料库，并在此基础上建立了汽车产品知识库。

（2）研究了停用词对文本情感倾向性分类的影响。选用信息增益、互信息和χ^2统计三种特征选择方法，布尔权重和频率权重两种权重计算方法，并选用支持向量机作为分类器进行了实验研究。实验结果表明，当选用不同的停用词表时，它们对文本情感分类的影响不尽相同，停用词表 5（不去掉停用词）、停用词表 1（仅选用形容词、动词和副词）对情感分类作用较大，整体性能效果较好。

（3）提出了基于类别区分能力的混合特征选择方法，并测试了其对文本情感分类的作用。该方法是基于词汇的类别区分能力与信息增益相结合的特征方法，讨论了在不同的特征选择方法和不同维数特征空间下对文本情感分类结果的影响。实验结果表明使用混合的特征选择方法要优于使用单一的信息增益方法。

（4）从特征选择和维数压缩的角度，提出了基于粗糙集理论的特征选择方法。通过对情感分类问题的分析并结合粗糙集理论，将属性离散化方法用于文本情感分类中的特征选择、维数压缩，利用支持向量机作为分类器进行分类实验。实验结果表明，该方法具有良好的特征可解释性和较好的特征维数压缩效果。

关键词：文本情感分类；特征选择；停用词；支持向量机；粗糙集

Abstract

With the rapid development of information technology, the Internet will usher in an unprecedented new era. The traditional subject classification has been far and away meeting the people's needs. Now, more and more users would like to get subjective information, such as the reflection of public events, the focus news follow-up report, products users' feedback, poll information. Therefore, texts information from web media are concerned and required by more and more enterprises and individual. There are too many comments on the event appearing on the web everyday. Obviously, it's impossible to track and analyze these reverent comments only depending on manual work. So, more and more people begin to study the subjective sentiment orientation of views text.

In this thesis, the feature selection of the text sentiment orientation classification has been researched as follows.

(1) Construction of the corpus for the text sentiment classification on the basis of the automobile products' comments, and founding automobile products' knowledge base.

(2) Research on the influences of stop words for text sentiment classification. By choosing the three kinds feature selection methods (Information Gain, Mutual Information and Chi-square Statistic), two kinds weighing assignment methods (Boolean Weighting and Frequency Weighting), and Support Vector Machine on text sentiment classification. The experiment results indicate that using different stop lists, their influence is different for text sentiment classification. The stop list 5 (none stop words), stop list 1 (only excluded adjective, verb, adverb) have greater impact on text sentiment classification.

(3) Research on the hybrid feature selection method based on category distinguishing ability and the effect for text sentiment classification. The feature selection method based on the category distinguishing ability of words and information gain is proposed. We discuss the impact of varying the

feature dimension and different methods for text sentiment classification. The experiment results indicate that the hybrid method is superior to that directly using information gain.

(4) From the views of feature selection and dimension reduction, a method for feature selection was presented by using rough set theory. The method of attribute discretization is applied to the feature selection and dimension compression. By using of the support vector machine, the experiment results indicate that the method of rough set has well solubility of feature and better effect of feature dimension reduction.

Key words: Text sentiment orientation classification; Feature selection; Stop words; Support vector machine; Rough set

第一章 绪论

1.1 文本情感分类的意义

近年来，随着信息技术的迅猛发展，互联网迎来了前所未有的新局面。至2007年，我国网民已经达到1.37亿，这为互联网在我国的持续发展奠定了基础。搜索引擎的发展改写了传统门户确立的商业规则，智能化搜索、社会化搜索、垂直搜索等成为新的亮点；博客、群组、社区等新兴的Web2.0服务领域，使用户全面体验虚拟信息生活、社会生活和物质生活，上述这些都共同影响着我们的生活方式。互联网越来越成为人们获取信息的重要来源，以往人们最为常见的应用是针对网络信息客观内容的获取，例如将网络信息按照“教育、经济、政治、环境、计算机、体育、汽车”等按主题进行分类[1-4]，用户得到的是属于客观的事实，显然这些已经远远不能满足人们的需求，用户希望得到更多的主观性的信息，如：公共事件的社会反映[5]、焦点新闻的追踪报道[6]、产品的跟踪和质量评价[7-9]、影视评论[10]、博客评论[11]、信息过滤[12]等。正是在这样的情况下以网络为传播媒介的文本信息越来越受到人们的关注，同时人们对网络信息处理提出了更高的要求。通过网络，用户不仅希望了解事件的本身，更加希望可以了解人们对事件的立场、观点和看法等主观信息。

文本内容的语言表述体现了作者对描述对象的情感（sentiment）。它涉及人们的观点、看法和评价，包括人类行为相对于社会标准的评价，产品相对于国家和行业的强制标准、审美观的评价等。大量的文本不仅包含了主题信息，同时还包含了立场、观点、看法、情绪、好恶等主观信息，这些隐含于文本中带有主观色彩的信息通常称为情感信息。文本的情感倾向是指文本所反映的正面的（positive）或者反面的（negative）倾向性以及情感倾向的强度。

对文本情感问题的研究，有的也称其为情感分类（sentiment classification）[13-17]或情感分析（sentiment analysis）[18-20]，包含分析挖掘文本中对于某一事件大量的立场、观点、看法、情绪、好恶等隐含的主观信息。文本情感分类可以广泛地应用于信息检索、文本过滤、在线讨论跟踪、产品质量评价、民情民意调查分析以及聊天系统等方面。由于网上文本的形式、内容的任意性和开放性，致使文本的情感倾向性研究特别困难，通常需要文本挖掘、信息检索、模式识别、机器学习、统计学、自然语言处理、语言学、本体学和认知科学等多学科手段的交叉。

1.2 国内外研究现状

近年来，国外对于英文文本的情感倾向性研究比较多。在人工智能、信息检索、数据挖掘、自然语言处理以及Web应用等领域的会议和期刊中都收录了文本情感倾向分析的相关论文。B. Pang等[14]使用特征词袋框架技术对文档进行表示，以特征在文档中出现的频数（或频率）作为权重，分别使用Naive Bayes网络、最大熵模型和SVM三种分类方法对评论文本的情感分类进行了研究。Jin Cheon Na等[21]采用支持向量机作为分类器，选用动词、形容词、副词作为特征，使得分类正确率可以达到76%,如果加入否定短语可以使分类正确率达到79.33%。P. Turney [22]利用目标短语与“excellent”和“poor”的互信息差度量该短语的语义倾向，评论中短语的平均语义倾向作为该评论的语义倾向。Yi等[23]使用语法分析器对句子进行语法分析，然后参照情感词汇表（the sentiment lexicon）和情感模式库（the sentiment pattern database）对句子进行情感分类，并将其运用到文本的情感分类。B. Pang等在文献[24]中研究的情感分类的类别采用了多种等级类别。该文献首先采用了人工的方式对其进行分类，然后应用meta-algorithm，通过n-gram分类器给出输出。该方法在使用新的相似性度量后极大改进了多类和回归的SVM的方法。Aidan Finn等[25]使用了词、词性和统计数据三种特征利用C4.5构建分类器，对文档的客观性和主观性分类进行了研究，发现对于单一主题的流派分类，在三种方法中词作为特征得到的分类效果比较好，但将三种特征结合起来可以取得比任何一种特征更好的分类效果。Janyce M.Wiebe从大规模的语料库中学习形容词的主观性[16]，利用分布相似性对词语进行聚类。Dave等人[18]以特征选择从语料中选择出常见情感词，然后通过Bayes网络来发现各词汇与文档的情感分类之间的关系，并据此计算各情感词的得分，并作为该词汇的语义方向。Turney提出了基于PMI-IR的无监督情感分类方法[26]。利用POS标注器识别并抽取包含形容词、副词或名词的短语，再运用PMI-IR方法计算短语的情感倾向（SO），然后,把每篇文本中的所有具有情感倾向的短语的情感倾向值相加，根据得到的值来判断文本的情感倾向。Turney和Littman将单对基准词扩展到多对基准词[17]，使用PMI-IR和LSA算法度量了给定词汇与正面基准词和反面基准词的关联程度，并对两个数值进行比较，最终确定词汇的语义倾向。

目前国内关于中文文本情倾向性研究刚刚起步。香港城市大学Tsou等[27] 采用三个衡量指标，即极性元素分布、极性元素密度和极性元素语义强度，对出现在中国四地（北京、香港、上海、台北）报刊上的四位政治人物（克里、布什、小泉纯一郎、陈水扁）的褒贬性新闻报道进行了分类研究。刘建等[15]提出了文本情感分类的

Super parsing方法，并将其应用于涉及多个对象的文本情感分类与意见抽取问题。Songbo Tan[28]等通过使用四种特征选择方法（MI、IG、CHI和DF）和五种分类方法（centroid classifier、KNN、 winnow classifier、 Naive Bayes 和 SVM）对中文文本倾向性分类问题进行了研究，实验表明IG在特征选择中表现较好，SVM分类性能较好。Qiang Ye[10]通过增加分词技术将文献[17]提出的电影评论情感分类方法应用于中文电影评论，但并未对特征选取等影响情感分类结果的重要因素进行深入的研究。我们也提出了一种混合的特征选取方法[29]，并利用支持向量机作为分类器对汽车产品评论做了情感倾向分类研究。朱嫣岚等[30]基于HowNet,提出了两种词汇语义倾向性计算的方法，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，该方法在汉语常用词中的判别精确率较高。李荣陆[1]使用具有情感色彩的形容词和少量的名词作为文本的特征，并采用判别形容词的语义倾向、利用互信息与潜在语义分析的方法计算词汇权值、人工确定文本特征和词汇权重三种方法，另外还采用了语义模式的方法。柴玉梅等对Web文本的褒贬倾向性进行了研究[31]。该文介绍了Web文本褒贬倾向性分类的原理和实现方法，特征选用已有的特征选择方法与褒贬特征提取技术进行选择，使用几种分类算法实现了名人网页的褒贬倾向性分类。本文将对中文文本情感分类问题中特征选择进行相关的研究。

1.3 课题的研究难度

文本的情感倾向表述与语言学角度的语义褒贬现象有关，同时涉及到现代语言学领域的许多问题。目前，关于中文情感分类的研究思路和技术尚处于起步阶段，其难度是显而易见的，具体来说有以下几个方面:

（1）领域主题相关

某些词在特定领域会呈现出不同的情感倾向。如:“简洁圆滑的新奥迪 A6L 与刚毅锋利的凯迪拉克 CTS 比美，两种截然不同的风格激烈碰撞，让人目不暇接，实在各有千秋。”“圆滑”这个词语根据《现代汉语词典》中的解释，表达的是反面的倾向，但是在汽车评论中形容汽车外形时，它表达的却是正面的情感倾向。

（2）情感倾向相关

某些词语虽然能够呈现出不同的情感倾向，但对于情感倾向的确定却是根据其不同的上下文语言环境来判定。如:“大”这个词语在下面的两句话中就表达出不同的情感倾向。

①“令人惊喜的是 QQ 的内部空间比我想象的要大。”

②“发动机的噪声出奇的大。”

在第①这个句子中“大”表达的是正面的情感倾向，而第②这个句子里表达的却是反面的情感倾向。

（3）内容相关

有些评论语料中会对多主体内容进行对比评价，这就加大了我们研究问题的难度，如：“帕萨特和雅阁，安全设备比较……帕萨特的防盗系统是密匙式的，夸张地说，没有原配钥匙，只有动用拖车才能够盗走；雅阁的防盗，几乎是不设防。尽管如此，我还是喜欢我的帕萨特！”，从作者来看是对帕萨特的评论，但是通篇都是对比，这给特征选择和情感分类带来了许多难度。

基于以上的难点分析，本论文采用网上的真实汽车评论语料，从特征选择的角度对文本情感分类问题进行了深入研究。

1.4 本文的研究工作

（1）为了开展情感倾向性分类问题的研究，建立了以汽车产品评论为主的中文文本情感语料库，并在深入分析评论语料的基础上，建立了汽车产品知识库。

（2）研究了停用词对文本情感分类的影响。停用词是指在文本中出现频率很高，但实际意义又不大的词，主要指副词、虚词、语气词等。如“是”、“的”等。本文通过选用信息增益（IG）、互信息（MI）和χ^2统计三种特征选择方法，布尔权重和频率权重两种权重计算方法，并选用了支持向量机（SVM）作为分类器，考察了停用词对文本情感分类结果的影响。

（3）提出了基于类别区分能力的混合特征选择方法。由于使用单一的特征选择方法不能很好的适应文本情感分类，本文采用了混合特征选择方法，该方法是基于词汇类别区分能力并结合信息增益（IG）方法，讨论了不同特征选择方法和不同维数特征空间对分类结果的影响，并将这些特征选择方法用于网上汽车产品评论文本的情感类别的判断。

（4）统计机器学习的方法在解决情感分类问题中突出的一个问题是，具有情感倾向性的词汇往往出现的次数信息不足，因而情感词汇在特征选择中很容易被漏掉，导致其他不具有情感词汇被选为特征，这也是造成分类可解释性差的原因之一，也就是说统计机器学习方法解决由少量情感词汇决定文章情感类别问题时非常困难。为此，本文提出了基于粗糙集理论的特征选择方法，通过将情感分类问题与粗糙集理论有机结合，研究了基于粗糙集理论的方法在文本情感分类中特征选择、维数压

缩等方面性能。

1.5 论文的组织结构

第一章，介绍本文的研究意义以及国内外研究现状。

第二章，介绍了面向中文文本情感分类数据资源建设。

第三章，利用中文文本主题分类的相关技术（互信息、信息增益方、χ^2统计方法、权重计算方法和支持向量机分类器），研究了五种停用词表对中文文本情感分类的影响。

第四章，基于混合特征选择方法的文本情感分类的研究，采用多种特征选择方法，研究对于汽车产品评论文本的正反两种倾向类别的判断。

第五章，选用粗糙集理论方法，利用属性离散化方法，实现了基于粗糙集理论的特征选择方法。对该特征选择方法在中文文本情感分类中的作用进行了深入研究。

第六章，全文结论与下一步的工作。

第二章 面向中文文本情感分类的数据资源

论文的研究工作是在中文文本情感语料库的基础上展开的。语料内容选择是否恰当，不仅对文本分类器的性能有较大的影响，而且直接影响着对问题的研究。语料应该能够代表研究问题中所包含的特点和突出现象，应该是公认的语料。由于目前中文文本的情感倾向分类刚刚起步，还没有公开的语料等资源可利用，我们自己构建了两个关于汽车产品评论的语料库。在深入分析语料的基础上，认为汽车产品评论能真实反映出情感倾向性分类问题中特有现象和突出问题，并在此基础上针对性的建立了汽车产品知识库和汽车产品评论情感词汇库。

2.1 语料收集

语料库一（corpus 1）：语料全部来自汽车点评网[32]，评论发表时间集中于2006年6月至8月，精选并收集了国内外11种品牌的轿车，总计400篇约41万字。为了反映网站评论的真实情况，其中正面、反面的语料约占网站评论总量的10%，这样导致正面、反面语料数量不一致，比例约为5:3。其中，每篇文本除了标题、正文外没有任何其它附加信息。

语料库二（corpus 2）:在语料库一的基础上进一步扩大规模，增加了 2006 年 9 月至 2007 年 2 月的相关评论，总计 1006 篇约 100 万字。为了研究的需要，这里我们将正面、反面的语料比例进行了平衡，比例约为 1:1，并且新增的语料未进行精选，突出评论文本的一般性。与语料库一相同，每篇文本除了标题、正文外没有任何其它附加信息。

2.2 语料分析

由于本文研究的问题与传统中文文本主题分类问题有很大的不同，这就导致了语料与传统主题分类的语料，如：环境、计算机、体育、经济等有所不同。对于汽车评论，判断其情感倾向的一些常见要素有：价格、售后服务、汽车燃油经济性、采用技术先进性、汽车行驶可靠性、安全性、汽车折旧残值等。对于购买小型车的用户而言，在价格、售后服务、汽车燃油经济性、采用技术先进性方面，各种竞争车型的差别不大，但行驶安全性、可靠性和残值却尤为重要。安全性是购买小型车消费者需要考虑的首要因素。可靠性是汽车使用过程中的一项重要因素。小型车更新的周期较短，残值高的车在转手时能卖个好价钱。另外，市场占有率、品牌认知度等因素都应该是评论发表者表达情感倾向的出发点和基本依据。

然而，真实的语料并非全部如此，这些评论大致分为两类：一是显性的，即无论是从题目、内容用语，还是结论都明确的表明作者的情感倾向和态度；相反另一类是隐性的，作者的情感倾向和态度不直接，这对问题的研究有一定的影响。从网上发表评论的作者来看，基本可以分为两大类：一是对汽车比较了解的专业人士或媒体网站，他们的评论大多从专业或技术的角度，这类评论一般用词恰当、准确，处处都突出鲜明的观点，因此基本符合上述要素，这类评论一般比较容易识别情感倾向；二是汽车爱好者，人群主要集中于车主和即将购车的人，这类评论在收集的语料中占绝大多数，他们对汽车的评论基本停留于感性认识，或者略带偏颇有失公正，大多是从非专业的角度进行评论，因此，驾乘感受成了评论的主要依据，评论一般篇幅较长而且影响情感类别判定的干扰因素较多。从评论表现形式来看，具体有如下几种情况：

媒体评论，如："高档与高性价比一肩挑"，"作为国内中高档车市场的主力干将，别克君威一直口碑素著"，"奇瑞 A520 新时代的大众情人"，"标致 206, 真的很标致"，"郁闷，严重怀疑 bmw3 质量，首保还没去先进修理车间了!!!"，"我后悔拥有 307"，"伪劣广本 3.0 劣迹补充"，"polo 的烂空调"。

专业评论，特点是汽车各项参数居多，对判断情感有一定的作用，但是很难从参数上直接得出情感倾向性信息，如："固特异 205/55R16 的宽扁平比"，"宝来 1.8T，配载了大功率直列 4 缸，5 气阀涡轮增压发动机（5vTurbo），最大功率为 110KW（150 马力），排量为 1.8 升，压力化为强大的动力，在 1750 转/分一直到 4600 转/分的转速下可持续提供足够 210Nm 扭矩"。

日志式评论，一般以时间为序，篇幅很长，且内容很散乱，褒贬之语反复交替出现，如："买了奇瑞 A520，跟 XDJM 做个汇报，不开不知道，油耗惊人地少！1、4 月 28 日提车时起始里程 15 公里，加油 51.4 升。2、新车磨合期，缓加速，尽量不踩刹车，……。3、其余 258 公里都是城市道路，开空调，缓加速，尽量不踩刹车。……小结：自己的爱车自己爱护，新车磨合期，缓加速，尽量不踩刹车，跑长途实属有事，平均油耗比较满意，动力感觉不错，四个大人两个小孩还有 50 公斤货物爬比较陡的坡时跑 3 档感觉轻松，没发现任何问题"。

大众化评论，这类评论基本上有两种形式：一是申明观点、比较、下结论的格式，但大多数情况是观点结论与中间评论内容不符，例如，虽然对该车有很多不满，但结论仍为正面的，这对研究情感的判定是个较大的影响。如："我的 QQ 车是 2 月

19 号买的，到现在开了 560 多公里了，我的初步感觉是：一、非常满意的地方：1、车的形态自然不必说了，……；2、颜色漂亮鲜艳，……；3、动力不错，……；4、发动机……；5、各种配置……；二、不满意的地方：1、内装饰粗糙，比如……；2、减震不太好，……!! 3、杯子座也太小了！ 4、挡不好挂，……。我是根据我开我的小 Q500 多公里初步感觉实事求是的说的这些，总体上我是喜欢 QQ 的！我也希望奇瑞能够多听听车主们的意见，不断改进，我支持奇瑞！"；二是对比类型的，这种类型评论通常是多种车型进行对比，褒贬的情感用语反复出现于多种车型，给评论情感倾向性判定带来很大的挑战；如："帕萨特和雅阁，安全设备比较：1，帕萨特的 ABS 含 EBD，EBA；雅阁的也是 ABS 含 EBD；2，帕萨特的 1.8T 含 EDS；雅阁 2.4 没有类似 EDS 的东东；3,帕萨特的 V6 含 ESP；雅阁的 V6 没有类似的东东；4，帕萨特的刹车系统从不跑偏；雅阁 2.3 的刹车跑偏是一个通病；5，帕萨特自动波的车有真空泵提供刹车安全保障； 6，帕萨特的后悬挂是非独立悬挂，注重牢靠； 雅阁的后悬挂介于帕萨特和 JW 之间……；21，帕萨特的防盗系统是密匙式的；雅阁的防盗，几乎是不设防。尽管如此，我还是喜欢我的帕萨特！"。

基于以上的分析，语料中评论倾向判别比较复杂，对于正面、反面类别的判定是由三个判断者人工进行判别，以减少人为的误差。

2.3 语料预处理

对上述语料库一（corpus 1）和语料库二（corpus 2）的评论文本分别进行如下处理：

删除非文本信息

扫描语料集，删除语料中存在的乱码、标识符、图形等非文本信息。

将数据集进行大字符集的扩展

原始语料是基于 GB 汉字编码，对于 GB 码以外的汉字，使用了特定的符号表示，因此，将这些符号转换成 GBK 汉字，除此以外的其他字符全部删除，目的是将数据集扩展，以适应之后的处理。

删除超短文本和超长文本

在语料中存在一些很短的评论，在这种情况下，如果保留这些文本，在降维以后，这些文本很容易变成空文本，无法获得信息进行情感分类。同样，对于超长的文本，其所含的特征信息可能会掩盖相关评论中的信息，同时基于 2.2 节分析较长的评论文本通常对正确分类干扰较大。因此，长度小于 20 和大于 3000 的评论文本都

被删除。

进行语料的切分处理

这里我们采用山西大学的汉语分词与词性自动标注系统 FC2000，对全部语料进行分词、词性标注处理，其分词结果形式如："奥迪 nz 公司 n 将 p 具有 v FSI ws 燃油 n 直 a 喷 v 技术 n 的 u 3.2 m V ws 6 m 发动机 n 配备 v 在 p 了 u 国产 n 奥迪 nz A ws 6 m L ws 07 m 年 nt 型 k 3.2 m FSI ws 和 c 3.2 m FSI ws quattro ws 车型 n 上 nd 。w"。

通过观察上面很小的一个例子，我们会发现经过分词的语料存在许多错误，如："3.2 m V ws 6 m"，"A ws 6 m L ws 07 m 年 nt 型 k"，"3.2 m FSI ws quattro ws"、"四 m 气门 n"、"进气 n 歧 n 管 v"、"进 vd 排气 v 凸轮轴 n"，"V ws 6 m 发动机 n"、"6500 m 转 v / ws 分钟 q"、"188 m 千瓦 q （ ws 255 m 马力 q ） ws"、"无 v 级 n / ws 手动 f"、"里 q / ws 小时 nt"、"7.4 m 升 q / ws 百 m 公里 q"。在汽车评论语料中这样的分词错误非常多，严重影响我们的研究。为此，我们采用建立汽车产品知识库的方法解决语料分词中错分的问题。

2.4 汽车产品知识库和情感词汇库

依据上节分词系统对语料分词后，我们发现由于分词系统的领域无关性，以及汽车评论中车名，专用术语等难以辨认，分词中存在着大量的错误，主要包括车名切分错误，如"标致"词性标注为 a，实际为 n 表示汽车名称，"宝""马"被切分开，"马""自""达"被切分开；部件或专用术语切分错误，如"变速""箱"、"性价""比"、"安全""气囊"、"手动""档"、"最大""功率"、"倒车""雷达"等错误切分。因此，我们建立了汽车产品知识库，该知识库主要收集了汽车产品领域相关的术语和指标名称，及具体参数形式。具体过程如下：

（1）收集了语料相关的汽车品牌、车系、型号。见图 2-1

（2）收集汽车零部件、性能指标等要素。见图 2-2

在语料中存在用语不规范现象，如："标致 307"常被表示为"小狮子"，"帕萨特"表示为"pst"、"帕帕"、"小帕"等昵称；"ABS"、"GPS"、"EPS"、"Santana"、"BUICK"等术语名称中英文混用，我们对这类词汇与相关汽车产品名称和零部件进行了关联，使得研究当中可以达到同一实体不同表达形式的相同识别，在分类的过程中可避免因此而引起的错误。

最后，我们利用汽车产品知识库，对评论中相关术语进行校对，基本消除了汽

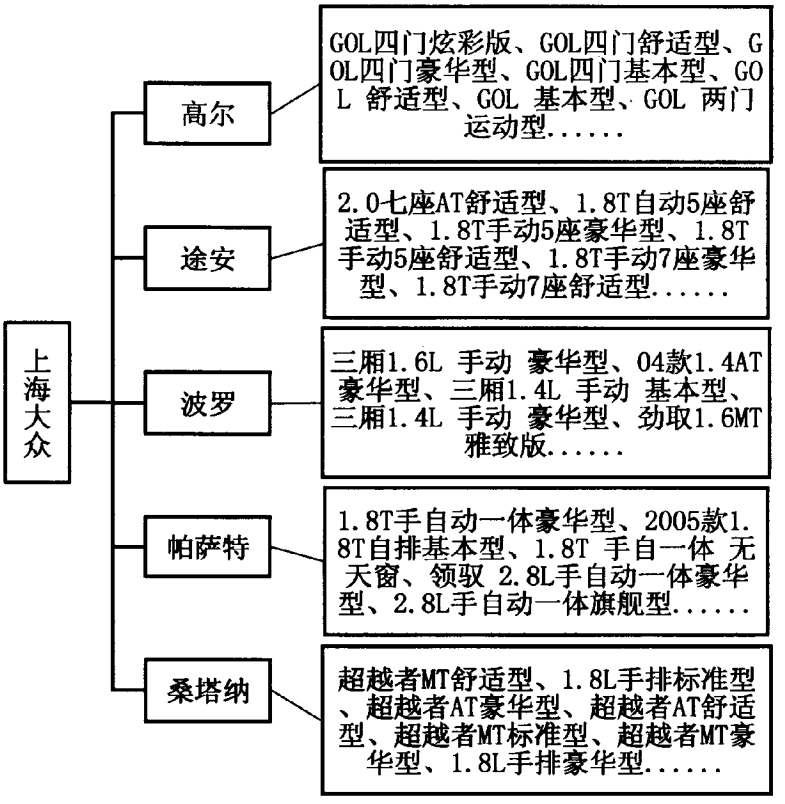

图 2-1 汽车知识库示例

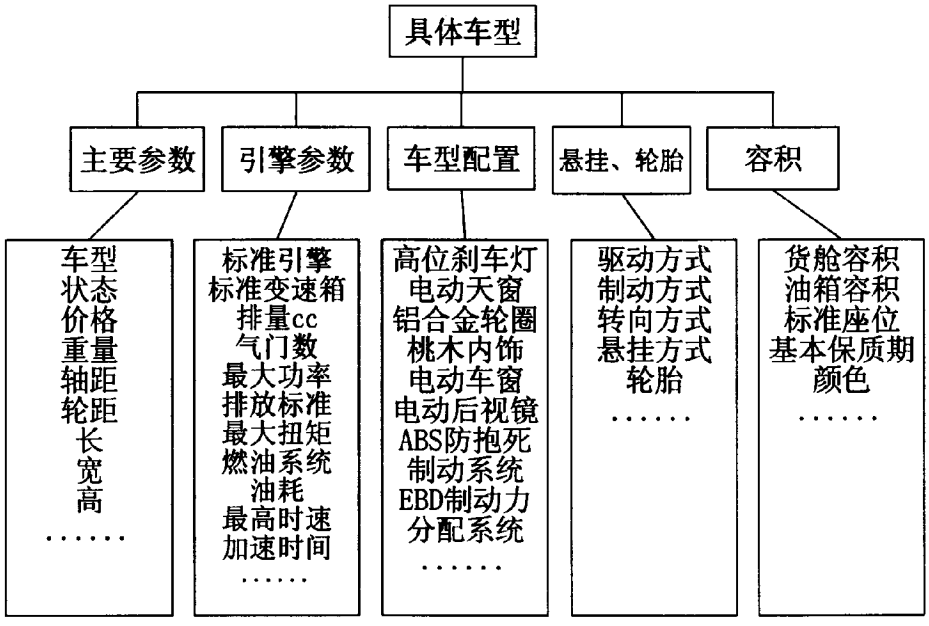

图 2-2 汽车参数指标要素

车产品及其领域相关的术语分词错误。这对后期的研究至关重要。只有关键词汇切分正确的前提下，才能很好研究情感分类问题。

在汽车评论中存在较多的情感词汇，他们对判定评论情感的倾向作用较大，这些词汇的倾向性非常明确，在评论文章中所表达的情感倾向清晰，如："好"、"漂亮"、"著称"、"庄重"、"尊崇"、"尊贵"、"卓绝"、"拙劣"、"圆满"、"圆润"、"优点"、"优惠"、"优美"、"优胜"、"优雅"、"优异"、"优越"、"优质"、"遗憾"、"遗漏"、"厌恶"、"压迫"、"压抑"、"虚荣"、"虚伪"、"痛心"、"偷盗"、"偷懒" 等，十分有必要对这类词汇进行细致的归纳和总结。为此，我们对目前已有的情感词汇语言资源进行了整理，主要包括：GI 词典中文翻译版、《学生褒贬义词典》[34]、知网[35]、《褒义词词典》及《贬义词词典》等。通过将上述词汇字典进行综合，建立了汽车产品评论情感倾向词汇库。

2.5 实验平台

本论文的主要实验都是在 Windows XP 操作系统下完成，采用 Delphi 作为开发工具，SQL2000 作为数据库，软件设计采用面向对象的方法。采用 Matlab 6.5 SVM 软件包计算的数据结果进行方法对比。硬件配置为：Pentium III 933 MHZ，256MB RAM，40GB HD。

2.6 本章小结

本章我们构建了两个关于汽车产品评论的语料库。在深入分析语料的基础上，认为汽车产品评论能真实反映出情感倾向性分类问题中特有现象和突出问题，并在此基础上针对性的建立了汽车产品知识库和汽车产品评论情感词汇库，在随后各章节的研究中我们都是在此数据资源的基础上开展的。

第三章 停用词表对中文文本情感分类的影响

传统的中文文本分类通常是基于主题分类，停用词表对其分类结果有较大的影响[36]。G.W.Hart 发现[37]，在典型英文段落中所用词的 50%可以包含在一个具有 135 个词的普通词表中，应在文本分析预处理中去除；Yang 和 Pedersen 认为[38]，若对停用词按照其出现的文本频数降序排列，仅用前 10 个停用词降低特征向量空间维数，不会产生负面影响；用前 100 个停用词降低特征向量空间维数，所产生的负面影响非常小，但再大一些，效果会有明显的影响。与此同时，Silva 验证了应用停用词表降低特征空间的维数，对提高文本分类器的准确率会产生积极的作用[39]；顾益军等提出联合熵自动获取停用词表[36]。

对汽车评论语料的情感倾向性进行分类，可以将其看成判断评论正面倾向、反面倾向的二分问题：假设预定义的文本类型集为 $Tain=\{P,N\}$，其中 P 表示对相关汽车持正面态度的评论，也可称正面的（Positive），N 表示对相关汽车持反面态度的评论，也可称反面的（Negative）。待分类的文本集为 $Test=\{d_1,d_2,.....d_n\}$，情感分类的任务就是将文本集 $Test$ 中的文档 $d_i(i=1,2,.....,\ n)$ 自动判断为正面或者反面。

对于情感分类来说，将什么样的词作为停用词以及停用词对其分类结果的影响还没有看到相关的报道。本章选用了五种停用词表，采用常用的三种特征提取方法[1]：信息增益（IG）、互信息（MI）和 χ^2 假设检验（CHI），两种权重的计算方法：基于文档和基于词频，利用支持向量机分类方法[39]，分别考察了五种停用词表对汽车评论的情感类别判断的影响，为后续情感分类问题的进一步研究提供了参考和依据。

3.1 支持向量机

支持向量机（SVM）[40]是建立在小样本统计学习理论的 VC 维理论和结构风险最小化原理基础上的一种新型机器学习方法。支持向量分类机能寻找一个满足分类要求的最优超平面，使其在保证分类精度的同时最大化超平面两侧的空白区域，使得 SVM 分类器的分类结果不仅在训练集上得到优化，而且在整个样本集上的经验风险最小。SVM 的基本思想是：对于两类待分样本，设 H 为把两类待分样本没有错误地分开的分类超平面，H_1、H_2 分别为过两类样本中离分类超平面最近的点且平行于分类超平面的超平面，H_1 与 H_2 之间的距离叫做分类间隔（Margin），如图 3-1 所示。所谓最优分类超平面是指能将两类文本正确分开（训练错误率为 0），且使分类间隔

最大的超平面。前者是保证经验风险最小（为 0），后者实际上使推广性的界中置信范围最小，从而使真实风险最小。

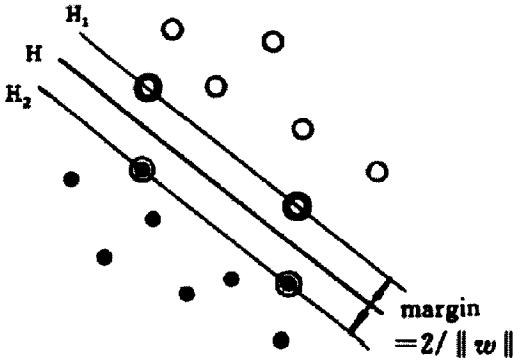

图 3-1 最优分类超平面

设样本集(x_i,y_i)，$i=1,\cdots,n$，$x_i\in R^d$，$y_i\in\{-1,+1\}$是类别标号。d 维空间中线性判别函数的一般形式为$g(x)=w\cdot x+b$，分类超平面方程为：$w\cdot x+b=0$，将判别函数归一化,使两类所有样本都满足$|g(x)|\geq 1$,使离分类超平面最近的样本的$|g(x)|=1$，此时分类间隔等于$2/\|w\|$，使分类间隔最大等价于使$\|w\|$（或$\|w\|^2$）最小。要求满足分类超平面对所有的样本正确分类，就是要求它满足：

$$y_i[(w\cdot x_i)+b]-1\geq 0, i=1,\cdots,n \tag{3.1}$$

因此，满足条件（3.1）且使$\|w\|^2$最小的超平面称为最优分类超平面，H_1、H_2上的训练样本称作支持向量。

利用最优化理论求得w的解w^*，然后利用支持向量求出b的解b^*，这样就得到了最优分类函数

$$f(x)=\mathrm{sgn}\{(w^*\cdot x)+b^*\} \tag{3.2}$$

根据上述公式（3.2）（最优分类函数）就可以确定输入样本的类别。

3.2 特征选择方法

特征选择就是从特征集$T=\{t_1,...,t_s\}$中选择一个真子集$T'=\{t_1,...,t_{s'}\}$（s'<<s）。其中：s为原始特征集的大小，s'为选择后的特征集大小。选择的依据是特征对分类作用的大小，通常用一个统计量来度量，选择没有改变原始特征空间的属性，只是从原始特征空间中选择了一部分重要的特征，组成一个新的低维空间。

本章采用信息增益（IG）、互信息（MI）、χ^2统计量（CHI）[5]等常见的特征

选择方法。首先对候选特征计算其度量值，然后根据预先设定的阈值T，将度量值大于T的特征选为有效特征。

假设c为文档类变量，C为文档类的集合，d为文档，f为特征。

（1）信息增益（Information Gain）

它表示特征在文本中出现或不出现为确定文本的类型所提供信息量的大小。对于特征f，其信息增益$IG(f)$被定义为：

$$\begin{aligned} IG(f) &= H(C) - H(C|f) \\ &= -\sum_{c\in C} p(c)\log(p(c)) + p(f)\sum_{c\in C} p(c|f)\log(P(c|f) + P(\bar{f})\sum_{c\in C} p(c|\bar{f})\log(P(c|\bar{f}) \\ &= \sum_{c\in C}(P(c,f)\log(\frac{P(c,f)}{P(c)P(f)}) + P(c,\bar{f})\log(\frac{P(c,\bar{f})}{P(c)P(\bar{f})})) \end{aligned} \tag{3.3}$$

其中：$P(c,f)$为c中出现特征f的文本数除以训练集中出现f的文本数；$P(\bar{f})$为训练集中不出现特征f的文本数除以训练集的大小；$p(f)$为训练集中出现f的文本数除以训练集的大小；$p(c)$为训练集中属于类型c的文本所占的比例；$p(c|\bar{f})$为类型c中出现f的文本数除以训练集中出现f的文本数。

IG定义为某一特征在文本中出现前后的信息熵之差，有利于高频特征。在进行特征选择时，应选择信息增益大的特征。

（2）互信息（Mutual Information）

在统计学中，互信息用于表征两个随机变量间的相关性。对于特征f，其互信息$MI(f)$被定义为：

$$MI(c,f) = \log(\frac{P(c,f)}{P(c)P(f)}) \tag{3.4}$$

其中：$P(c,f)$为c中出现特征f的文本数除以训练集的大小；$p(f)$为训练集中出现f的文本数除以训练集的大小；$p(c)$为训练集中属于类型c的文本所占的比例。

MI来源于信息论，表示特征与类型之间的相关程度。当特征出现只依赖于某一类型时，特征与该类型的互信息很大；当特征与类型相互独立时，互信息为0；当特征很少在该类型文本中出现时，它们之间的互信息为负数，即负相关。

（3）χ^2统计

χ^2统计也是用于表征两个随机变量间的相关性的统计量，但它比互信息更强，因为它同时考虑了特征出现与不出现两种情况。对于特征f，其χ^2统计值被定义为：

$$\chi^2(c,f)=\frac{(P(c,f)P(\bar{c},\bar{f})-P(c,\bar{f})P(\bar{c},f))}{P(c)P(f)P(\bar{c})P(\bar{f})} \tag{3.5}$$

其中：$P(c,f)$为c中出现特征f的文本数除以训练集的大小n；$P(\bar{c},\bar{f})$为训练集中不出现特征f并且不属于类型c的文本数除以训练集大小n；$P(c,\bar{f})$为训练集中出现特征f并且不属于类型c的文本数除以训练集大小n；$P(c,\bar{f})$为训练集中不出现特征f并且属于类型c的文本数除以训练集大小n。

上述公式（3.3）～（3.5）中的概率值的估算分别采用如下公式：

$$P(c)\approx\frac{N_c}{N} \tag{3.6}$$

$$P(c,f)\approx\frac{N_{fc}}{N} \tag{3.7}$$

$$P(f)\approx\frac{N_f}{N} \tag{3.8}$$

这里，N_{fc}为类c中特征f出现的频数，且

$$N_f=\sum_{c\in C}N_{fc} \tag{3.9}$$

$$N_c=\sum_{f\in F}N_{fc} \tag{3.10}$$

又有，

$$P(c,\bar{f})=P(c)-P(c,f) \tag{3.11}$$

$$P(\bar{f})=1-P(f) \tag{3.12}$$

$$P(\bar{c})=1-P(c) \tag{3.13}$$

$$P(\bar{c},f)=P(f)-P(c,f) \tag{3.14}$$

$$P(\bar{c},\bar{f})=P(\bar{f})-P(c,\bar{f}) \tag{3.15}$$

据此我们有，若$P(c,f)=0$，则令$P(c,f)\log(P(c,f)/(P(c)P(f))$和$\log(P(c,f)/(P(c)P(f))$为0。

3.3 权重计算方法

根据3.2介绍的三种特征选择方法和预先设定的阈值，可以抽取出对文本情感倾向分类起作用的特征，可以选择出对文本分类起作用的特征，但不同的特征对文档类别的重要程度和区分度是不同的。要将待处理的文本表示为向量的形式，计算特征在此文本中的权重（特征的取值），向量的维数即为特征的个数。本文中，我们采用布尔权重（Boolean Weighting）和频度权重（Frequency Weighting）两种权重形式：

（1）布尔型特征权重：布尔权重的特征权重计算只存在0和1两个值，描述特征的过程中会丢失大量的信息。

$$w=\begin{cases}1 & \text{若特征 } c\text{出现在}f\text{中}\\ 0 & \text{否则}\end{cases}$$

（2）词频型特征权重：认为特征在文本中出现次数越多，就越重要。以特征c出现在文档f中的频次t作为该特征的权重，即

$$w=t。$$

3.4 停用词的选择

传统文本分类中的停用词是指对文本主题没有描述能力和区别能力的词，是一些噪声词。而对情感分类问题来说，特征的选取就是要选择那些既带有情感色彩又具有类别区分能力的词。在英文中，人们首先选择名词（n）、动词（v）、形容词（a）、副词（d）确定为候选特征[14,23,41]，而中文[42]具有情感色彩的确定为名词（n）、动词（v）、形容词（a）、副词（d）、区别词（f）、叹词（e）、拟声词（o）、代词（r）、成语、简称等作为具有情感色彩的词。为了测试选用不同的候选特征对于情感倾向性分类的影响，我们构造出下面五种不同的停用词表作为选择候选特征的依据。

（1）Stoplist 1 不含动词（v）、形容词（a）、副词（d）的停用词表。即将动词（v）、形容词（a）、副词（d）作为候选特征。

（2）Stoplist 2 不含名词（n）、动词（v）、形容词（a）、副词（d）的停用词表。即将名词（n）、动词（v）、形容词（a）、副词（d）作为候选特征。

（3）Stoplist 3 不含名词（n）、动词（v）、形容词（a）、副词（d）、区别词（f）、叹词（e）、拟声词（o）、连词（c）的停用词表。由于汉语的句式对情感分类有很大的影响，比如转折句中情感描述主要在于后半分句，因此，候选特征将名词（n）、动词（v）、形容词（a）、副词（d）、区别词（f）、叹词（e）、拟声词（o）、连词（c）作为候选特征。

（4）Stoplist 4 主题停用词表：采用李荣陆[1]的停用词表。

（5）Stoplist 5 无停用词表，即将所有的词作为候选特征。

3.5 中文文本情感倾向性分类的步骤

对于中文文本情感分类，其步骤如下：

Step1 将语料库文本进行分词、词性标注预处理；

Step2 将预处理文本集随机分为训练集和测试集，比例约为4:1；

Step3 分别使用上述的五种停用词表，得到候选特征；

Step4 对每一种候选特征，通过特征选择方法，计算其度量值，然后根据设定的阈值T，将度量值大于T的候选特征选为文本的情感倾向分类特征；

Step5 利用权重计算方法，得到每个文本中的情感特征权重；

Step6 利用情感特征与权重，将文本表示成特征向量形式；

Step7 利用文本分类器（支持向量机）训练文本；

Step8 利用训练过程得到的文本情感倾向分类器，对测试集文本的情感倾向类别进行判断。

3.6 实验结果与分析

3.6.1 评价指标

评价指标是在测试过程中所使用的一些用来评价分类性能的量化指标，通常采用的分类评价指标有查全率（Recall，简记为r）、查准率（Precision，简记为p），（以后各章节相同）其定义如下：

反面查全率 $RN=\frac{a_1}{c_1}$　　反面查准率 $PN=\frac{a_1}{b_1}$

正面查全率 $RP=\frac{a_2}{c_2}$　　正面查准率 $PP=\frac{a_2}{b_2}$

总体查全率 $F1=\frac{a_1+a_2}{c_1+c_2}$　　总体查准率 $F2=\frac{a_1+a_2}{b_1+b_2}$

其中：a_1表示系统判断为反面的文档与实际应为反面的文档相等的数量，a_2表示系统判断为正面的文档与实际应为正面的文档相等的数量；c_1表示实际应为反面的文档数量，c_2表示实际应为正面的文档数量；b_1表示系统判断为反面的文档数量，b_2表示系统判断为正面的文档的数量。由于本文判断的是正反两类问题，因此，$c_1+c_2=b_1+b_2$，即$F1=F2$，这样混合两种类别的查全率和查准率应为相等。在实验中，我们仅考虑$F1$。

3.6.2 中文文本情感倾向性分类实验与分析

为了减少训练语料与测试语料对测试结果的影响，本实验采用五次交叉检验，考察五种停用词表对汽车评论情感类别（正面、负面）判断的影响。实验采用特征维数为4000维，其结果见表3-1，为了反映其结果的趋势，将各分类果的F值绘成图3-2。

由图3-2可以看出，在4000维是要总体优于其他情况，在2000和3000维的特征空

表 3-1：五种停用词表、三种特征抽取方法以及两种权重计算的情感分类正反面、综合的 F 值

停用词表	评价 F	基于文档			基于词频		
		IG	MI	X^2	IG	MI	X^2
1	正面	83.90	83.33	82.24	81.12	84.50	81.43
	反面	68.32	65.97	63.41	62.18	68.64	63.42
	综合	78.48	77.11	77.70	74.44	78.94	76.12
2	正面	82.35	82.22	82.86	83.25	81.52	82.73
	反面	64.63	66.03	64.56	65.91	65.57	63.95
	综合	76.73	74.95	76.95	76.40	76.45	75.18
3	正面	82.62	81.19	83.46	82.45	82.27	80.90
	反面	65.85	63.28	66.62	65.85	65.64	59.64
	综合	76.19	76.19	76.69	77.11	75.40	76.43
4	正面	81.63	80.39	83.32	80.60	80.51	81.77
	反面	64.20	60.51	66.88	60.36	61.98	62.32
	综合	75.40	73.68	77.70	73.57	74.18	73.94
5	正面	85.36	78.31	83.80	83.23	77.97	84.11
	反面	71.29	52.86	68.77	66.00	51.98	65.92
	综合	80.31	70.11	78.45	77.23	69.43	78.21

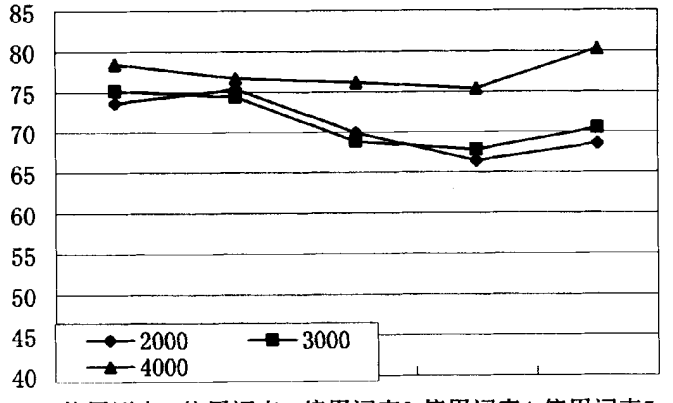

图3-2 不同维数对停用词表的情感分类结果F值

表 3-2：五种停用词表正反面特征同现情况

停用词表	正、反面同现特征数
1	1413
2	1582
3	1582
4	1453
5	1656
1-5	524

间下，特征携带的信息不足，这也反映出分类器在文本情感分类中的不足，不能更好地适应在低维特征空间分类。

由表 3-1 和图 3-3 可以看出：

（A）从语料的正反查全率、查准率来看，正面查全率、查准率比较高，主要原因一是正面的语料规模比反面语料的规模大。我们已做了测试，当反面评论数量增加时，反面的查全率和查准率、总体评价性能都会得到相应的提高；二是语料自身的一些特点（如：评论本身正面、反面用语交替出现，评论作者有失公正的评价，即虽然该车问题很多，但依然持正面的积极评价）也是导致这一结果不可忽视的原因。这也说明采用统计机器学习的方法其结果可解释性较差，在情感倾向分类这个较新的问题中还有很多问题需要深入研究。

（B）从特征抽取方法看，基于 IG 特征抽取方法在各种停用词表的情况下，分类效果最好，χ^2 次之，MI 方法最差。

（C）从权重计算看，利用停用词表 1 得到的分类结果，词频型与布尔型一致，其余的停用词表，词频型权重比词频型权重的整体分类效果要好，这与英文的情感分类问题研究结果的认识是一致的[15]。

（D）以停用词表 2 和停用词表 3 作为停用词的筛选，各种特征选取方法得到的分类结果相差不大，大部分集中于 74%－77%，说明去掉这两种停用词后对特征选择方法影响不大；对于不使用停用词表（停用词表 5）时各个特征抽取方法得到的分类结果波动最大，最好的性能达到了 80.3%，最差的为 69.43%，相差超过了 10%；对于选用停用词表 4 和停用词表 1 时各个特征抽取方法得到的结果波动低于无停用词表（停用词表 5）的情况，但使用停用词表 1 时各个特征抽取方法得到的性能比选用停用词表 4 的整体效果好。总的来说，不使用停用词表（停用词表 5）与停用词表

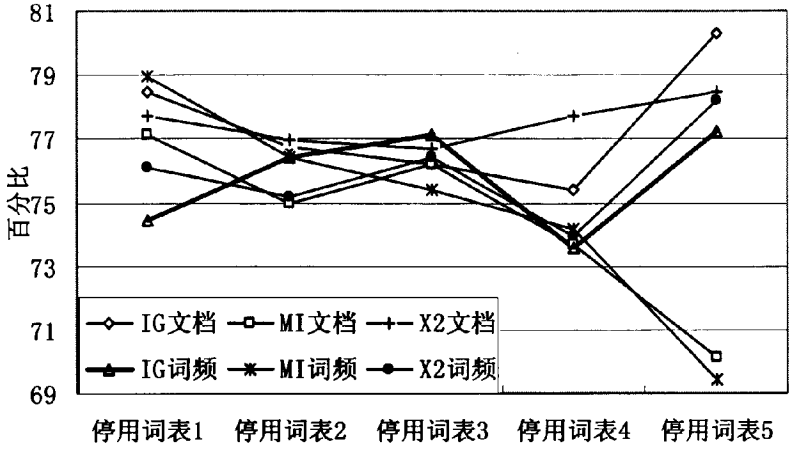

图 3-3：五种停用词表、三种特征抽取方法以及两种权重计算的情感分类结果

1（不含 avd）时各个特征抽取方法得到的结果较好，这与人的直觉认识是一致。

（E）从获取的特征，在五种停用词表中正面特征有 1571 个词同现，反面特征中同现 1329 词。如表 3-2 所示，无停用词表（停用词表 5）时正面反面共现特征为 1656 词，主题停用词表（停用词表 4）中共现 1453 词，不含 navd 停用词表（停用词表 2）共现 1582 词，不含 navdcef 停用词表（停用词表 3）1582 词，不含 avd 停用词表（停用词表 1）1413 词，综合上述情况，正面和反面共现 524 词，可见这 524 个特征词对情感分类的作用是非常大的。利用五种停用词表得到候选特征后，再使用信息增益进行特征选择，得到了 524 个特征的交集，部分特征为“拒绝、宽敞、失望、简洁、愉悦、投诉、宽大、张扬、灵敏、优雅、平稳、灵巧、方便、优异、适中、非常、精细、流畅、良好、烦、惬意、喜欢、饱满、富有、精细、青睐、新颖、圆润、满足、轻松、出色、维修、无奈、别致、隐藏、安全、严重、实用、智能化、强劲、清新、享受、信赖、足够”等，表明这 524 个特征不论采用什么样的停用词表均被选为特征，即它们对情感分类的作用是比较大的。

3.7 本章小结

相对文本主题分类而言，情感倾向性分类问题更为复杂，选择刻画具有情感倾向性类别的特征更加困难。本章采用网上下载的中文汽车评论语料库（corpus 1）作为训练与测试语料，针对三种特征抽取方法信息增益（IG）、互信息（MI）与 χ^2 假设检验（CHI），两种权重计算方法文档频率与词频统计，以及特征维数对文本情感分类的影响进行了比较研究，在五种不同的停用词表中利用支持向量机分类器对文本

情感分类进行了实验。结果表明，基于传统的主题分类的关键技术可以应用于文本情感倾向分类，信息增益（IG）和布尔型对情感分类的效果整体比较好，这与英文文本的情感分类报道结果相一致。当选用不同的停用词表时，它们文本情感分类的影响也不尽相同，停用词表 5（不去掉停用词）、停用词表 1（仅选用形容词、动词和副词）对情感分类作用较大，整体性能效果较好。通过实验同时发现对于情感倾向性的判定中，特征词应该是形容词、副词、动词、助词、感叹词等与情感表达相关的带有较强情感色彩的短语或词汇，但实验的结果是“路程、外方、GPS、眼球、桑塔那”等名词居多，它们与主题相关。另一个现象是很多正面的形容词、副词、动词，如“巨大、完好、有效、风范、热销”同时也出现在反面评论中，使得文本情感分类的各项性能指标均低于已有相关主题文本分类报道的结果[1]，说明了中文文本情感分类比主题分类更复杂。通过对于汽车评论的情感分类，帮助用户对所关注的汽车表现有一个整体了解，而且还可对汽车产品起到推荐作用。

为了在主题分类的关键技术的基础上，提高文本情感倾向分类的效果，我们在第四章和第五章中采用混合的分类方法和基于粗糙集的分类方法研究文本的情感倾向性，同时将研究具有情感倾向的词典作为情感特征来源，并用于文本情感倾向分类，提高其分类性能。

第四章 文本情感分类的混合特征选择方法

特征选择作为文本情感分类的关键技术之一，本章对特征选择方法进行了进一步的研究，提出了混合的特征选择方法。该方法是基于情感词汇库和上一章介绍的信息增益（IG），讨论了特征空间不同维数对分类结果的影响。

在第三章中使用的信息增益（IG）、互信息（MI）、χ^2统计量（CHI）的特征选择方法，这些方法虽然在英文文本分类问题中表现良好，但在我们的实验中它们的表现却不理想。经过分析发现，造成这种差别的原因来自两方面：评论文本中低频词较多且很多被选为特征；文本向量表示的空间候选特征较多，造成了维数空间的过高，这与文献[43]的认识是一致的。从方法上来看，信息增益（IG）、互信息（MI）、χ^2统计量（CHI）三种特征选择方法本质上都是通过不同的方式使用词汇的类别信息。具体来说， IG计算$p(c|f)$和$p(c|\overline{f})$的值，并同类别的概率一起度量词汇携带的类别信息；对于MI，它的定义等价于：$\log(p(f|c))-\log(p(f))$ ，其中$\log(p(f|c))$即为类条件概率；χ^2的定义式中的$P(c,f)$、$P(\overline{c},\overline{f})$、$P(c,\overline{f})$和$P(\overline{c},f)$都表示在训练文档中词条的类别信息。从实验来看，也验证了方法的词汇类别信息的利用方式。

当训练语料库的规模较小没有达到一定规模的时候，特征空间中必然存在很多数量的出现频数很低（比如低于五次），甚至不出现的特征词汇，因为它们较低的出现频数，必然只属于少数的类别，造成特征空间的稀疏。而使用类别信息的统计方法必定认为这些低频词携带较为强烈类别信息，从而对它们有不同程度的依赖，同时文本长度的不一，更加重了这一问题。事实上，这些低频词中只有不到20%的词确实带有较强的类别信息，大多数的词都是噪音词，它们对分类并不能起到积极的作用，不应该被选为特征。

在英文文本的分类问题中，通常取单词和短语作为特征，特征空间的维数相对较少。在训练语料的规模适度大（这样的规模较容易达到）的情况下，大多数词条都可以获得较高的频数，使得IG、MI和CHI对低频词的依赖性得到减弱。

在中文文本中分类应用中，通常将单个的词条作为特征。如果不考虑停用词等未登陆词，可以近似认为特征空间维数等于分词词典中的词条数目。在中文处理中，通常采用的分词词典的规模一般在5万到25万词条之间[16]。也就是说，中文的特征空间维数比英文更高，而且可能高很多。在相同规模训练语料条件下，更高的维数必然导致更多的低频词出现。在这样的情况下使用IG、MI和CHI进行特征选择，由于

它们对低频词的依赖，必定会将更多的低频词作为特征使用，从而导致了分类效果的低下。

根据分析，要使用类别信息提高分类效果，必须消除相应的特征选择方法对低频词的依赖。一个解决方法是增加训练语料的规模，使得所有或至少绝大多数词条在语料中的出现频数都超过一定阈值，但是由于中文特征空间的维数如此之高，要达到该目的训练语料至少要达到GB级。在实际设计文本分类器的时候，如此大规模的训练语料通常难于获取（只能通过搜索引擎查询返回值，可以认为是在无限大语料搜索）。再者IG、MI和CHI的计算复杂程度都为$O(N^2)$，统计如此大规模的语料将会花费很高的计算成本（高性能计算设备，时间和空间）。

在只有通常规模训练语料的条件下，充分利用类别信息的一个可行方法是使用组合的特征选择方法。

4.1 混合特征选择方法

在传统的主题文本分类中，特征的选择是在经过停用词筛选后的剩余词汇中产生的。然而对于情感倾向性分类问题，应该选择尽可能少而准确且与文本表示的情感概念相关的文本特征。选择什么样的文本特征由具体的度量和特征情感倾向确定。目前人们常采用多种特征选择融合的方法。我们基于词汇的类别区分能力和信息增益（IG）进行选择。

为了选出类别区分性好的候选特征，我们设计了五种与类别信息相关的候选特征选取方法。如下：

P——正面类别

N——反面类别

PIW——属于P类且包含特征项 f 的训练文本个数

NIW——属于N类且包含特征项 f 的训练文本个数

PEW——属于P类且不包含特征项 f 的训练文本个数

NEW——属于N类且不包含特征项 f 的训练文本个数

M——属于P类的训练文本个数

N——训练文本总数

其中，f 为某个特征项。显然，N-M为属于N训练文本个数，$PIW + PEW = M$，$PIW + NIW + PEW + NEW = N$。

Scheme1 利用信息熵来描述特征 f 在语料中正面和反面的分布情况。假设特征 f 在正面文档和反面文档中的分布是均匀时，特征 f 对分类结果没有贡献。对分词后的词语，计算其在训练语料中的熵值，将熵值较大的词语去掉，剩余的词语作为候选特征。

某个特征 f 的熵计算公式如下：

$$H(f) = -(p(P/f)\log p(P/f) + p(N/f)\log p(N/f))$$

$$\approx -(\frac{PIW}{PIW+NIW}\log\frac{PIW}{PIW+NIW} + \frac{NIW}{PIW+NIW}\log\frac{NIW}{PIW+NIW}) \quad (4.1)$$

Scheme2 通常训练语料中出现次数较低的词被认为对分类没有贡献，我们将频次为 1 的词语去掉，剩余的词语作为候选特征。

Scheme3 融合 Scheme 1 和 Scheme 2，即去掉频次为 $DF=1$ 并且熵值大的词语，剩余的词语作为候选特征。

Scheme4 如果某类别中大多数文本具有这一特征，则该特征称为某类别的代表性特征；如果其他类别的大多数文本均不具备该特征，则称该特征为该类别的鉴别性特征，因此我们选择那些既具有代表性又具有鉴别性的特征。本文某个特征项 f 与类别的相关性度量用频率差 FD（frequency difference）：

$$FD(f) = (p(f|P) - p(f|N))^2$$

$$\approx (\frac{PIW}{PIW+PEW} - \frac{NIW}{NIW+NEW})^2$$

$$= (\frac{PIW}{M} - \frac{NIW}{N-M})^2$$

$$= (\frac{PIW\cdot NEW - NIW\cdot PEW}{M\cdot(N-M)})^2 \quad (4.2)$$

其中 M 和 N-M 独立于特征 f， 可将公式（4.2）简化成公式 （4.3）。

$$FD(f) = (PIW\cdot NEW - NIW\cdot PEW)^2 \quad (4.3)$$

$FD(f)$ 越大，特征项 f 与类别正反面的区分能力越强。

Scheme5 Fisher 鉴别量，是机器学习中非常有用的一种方法，其思想是希望所选择的特征使 Fisher 准则函数达到最大，即使得类间的距离尽可能的大而类内的距离

尽可能的小，选择特征的 f 具有尽可能大区分能力。同时意味着将整个空间压缩到特征项 f 所在的低维空间，从而达到降维的目的。

对于本文只有正面和反面两种类别，基于 Fisher 鉴别量的文本情感分类候选特征项 f 的鉴别能力的度量计算公式如下：

$$FD(f)=\frac{(E(f/P)-E(f|N))^2}{D(f|P)D(f|N)} \tag{4.4}$$

其中，$E(f/P)$和$E(f/N)$表示特征 f 对正面类别 P 和反面类别 N 的条件均值，$D(f|P)$ 和 $D(f|N)$ 表示特征 f 对正面类别 P 和反面类别 N 的条件方差，$(E(f|P)-E(f|N))^2$ 表示特征项 f 与正、反面类间散度（scatter disperse），$D(f|P)\cdot D(f|N)$表示特征项 f 与正、反面类内散度（scatter disperse）。

比值$\frac{(E(f|P)-E(f|N))^2}{D(f|P)\cdot D(f|N)}$越大，特征项 f 的区分能力就越强。

为了简化计算，将公式（4.4）近似为：

$$FD(f)\approx\frac{(PIW\cdot NEW-NIW\cdot PEW)^2}{PIW\cdot NIW\cdot PEW\cdot NEW} \tag{4.5}$$

4.2 文本情感倾向性分类的步骤

通过上一章的实验表明，信息增益（IG）特征选择方法的结果优于互信息（MI）和 χ^2 检验法，布尔权重方法优于词频权重的方法，这与相关英文情感倾向性分类结果一致。本章的研究将在上一章结果的基础上进行。整个实验分为训练和测试两个部分，语料采用 corpus 1。实验步骤如下：

Step1 将语料库文本进行分词、词性标注预处理；

Step2 将预处理文本集随机分为训练集和测试集，比例约为4:1；

Step3 按照词汇的类别区分能力分别使用上述 Scheme 与 IG 进行特征选择；

Step4 采用布尔权重利用向量表示每一篇文本；

Step5 利用支持向量机（SVM）进行训练；

Step6 利用训练过程得到的文本情感倾向分类器，对测试集文本的情感倾向类别进行判断。

4.3 实验结果与分析

4.3.1 评价指标

沿用上一章评价指标，采用正确率，召回率（Recall，简记为 r）、精确率（Precision，

简记为 p），F 值。根据其分类类别分为正面的、反面的、综合的。

正确率定义为系统分类正确的正面和反面的文本数与系统自动分类的文本数之比。

精确率定义为系统对该类分类正确的文本数与系统自动分类该类文本数之比。

召回率定义为系统对该类分类正确的文本数与实际文本中存在的该类文本数之比。

F 值均衡精确率和召回率，为 2×（精确率×召回率）/（精确率+召回率）。

4.3.2 混合特征选择方法的实验结果与分析

在本章实验中，我们仍采用上一章的评价指标，包括查全率（Recall，简记为 r）、查准率（Precision，简记为 p）、F 值，采用五次交叉方法对正面、反面文档进行测试。实验中采用两种特征选择方法：一种是基于信息增益（IG）；另一种是混合的特征选择方法即基于特征词汇类别区分能力和信息增益（IG）。根据 4.1 中介绍，特征词汇的类别区分能力可以由对应的五种 schemes 度量。因此，这里有五种对应的方法 Approach I（Scheme i + information gain， i = 1，2，……，5）分别在特征选取 1000，2000 和 3000 维下被考察，研究维数对分类结果的影响。实验结果见图 4-1 和表 4-1。

在图 4-1 中，特征选择方法 2，方法 3 和方法 5 的情感倾向性 F 值随着特征维数的增长而得到提高，但是在 F 值的刻画下方法 1 和方法 4 却产生了波动。在特征维数在 3000 时，五种特征选择方法的 F 值均达到各自的最大值，方法 5 的 F 值达到五种方法的最大值，超过了 80%。尽管与方法 5 特征选择上相差不大，但方法 4 的 F 值最差，相差近 1%。

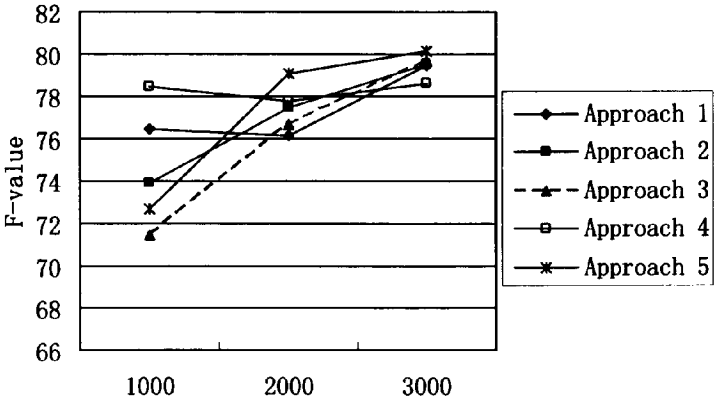

图 4-1. 维数对五种方法的影响

在表 4-1 中可以看出，正面文档的 F 值要由于相应反面文档的 F 值。采用混合特征选择方法的 F 值在 2000 维和 3000 维下要明显优于基于信息增益（IG）特征选择方法的结果。这也表明随着语料库规模的扩大，特征空间的扩大，混合特征选择方法的性能要优于信息增益（IG）特征选择方法。混合特征选择方法更适用于大规模语料的特征选择。总之，情感倾向性分类的结果随着特征区分能力的增强而变得更好，也表明混合特征选择方法在中文文本情感倾向性分类问题中是一种有效的方法。

表 4-1. 实验结果

Feature Dimension	Document Type	Evaluating Criterion（F-value）					
		Approach					Information Gain
		1	2	3	4	5	
1000	Positive	83.18	79.76	78.97	84.20	78.98	84.40
	Negative	62.10	65.13	56.28	66.98	62.00	67.11
	Total	76.46	73.96	71.46	78.46	72.70	78.60
2000	Positive	82.45	83.35	80.17	83.56	84.60	78.32
	Negative	63.72	66.42	64.51	66.75	67.95	57.44
	Total	76.19	77.48	76.70	77.77	79.10	68.55
3000	Positive	84.59	84.73	84.78	84.04	85.27	81.62
	Negative	70.14	70.78	70.28	68.96	70.84	63.53
	Total	79.46	79.52	79.72	78.60	80.18	75.20

由上述实验可知：维数空间较大时，方法 5 的分类性能较稳定，表明该方法可以以较快、较小的代价选择出最具有表征文本类别倾向性的特征来。这对于解决文本分类问题中维数灾难，压缩空间是非常有益的。

4.4 本章小结

通过本章的实验说明，情感倾向性分类问题更为复杂，选择刻画具有情感倾向性类别的特征更加困难。本章采用网上下载的中文汽车评论（corpus1）作为训练与测试语料，采用基于类别区分能力和信息增益（IG）的混合特征选择方法，在不同的特征维数空间下对文本情感倾向性分类进行了研究。实验结果表明混合特征选择

方法中候选特征 4 的情感分类效果最好，而基于类别区分能力度量的候选特征选择方法进行文本情感分类，方法 5 的情感分类结果优于其它四种方法，而五种基于类别区分能力度量的候选特征选择方法都优于单独使用信息增益（IG）特征选择方法。实验中的存在一些现象也值得我们进一步的深入研究。

（1）在混合特征选择方法中，尽管一些词汇是具有情感倾向性的被选为特征，如："严重、先进、喜悦、偷工减料、强烈、轻微、耐用、勉强、快速、华丽、脆弱、高档、刺激、满意、安全、青睐、澎湃、新颖、圆润、满足、轻松、出色、无奈、别致、严重"等，同时另外一些词汇与情感倾向性分类无关也被选为特征，如："距离、变速杆、桑塔那、高度、版本、玻璃、部分、部门、差别、厂家、车展、成绩、乘客、尺寸、出风口、道路、顿挫、范围、方案、方式、方向、风格、高速公路、工艺、公司、国际、红灯、环境、基础"，这些都对实验结果产生了较大的影响。

（2）在真实的网上评论中，如本章所用的语料库，具有反面情感倾向文本的数量小于正面情感倾向性文本数量，这也使得反面 F 值小于正面 F 值，同时导致了整体的 F 值较小。在随后的研究中我们将平衡正面和反面的文本数量，进行深入的研究，结果表明在平衡语料库中反面的 F 值会等到提高，整体的 F 值也会得到提高。

第五章 基于粗糙集的文本情感分类特征选择方法

目前，文本情感分类特征选择方法大都采用传统主题分类中所使用的统计机器学习方法[10,14,24]，事实上由于该问题尚处于研究初期还没有专门解决这类问题的方法和工具。本文的之前章节也利用了多种统计方法进行特征选择对情感分类问题进行了研究，但统计方法的不足是分类结果的可解释性差，数据集要达到一定的规模，同时用统计机器学习的方法解决情感分类问题过程中遇到的一个突出现象是：具有情感倾向性的词汇相比较一般词汇往往出现次数信息不足，导致了具有情感倾向性的词汇在特征选择中很容易被漏掉，造成情感分类结果较差，这也是造成可解释性差的原因之一。上述这些问题都是由方法本身及所研究问题的特点所决定的。

本章将粗糙集理论与文本情感倾向性分类问题结合，实现了基于粗糙集模型的文本情感倾向性特征选择方法，为今后从不同的角度研究中文文本情感倾向性问题提供了新的参考依据。该特征选择方法首先利用向量空间的文本特征表示方法，然后对文本特征的权重采用基于粗糙集的 MD 算法进行离散化，最终获得具有文本情感类别的特征。实验结果表明，基于粗糙集的文本情感特征选择方法是可行的，相比较统计特征选择方法，具有良好的可解释性和低维特征空间适应性。

5.1 粗糙集理论简介

粗糙集理论是20世纪80年代初由波兰数学家Z.Pawlak首先提出的，用于处理不确定知识的数学理论，它能有效的分析和处理不精确、不一致、不完整等各种不完备信息，并从中发现隐含的知识，揭示潜在的规律。

近年来，作为一种新兴的归纳学习方法，粗糙集理论以其“不需对数据的任何先验假设”、“可提供非完备，非协调等不确定性知识获取方法”、“所获知识具有较好的直观可理解性”等显著优势获得人们的广泛关注。

粗糙集理论是以不可分辨关系划分所研究论域的知识，形成知识表达系统，利用上、下近似集逼近描述对象，通过知识约简，从而获得最简知识，下面介绍其基本概念与MD离散化算法。

5.1.1 基本概念

（1）知识、分类及不可分辨关系

知识是基于对对象进行分类的能力，由分类模式组成，利用不同的属性知识描述，可以产生不同的分类。分类主要用来产生范畴，这些范畴构成知识模块，即分

类的类。

不可分辨关系是粗糙集理论的基石。粗糙集理论中以等价关系代替分类，当用R表示论域U中对象之间的等价关系时，则U/R表示U中的对象根据关系R构成的所有等价类。若$P\in R$,且$P\neq\varnothing$，则$\cap p$（p中全部等价关系的交集）也是一种等价关系，称为P上的不可分辨关系，且记为$IND(p)$：

$$[X]_{IND(P)}=\cap[X]_R, P\subset R \tag{5.1}$$

不可分辨关系是对象p由属性集表达时在论域U中的等价关系。它揭示出知识的颗粒状结构。

（2）边界与粗糙度

粗糙集理论中的不确定性和模糊性是一种基于边界的概念，即一个模糊的概念，具有模糊的边界。每一个不确定概念由一对称为上近似和下近似的精确概念来表示：

设给定知识库$K=(U,R)$，对于每个子集$X\in U$和一个等价关系$R\in IND(K)$，可以根据R的基本集合描述来划分集合X：

$$R\_(X)=\cup\{Y\in U/R\mid Y\subset X\} \tag{5.2}$$

$$R^{-}(X)=\cup\{Y\in U/R\mid Y\cap\neq\varnothing\} \tag{5.3}$$

式中，$R_(X)$和$R^{-}(X)$分别称为X的R下近似和R上近似。

集合$Bn_R(X)=R^{-}(X)-R_(X)$称为X的边界域；$Pos_R(X)=R_(X)$称为X的R正域；$Neg_R(X)=U-R_(X)$称为X的R负域。

$R_(X)$是由那些根据知识R判断肯定属于X的元素组成的集合；$R^{-}(X)$是由那些根据知识R判断可能属于X的元素组成的集合；$Bn_R(X)$是由那些根据知识R判断可能属于X又可能属于$U-X$的元素组成的集合；$Neg_R(X)$是由那些根据知识R的判断肯定不属于X的元素组成的集合。论域中等于某个等价类或某些等价类的并集的集合称之为可定义集（精确集），反之称为不可定义集，也称粗糙集。

集合的不确定性是由边界域的存在造成的，由等价关系R定义这种不确定程度（粗糙度）。$Card(\cdot)$表示集合元素数，且$X\neq\varnothing$：

$$P_R(X)=1-\frac{card(R\_(X))}{card(R^{-}(X))} \tag{5.4}$$

（3）信息系统与决策表

信息系统[44]（IS）是一个三元组$S=(U,AT,f)$，其中U是一个非空有限对象集，AT是一个非空有限属性集，f是对象集到属性值域的映射，$f_a:U\rightarrow V_a$，对于任

意$a \in AT$，其中V_a叫做属性a的值域，$Inf(x) = \{(a, f_a(x)) | a \in AT\}$被称为$x$的一个信息向量。对一个系统，如果任一对象在任一属性下的值都是确定的，则该信息系统被称为是完备的，反之信息系统被称为是不完备的。

每一个属性子集$A \subseteq AT$决定了一个二元的不可区分关系$IND(A)$：

$$IND(A) = \{(x, y) \in U \times U \mid \forall a \in A, f_a(x) = f_a(y)\} \tag{5.5}$$

关系$IND(A)$（$A \subseteq AT$）是一个等价关系，决定了样本空间U的一个划分。

通常，信息系统中的属性并不是同等重要的，某些属性对于分类学习可能是冗余的，因此，需要在保持系统分类能力不变的条件下，删除那些冗余属性，这个过程称为属性约简。

设$S = (U, AT, f)$是一个信息系统$A \subseteq AT$，$a \in A$，如果$IND(A) = IND(A - \{a\})$，则称a为A中不必要的，否则称a为A中必要的。如果每一个$a \in A$都为A中必要的，则称A为独立的，否则称A为依赖的。

设$S = (U, AT, f)$是一个信息系统，$P \subseteq A \subseteq AT$。如果$P$是独立的，且$IND(A) = IND(P)$，则称$P$是$A$的一个约简。$A$中所有必要属性组成的集合称为$A$的核，记作$core(A)$。

（4）区分矩阵与区分函数

在粗糙集理论中，系统的区分能力可以用区分矩阵（Discernibility Matrix）及区分函数（Discernibility Function）[44]的形式来表示。

设$S = (U, AT, f)$是一个信息系统，$A \subseteq AT$，$|U| = n$，A的区分矩阵是一个$n \times n$矩阵，其中任一元素定义为

$$\alpha(x, y) = \{a \in A \mid f(x, a) \neq f(y, a)\} \tag{5.6}$$

因此，$\alpha(x, y)$是A中能区分对象x和y的所有属性的集合。

下面介绍区分函数的概念，用Δ来表示。对每个属性$a \in A$，指定一个布尔变量“a”，若$\alpha(x, y) = \{a_1, a_2, ..., a_k\} \neq \phi$，则指定一个布尔函数$a_1 \vee a_2 \vee ... \vee a_k$，用$\sum \alpha(x, y)$来表示；若$\alpha(x, y) = \phi$，则指定布尔常量1。区分函数$\Delta$定义如下：

$$\Delta = \prod_{(x,y) \in U \times U} \sum \alpha(x, y) \tag{5.7}$$

如果布尔表达式是一个析取范式且包含最小数目的合取式，那么布尔表达式称为一个极小析取范式。这样区分函数具有如下性质：函数Δ的极小析取范式中的所有合取式是属性集A的所有约简。就是说，约简是能区别由整个属性集区别的所有的

对象的属性极小子集。

（5）连续值属性离散化的粗糙集表示

选取分割点本质意义就在于其区分不同类别样本的能力，样本空间连续属性值离散化还可以直观地描述为：离散化的过程就是要在属性空间内寻找一组最优超平面，它能将空间中所有不同类别的对象划分开来。因此，可以利用各属性上的分割点与不同类别样本对之间对应的二元关系来建立布尔函数，进而得到整个样本的区分函数，然后化简区分函数来得到相应的离散化取值。

连续值属性离散化方法的基本思想[45]，设一个具有连续值属性的决策信息系统$S=(U,AT\cup D)$，这里U为有限非空的样本集合，称为论域或对象空间，AT是样本空间的非空属性集合，D为决策属性集合，对于每个连续值属性$a\in AT$，其值域V_a是样本空间U在属性a上的取值范围，由实数域上的一段左闭右开的区间$[v_a,w_a)$来表示。对样本空间U的连续值属性离散化的结果就是要在每个连续值属性a的值域V_a中寻找一个恰当的划分P_a，且在划分P_a下的系统与初始系统具有相同的决策能力，P_a将属性值域划分为若干互不相交的子区间，对每个子区间以符号赋值，即得到一组V_a上的离散化取值。因为任何划分P_a是由一组值域V_a内的分割点序列$(v_1<v_2.....v_k)$确定的，所以，离散化就是要在每个连续值域V_a的划分点序列集合中选出一个恰当的划分点序列，进而形成满足系统需要的划分。如果划分后的决策表S^P的广义决策与初始系统S的广义决策相同，即$\partial_{S^P}=\partial_S$，那么称这些划分为协调的（consistent），且如果减少划分中的任意一个划分点，则划分就会变得不协调，则该划分是既约的（irreducible）。在所有的协调既约划分中，含划分点数最少的划分（即离散化后取值个数最少）被称为是最优划分，一个决策表的最优划分通常不是唯一的,因此，人们设计了各种算法以求得近似解，即寻找次优划分。下面介绍常用的MD离散化算法。

5.1.2 MD-离散化方法

MD-算法[46]是基于Johnson策略的离散化方法，其基本思想如下，每次在所有的候选划分点中选出能够识别出最多具有不同决策值的样本对的划分点，然后在论域中消除这个点以及被这个划分点识别出的样本对，形成新的论域，重复以上步骤，直至所有样本对均被区分，选中的划分点集合形成的划分就是所求的次优划分。

根据给定的决策表A构造一个新的决策表A^*，其中每一行代表了决策表中决策值不同的待区分的样本对，每一列代表一个候选划分点，若依据划分点划分可以辨

别某个样本对，则划分点对应的区间变量值为 1，此外，在其中增加一行 new 以及一个新的列 d^*，并作相应赋值。

MD-算法

Step1 构建区分矩阵 A^*，然后从中删除 new 行，得到初始表系统 B；

Step2 选择 B 中含'1'个数最多的列，即选取该列对应的划分；

Step3 从 B 中删除第 2 步中选中的列，并且删除该列中含有的'1'所对应的所有的行，创建新的系统 B；

Step4 如果 B 不为空，则转至第 2 步，否则停止。

由上述步骤算法执行过程考虑了各属性分割点在离散化过程中区分能力的相关性，因此，算法得到的离散属性值具有较高的精度和较小的数据规模。

5.2 基于粗糙集理论的特征选择方法

该特征选择方法采用上述基于粗糙集理论的 MD 算法，利用离散化方法获得评论文本中具有情感倾向性的特征，借助支持向量机的分类方法进行研究。实验结果表明基于粗糙集的文本情感倾向性特征选择方法是可行的，相比较统计特征选择方法，具有良好的可解释性和低维特征空间适应性。

5.2.1 获取候选特征的类别区分能力

根据我们的实验，发现Fisher判别函数方法在缩减特征空间的同时,能选出那些最具类别指示意义的特征。我们使用如下的Fisher判别函数 $F(\cap f)$ 来计算候选特征的类别区分能力。

$$F(\cap f)\frac{m\times n\times(n\times\sum_{i=1}^{m}\frac{w_{d_{pi}}}{v_{d_{pi}}}-m\times\sum_{j=1}^{n}\frac{w_{d_{Nj}}}{v_{d_{Nj}}})^2}{n^3\times\sum_{i=1}^{m}(m\times\frac{w_{d_{pi}}}{v_{d_{pi}}}-\sum_{i=1}^{m}\frac{w_{d_{pi}}}{v_{d_{pi}}})^2+m^3\times\sum_{j=1}^{n}(n\times\frac{w_{d_{Nj}}}{v_{d_{Nj}}}-\sum_{j=1}^{n}\frac{w_{d_{Nj}}}{v_{d_{Nj}}})^2} \tag{5.8}$$

这里，m 表示正面的文本数，分别为 $d_{p_1},d_{p_2},\cdots,d_{pm}$，每篇文本中出现的总词次分别为 $v_{d\,p_1},v_{d\,p_2},\cdots,v_{d\,pm}$，词汇 f 出现在每篇文本中的总次数分别为 $w_{d\,p_1},w_{d\,p_2},\cdots,w_{d\,pm}$。同理 n 表示反面的文本数，分别为 $d_{N_1},d_{N_2},\cdots,d_{Nn}$，每篇文本中出现的总词次分别为 $v_{d\,N_1},v_{d\,N_2},\cdots,v_{d\,Nn}$，特征项 f 出现在每篇文本中的总次数分别为 $w_{d\,N_1},w_{d\,N_2},\cdots,w_{d\,Nn}$。$\cap f$ 表示词汇正反面文本的词汇的交集。

5.2.2 建立决策表

通常，非完备决策表中的空值有两种情形：一种是“丢失（missing）”，即原本这个属性值存在的，只是当前没有，它可以是对应属性值域中的任何一个，一般用“*”表示；另一种是“缺失（absent）”，即这个属性值原本就不存在。在文本以向量形式的表示中，有些候选特征或特征在某些文本中根本不存在。由于我们每个候选特征或特征在向量中以权重表示，因此我们用 0 来表示。如果两个文本在某个候选特征或特征下均取 0 值，则这两个文本在此候选特征或特征下显然是不可区分的，因为它表明这两个文本对应的候选特征或特征是不存在。

由此可见，带有情感倾向强度的决策表是一个非完备决策表，可将其完备化变成了一个完备化的决策表。

5.2.3 特征选择

特征选择的过程是基于 MD 算法，前面已经对算法有介绍，下面给出算法的 Delphi 语言描述。

算法：训练器实现

输入：区分矩阵 DT

输出：生成新的规则

```
Procedure Main.trainning(Sender: TObject);
var
    fetr:= array of String ;//表示候选特征集合
    DTT,cut: array of array of Real; //表示节点 cut 集
begin
    try
        Init(fetr);
        Init(DTT);
        Init(cut);
   finally
         FreeAndNil();
   end.                          //进行初始化
Main. Sort(fetr);//对候选特征排序
for j:=0 to high(DT.Line) do
    begin
```

```
        for i:=0 to high(DT.Row) do
            begin
                Main. sort(DT);//对 DTT 中的每个属性排序
            end;
        for i:=0 to high(DT.row) do
            begin
                cut[i1:=0][j]:=Main. Middle(DT[i][j]); //计算节点值
                Inc(i1);
            end;
    end.
Main. Change(DTT,DT,cut);//进行 DT 变换
R:=Main. Reduct(DTT);//对 DTT 进行约简
Main. Wrt(R,'Rule-DB'); //将特征结果写入候选特征库
    Try
        Main. Init(DT);//重新初始化 DT
    finally
        FreeAndNil();
   End.
Main. Map(DT,DTT, cut);//构造节点下的离散化 DT
    Try
        Main. Wrt(DT,'Rule-DB'); //d 得到分类规则结果
    finally
        FreeAndNil();
    end;
        Showmessage('完成');
end.
```

算法：Change()的具体实现

输入：DTmp，cutmp

输出：DTTmp

```
Procedure Main.Change (DTTmp: array of array of Real ; DTmp: array of array of
                        Real;cutmp: array of array of Real);
```

```
var
    T,DTTmpL:= array of array of Integer;//表示候选特征集合
    DTT,cut: =array of array of Real ; //表示节点 cut 集
begin
    Try
        Main. Init(T);// 初始化计数器
        Main. Init(DTTmpL);//初始化 DTTmp 行
        Main. Init(DTT);
        Main. Init(cut);
    finally
        FreeAndNil();
    end;
for j:=0 to high(cutmp.Line) do
    for i:=0 to high(cutmp.Row) do
      begin
        if (cutmp[i][j]=true) then
            begin
                T[0][j1]:=j; //记录每个属性产生的列的个数
                T[1][j1]:=cutmp[i][j];//线性化 cutmp
                Inc(j1);          //计算 DTTmp 行
            end;
      end;
    DTTmpL :=Main. Rank(i,2);//计算 DTTmp 行文本和对应决策
    for j:=0 to Dec(j1) do
      for i:=1 to   high(DTTmpL) do
          begin
              if (DTTmpL[i][2]<> DTTmpL[i][3])and
                (DTTmpL[i][0]> T[1][j])and (DTTmpL[i][1] >T[1][j]) then
                  DTTmp[i][j]:=0
              else if (DTTmpL[i][2]<> DTTmpL[i][3])and
                      (DTTmpL[i][0]<T[1][j])and (DTTmpL[i][1] <T[1][j]) then
```

```
                DTTmp[i][j]:=0
            else if (DTTmpL[i][2]= DTTmpL[i][3])and
                (DTTmpL[i][0]>T[1][j])and (DTTmpL[i][1] < T[1][j]) then
                DTTmp[i][j]:=1
            else if (DTTmpL[i][2]= DTTmpL[i][2])and
                (DTTmpL[i][0]< T[1][j])and (DTTmpL[i][1] > T[1][j]) then
                DTTmp[i][j]:=1
            else
                DTTmp[i][j]:=2;
        end;
end.
```

算法：Map()的具体实现

输入：DTmp ,DTTmp，cutmp

输出：DTmp

```
Procedure Main.Map (DTmp: array of array of Real; DTTmp: array of array of Real;
                    cutmp: array of array of Real);
begin
  for j:=0 to high(DTmp.Line) do
    for i:=0 to high(DTmp.Row) do
      if  DTmp[i][j]=0 then
        DTmp[i][j]:=-1   //表示第 j 列对应的特征在第 i 篇文档中没有出
                         //现，标记为-1
      else
      while t< high(cutmp.Line) do
        Begin
          Inc(t);
          If  Cutmp[t][0]<>i then
            Break    //如果第 i 个特征可能的 cut 值已经全部遍历
                     //完，则结束搜索，跳出进行下一篇文本的搜
                     //索计算
          else If (DTmp[i][j]< DTTmp[0][t]) and (Cutmp[t-1][0]<>i) then
```

```
                DTmp[i][j]:=0        //特征小于其对应的最小的 cut，
                                     //则赋值为 0
            else If (DTmp[i][j]> DTTmp[0][t]) and
                    (DTmp[i][j]< DTTmp[0][t+1]) then
                DTmp[i][j]:=t   //特征大于其对应的第 t 个 cut 值，
                                //但小于第 t+1 个，则赋值为 t
            else
                DTmp[i][j]:=t+1; //如果是最大的，则赋值为 t+1
        end;
    end.
```

5.3 文本情感倾向性分类的步骤

实验选用的评论文本主要来自 corpus 2 的 500 篇进行实验。详细步骤如下：

Step1 将语料库文本进行分词、词性标注预处理；

Step2 将预处理文本集分为训练集和测试集，比例约为 4:1；

Step3 计算候选特征的类别区分能力；

Step4 采用粗糙集理论算法进行特征选择；

Step5 利用支持向量机（SVM）分类器进行训练；

Step6 利用训练过程得到的文本情感倾向分类器，对测试集文本的情感倾向类别进行判断。

5.4 实验结果与分析

5.4.1 评价指标

根据上节对评论文本特征选择的步骤，得到的主要结果是对原数据利用属性离散化方法实现了数据的离散化，同时去除了冗余数据，对属性进行了一定程度的压缩获取特征。因此，我们度量整个情感分类的效果时主要从以下几个评价指标进行评价：

（1）离散化方法的数据压缩能力（离散化后的数据集规模与原始数据集规模之比）和约简协调率。

（2）最后系统分类性能的量化指标，通常采用正确率，召回率（Recall，简记为 r）、精确率（Precision，简记为 p），F 值。根据其分类类别分为正面的、反面的、综合的。

正确率定义为系统分类正确的正面和反面的文本数与系统自动分类的文本数之比。

精确率定义为系统对该类分类正确的文本数与系统自动分类该类文本数之比。

召回率定义为系统对该类分类正确的文本数与实际文本中存在的该类文本数之比。

F 值均衡精确率和召回率，为 2×（精确率×召回率）/（精确率+召回率）。

5.4.2 实验结果与分析

本章的实验语料来源于第二章中的 corpus 2 中的 500 篇评论文本，分为训练集和测试集，比例为 4:1，其中正、反面文档比例均为 1:1。设定阈值选取前 500 个词汇作为候选特征并采用离散化算法进行特征选择，进一步使用支持向量机分类器进行分类和测试。

（1）数据的压缩能力和约简协调率

通过文本表示以及文本的候选特征选择，400 篇训练集文本得到的候选特征为 500，这样得到的文本表示矩阵为 400×500，为了展示 400 篇文本的在未进行数据离散化前的数据形式，我们列出部分数据如下：确实、口水、独立、轻质、人性化、中规中矩、传统、出色、现代、协调、适用、附近、根本、差、全、个性化、肯定、长、客观、专业、匀速、首选、险峻、充裕、充沛、匀称、隽永、顽固、顽强、经意、出众、绝佳、真挚、破烂、吓人、杂、下贱、配套、具体、真正、在理、绝对、鲜活、闲、完善、鲜有、完整、鲜艳、鲜明。

利用离散化方法后的数据，由原来的 400×500 矩阵压缩成 211×127 的矩阵，数据的压缩比例为 13.39 %，具有较高的数据压缩能力。

实验中文本只有 1 篇不协调，协调率为：99.75%。其离散化的形式为：

表 5-1 离散化表示

U	精确	出色	绝佳	隐患	…	Category
Doc_1	-1	2	-1	-1	…	P
Doc_2	2	-1	1	-1	…	P
…	…	…	…	…	…	…
Doc_n	-1	1	-1	2	…	N

（2）分类性能

由于所选的语料均未经过精选，更接近真实的情况，所以我们任意选择了三名

本科生进行人工判别（Manual）将结果作为实验对比。我们将压缩前的500个特征（SVM-1）压缩后的127个特征（SVM-2）和第三章效果最好停用词表5分别在相同的维数下（500维特征记为SVM-3、127维特征记为SVM-4）选出的特征利用支持向量机分类器分类，并将结果进行对比，结果见表5-2。

由表 5-2 和图 5-1 可以看出：

（A）从人工判断的结果来看，*F* 值尚不到 90%，说明即使对人而言情感倾向性判断也比较困难，造成的原因主要来自两个方面：一方面是不同的人对情感的理解和情感倾向的认识不同，有些文档正，反面判定难以把握；另一方面来自语料的实际情况，产品评论格式比较自由，情感倾向较模糊。

表 5-2 实验结果对比

方法		Manual	SVM-1	SVM-2	SVM-3	SVM-4
正面	精确率	93	64	75	78	63.07
	召回率	84.55	74.42	66	67.24	82
	F 值	88.57	68.82	70.21	72.22	71.30
反面	精确率	82	78	69.64	62	52
	召回率	91.11	68.42	78	73.81	74.29
	F 值	86.32	72.90	73.58	67.39	61.18
综合	*F* 值	87.5	71	72	70	67

（B）可以看到与第三章测试结果变化趋势有所不同，由于正面反面文档比例相同，反面的查全率和查准率与正面查全率、查准率在整体变化上没有明显的差距，这里再次验证了我们前面的测试。本章实验总体评价性能低于第三章结果，造成这样的原因主要是特征维数空间非常低，反面文档难以判断。

（C）随着特征空间的不断降低，对比第三章中相关特征选择方法结果出现了降低，这表明传统特征选择方法在维数空间骤降的情况下，所选的特征已经不能满足分类所需的信息，只有在足够的特征空间下才能获得较好的结果。同时基于粗糙集的特征选择方法，在低维空间表现出良好的适应性，具有较好的分类效果，并且随着特征空间的进一步缩小，分类的效果得到了进一步的提升，而传统的特征选择方法则相反，随着特征空间的缩小效果越来越差。在低维空间下，基于粗糙集理论的特征选择方法分类效果整体优于停用词表5的情况，在500维时方法间的差距不大，但是随着特征空间的进一步压缩，特征选择方法的差距变得较大。

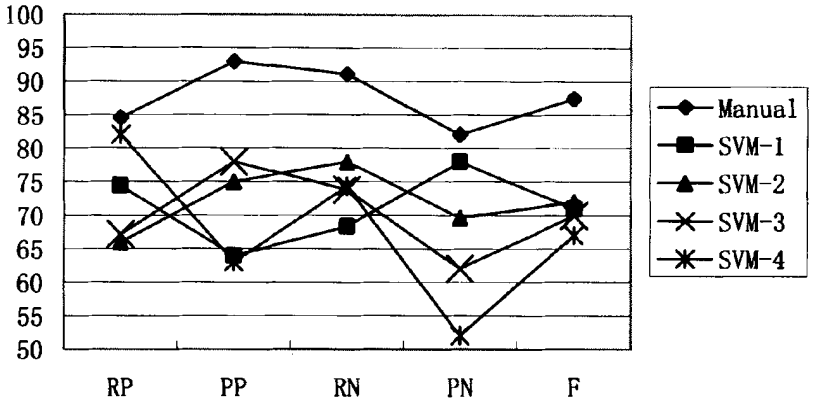

图5-1实验结果对比

（D）SVM-4 中，各项指标波动较大，表明传统的特征选择方法在低维空间，性能不够稳定，结合 SVM-3 可以看出，随着空间的压缩，分类性能会下降较快。而基于粗糙集的特征选择方法，适合于低维空间的使用，随着维数的降低，性能有所提高，但是分类效果要低于上述方法在高维空间的结果。

（E）使用停用词表 5 的方法获取的 127 个部分特征如下: 超车、显示屏、宽敞、当然、装备、符合、欧洲、中央、配合、有些、后座、A6、天窗、外形、竟然、油耗、虽然、千瓦、电动、风格、空间、排放、天籁、适应、设计、通过、看见、动作、时间、变速器、富康。从这些特征来看可解释型很差，与情感分类的关系很难理解。基于粗糙集方法选择的部分特征如下：欺骗、精确、最佳、隐患、潮流、气愤、圆润、脱落、害、智能、负责、羡慕、好听、好处、充足、报废、疲劳、吃亏、熟悉、合适、机会、辛苦、迟钝、悲哀、心疼、鄙视、不爽、不平、清楚。这些特征符合我们对情感倾向判断的认识，同时也能较好的反映评论的情感倾向。

总的来看，情感倾向性分类问题较传统的主题分类问题，更加复杂和困难。本章所采用的特征选择方法，在低维空间具有较好的适应性，更适用于大规模的分类应用，但是性能需要进一步的研究提高。

5.5 本章小结

针对统计机器学习方法在情感倾向性分类问题中存在的问题，本章将粗糙集理论与文本情感倾向性分类问题相结合，实现了基于粗糙集的特征选择方法，将其用于文本的倾向性判别。实验结果表明该分类方法是可行的，具有良好的可解释性和低维数空间的适应性，但该理论方法存在复杂度较高，部分计算过程存在指数阶时

间和空间复杂度。

尽管实验中存在不理想，如运算执行中存在复杂度较高的问题，但也看到其在理论上的优越性和对解决情感倾向性分类问题的有效性。

第六章 结论与展望

本章概括了全文所作的研究工作和研究成果，同时指出了研究中尚未解决的问题以及今后需要进一步研究的方向。

6.1 结论

近年来随着信息技术的迅猛发展，互联网迎来了前所未有的新局面。传统的基于主题的信息处理，已经远远不能满足人们的需求，用户希望得到更多的主观性的信息，正是在这样的情况下以网络为传播媒介的带主观信息的文本越来越受到人们的关注。对于这些仅靠人工发现和分析显然是行不通的，人们开始关注并研究评论文本的情感倾向性分析。

本文主要针对网上的汽车产品领域的评论文本进行情感倾向性分类研究工作。本文的主要研究工作有以下几个方面：

（1）建立了用于情感倾向性分类研究的资源

本文建立了两个用于情感分类的汽车产品评论语料库：精选的corpus 1和未精选的corpus 2。在深入分析了评论性语料的特点后，认为该语料能真实反映出情感倾向性分类问题中特有现象和突出问题，建立了受限领域汽车产品知识库和汽车产品评论情感词汇库，这些为顺利开展问题的研究提供了基本保障。

（2）研究停用词表对中文文本情感分类的影响

本文选用了五种停用词表，采用常用的三种特征提取方法[5]，信息增益（IG）、互信息（MI）和χ^2假设检验（CHI），两种权重的计算方法，基于文档和基于词频，利用支持向量机分类方法，分别考察了五种停用词表对汽车评论的情感类别判断的影响。实验结果表明基于传统的主题分类的关键技术可以应用于文本情感倾向分类，但分类效果不及文本主题分类。信息增益（IG）和布尔型对情感分类的效果整体比较好，这与英文文本的情感分类报道结果相一致，另外实验也反映出所选的特征与主题相关，不具有情感倾向性。这些都说明了中文文本情感分类要比主题分类更加复杂。

（3）提出了混合特征选择的方法

对于情感倾向性分类问题，本文提出了基于类别区分能力的混合特征选择方法。该方法按照词汇的类别区分能力和词汇所包含的情感倾向信息进行特征选择，讨论了在不同的特征选择方法和不同维数特征空间下对文本情感分类结果的影响。实验

表明混合特征选择方法的性能要优于信息增益（IG）特征选择方法，情感倾向性分类的结果随着特征区分能力的增强而变得更好，也表明混合特征选择方法在中文文本情感倾向性分类问题中是一种有效的方法。

（4）设计并实现了基于粗糙集理论的特征选择方法

针对统计机器学习的方法在解决情感分类问题中存在的问题和不足之处，本文提出了基于粗糙集理论的特征选择方法，并将此方法用于文本情感分类。通过情感特征选择、维数压缩等方面的实验测试，结果表明该方法用于文本情感分类特征选择是可行的，在低维空间下基于粗糙集理论的特征选择方法分类效果整体优于基于统计机器学习的特征选择方法，用于分类的特征具有良好的类别可解释性和低维数空间的适应性，因而更适合于大规模语料文本情感分类的特征选择。但实验结果同时也反映出粗糙集理论及其方法在实际应用中存在一定的问题，如：算法的复杂度较高，部分子计算过程存在指数阶时间和空间复杂度，对大规模计算存在着时间和空间上的不适应。

6.2 展望

通过对文本情感倾向分类的研究，取得了阶段性的研究成果，但得到的结果与文本主题分类结果相比，还是不太理想。就本文而言需要下一步深入研究的问题有以下几个方面：

（1）资源建设方面，带有情感倾向的词汇库资源除利用现有的静态资源外，还应加入动态的资源，如：情感倾向性受上下文环境影响的情感词汇，同时应该考虑对上述情感资源词汇的倾向性进行强度量化标记。

（2）情感词汇的特征选择方面，本文虽然取得了不错的研究成果，但是仅仅依靠情感词汇作为特征进行文本情感分类还是不够的，它不能解决短语所蕴含的情感信息，比如带否定词的短语、程度副词的短语等，因此，基于多粒度的情感信息获取方面应进一步深入研究。

（3）由于问题的复杂性，在基于粗糙集理论的特征选择方法的实现过程中，在进行类间比较时，时间复杂度比较高，应进一步研究高效的离散化算法，以便提高整个系统的性能。

总之，对中文文本的情感倾向分类的研究，国内目前都处于起步阶段，不论是相关理论还是处理方法可以借鉴和参考的都很少，有许多问题值得进一步深入研究。这就要求我们重新审视新问题的特殊性和复杂性，摆脱传统的主题分类理论和方法

在新问题中表现出的不适应，以新的思维去看待和研究这个新问题。

参考文献

[1] 文本分类若干关键技术研究.李荣陆.复旦大学.博士论文.2005.6.

[2] 李荣陆,胡运发. 基于密度的KNN文本分类器训练样本裁剪方法.计算机研究与发展. 2004.41(4):539-545.

[3] Y.Yang. An evaluation of statistical approaches to text categorization. Information Retrieval, 1999, 1(1): 76~88.

[4] Dejun Xue, Maosong Sun. A study on feature weighting in Chinese text categorization. Computational Linguistics and Intelligent Text Processing (CICLing-03), LNCS2588, Springer-Verlag, 2003: 592-601.

[5] N. Chambers, J.tetreault, and J. Allen. Approaches for automatically tagging affect. In Proceeding of the AAAI Spring Symposium on Exploring Attitude and Affect in Text: Theories and applications. 2004. Stanford ,US.

[6] 邹嘉彦.评述新闻报道或文章色彩-正负两极性自动分类的研究.自然语言理解与大规模内容计算-全国第八届计算语言学联合学术会议.清华大学出版社. 2005.21-23.

[7] Mingqing Hu and Bing Liu. Mining opinion features in customer reviews. In AAAI, 755-760. 2004.

[8] S.morinaga, K.Yamanishi,K.Tateishi and T.Fukushima. Mining product reputations on the Web. In Proceedings of KDD-02,8th ACM International Conference on Knowledge Discovery and Data Mining . ACM Press, 2002,341-349,Edmonton, CA.

[9] M.Hurst and K.nigam. Retrieving topical sentiments from online document collections. In Proceedings of SPIE, Document Recognition and Retrieval XI(2004), no. 5296 In Proceedings of SPIE, 27-34.

[10]Qiang Ye, Wen Shi, Yijun Li. Sentiment classification for movie reviews in Chinese by improved semantic oriented approach. In Proceedings of the 39th Hawaii International Conference on System Sciences, 2006.

[11]时达明，林鸿飞. 基于内容相关度和情感分析的Bolgger声誉度研究. 第三届全国信息检索与内容安全学术会议.2007.11.656－662.

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

[13] DAS, S. R., AND CHEN, M. Y. Yahoo! for Amazon: Sentiment parsing from small talk on the Web. In Proceedings of EFA 2001, European Finance Association Annual Conference (Barcelona, ES, 2001).

[14] B. Pang, L. Lee, and SVaithyanathan. Thumbs up? Sentiment classification using machine learning techniques. In Proceedings of the Conference on Empirical Methods in Natural Language Processing, 2002, 79-86.

[15]Jian Liu, Jianxin Yao, and Gengfeng Wu. Sentiment classification using information extraction technique. In International Symposium on Intelligent Data Analysis. Madrid. 2005, 3646:216-227.

[16]Qiang Ye, Wen shi, Yijun li. Sentiment classification for reviews: Comparison between SVM and semantic approaches. The fourth international conference on machine and cybernetics. Guangzhou, 2005:2341-2346.

[17]Jin Cheon Na, Christopher Khoo, Paul Horng Jyh Wu. Use of negation phrases in automatic sentiment classification of product reviews. Library collections, acquisitions &Technical services 2005(29):180-191.

[18]B. Pang, L. Lee. A sentimental education: Sentiment analysis using subjectivity summarization based on minimum cuts. In Proceedings of ACL-04, 42nd Meeting of the Association for Computational Linguistics, 2004:271–278, Barcelona, ES.

[19]L.Lee. A matter of opinion: Sentiment analysis and business intelligence (position paper).Written for and presented at the IBM Faculty Summit on the Architecture of On-Demand Business, May 2004.

[20]B.Liu,M.Hu,and J.Cheng. Opinion observer: analyzing and comparing opinions on the Web. In Proceeding. of WWW'05, the 14th International Conference on World Wide Web, Chiba, Japan.2005: 342-351.

[21]Jin Cheon Na, Christopher Khoo, Paul Horng Jyh Wu. Use of negation phrases in automatic sentiment classification of product reviews. Library collections, acquisitions &Technical services 2005,29, 180-191.

[22]Peter D. Turney and Michael L. Littman. Measuring praise and criticism: inference of

semantic orientation from association. ACM Transactions on Information Systems, Oct.2003,21(4): 315-346.

[23]J. Yi, T. Nasukawa, R. Bunescu, W. Niblack. Sentiment analyzer: Extracting sentiments about a given topic using natural language processing techniques. In Proceedings of the Third IEEE International Conference on Data Mining, November 19-22, 2003.

[24] B. Pang ,L.Lee ,and S.Vaityyanathan. Seeing stars: Exploiting class relationships for sentiment categorization with respect to rating scales. In Proceedings of ACL-05,43nd Meeting of the Association for Computational Linguistics 2005 ,Ann Arbor, US.

[25]A. Finn and N. Kushmerick. Learning to classify documents according to genre. Journal of the American Society for Information Science and Technology(JASIST),Special issue on computational analysis of style, volume 7,nember 5,March 2006.

[26]Peter D. Turney and Michael L. Littman. Unsupervised learning of semantic orientation from a hundred-billion-word corpus. Tech. Rep. EGB-1094, National research Council Canada,2002.

[27]邹嘉彦.评述新闻报道或文章色彩-正负两极性自动分类的研究.自然语言理解与大规模内容计算-全国第八届计算语言学联合学术会议.清华大学出版社.2005.21-23.

[28]Songbo Tan and Jin Zhang. An empirical study of sentiment analysis for Chinese documents. Expert Systems with Applications, 2007.

[29]Wang Suge,Wei Yingjie,Zhang Wu, Li Deyu, Li Wei. A hybrid method of feature selection for Chinese text sentiment classification. The 4th International Conference on Fuzzy Systems and Knowledge Discovery.24-27 August 2007, Haikou, China.

[30]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.2006.21(1): 14-20.

[31]柴玉梅，熊德兰，昝红英. Web文本的褒贬倾向性分类研究.计算机工程2006.(17):89-91.

[32]http://www.xche.com.cn/baocar/

[33]http://www.wjh.harvard.edu/~inquirer/

[34]张伟、刘缙等.学生褒贬义词典.中国大百科全书出版社.2004.

[35]HowNet [R]. HowNet's Home Page. http://www. keenage. com.

[36]顾益军,樊孝忠,王建华,汪涛,黄维金.中文停用词表的自动选取。北京理工大学学报. 2005.25(4):337~340.

[37]Hart G W. To decode short cryptograms[A] Communications of the ACM[C]. New York Association for Computing Machinery,1994:102~108.

[38]Yang Y. Pedersen J O. A comparative study on feature selection in text categorization[A] Proceedings of ICML-97,14th International Conference on Machine Learning[C].San Francisco Morgan Kaufmann Publishers Inc.1997:412~420.

[39]Silva C, RibeiroB. The importance of stop word removal on recall values in text categorization [J].Neural Networks,2003,3:20~24.

[40]M. Taboada, C. Anthony and K.Voll. Methods for creating semantic orientation dictionaries. In Proceedings of Fifth International Conference on Language Resources and Evaluation. Genoa, Italy.

[41]V. Hatzivassilolou Kathleen, R. Mckeown. Predicting the semantic orientation of adjectives. In Proceeding of the 35th Annual Meeting of the Association for Computational Linguistics and the 8th Conference of the European Chapter of the ACL. Association for Computational Linguistics, New Brunswick, 1997:174~181.

[42]王治敏，朱学峰，俞士汶 ．基于现代汉语语法信息词典的词语情感评价研究. Computational Linguistics and Chinese Language Processing. 2005:10(4):581~592.

[43]代六玲,黄河燕 ,陈肇雄.中文文本分类中特征抽取方法的比较研究. 中文信息学报.2004:18(11):26~32.

[44]张文修,吴伟志,梁吉业,李德玉.粗糙集理论与方法[M].第一版.北京,科学出版社,2000,3-39.

[45]Jan Komorowski, Lech Polkowski, Andrzej Skowron. Rough Sets:A Tutorial [EB/OL]. http://folli.loria.fr/cds/1999/library/pdf/skowron.pdf.

[46]S.H.Nguyen, H.S.Nguyen. Some efficient algorithms for rough set methods[C] // In

Proceedings of the Conference of Information Processing and Management of Uncertainty in Knowledge-Based Systems, Granada, Spain, 1996, 1451-1456.

[47]薛德军.中文文本关键技术研究.清华大学.博士论文.2004.6.

附录 I 信息处理用现代汉语词类标记集

n	名　词	q	量　词
n	普通名词	d	副　词
nt	时间名词	r	代　词
nd	方位名词	c	连　词
nl	处所名词	p	介　词
nh	人　名	u	助　词
ns	地　名	e	叹　词
ni	机构专名	o	拟声词
v	动　词	i	习用语
v	普通动词	j	缩略语
vl	判断动词	k	后接成分
vu	能愿动词	x	非语素字
vd	趋向动词	h	前接成分
a	形容词	wp	标点符号
m	数　词		

附录 II GI 词典示例

abandon	Negativ	放弃
abandon	Negativ	遗弃
abandon	Negativ	中止
abandonment	Negativ	放弃
abandonment	Negativ	遗弃
abandonment	Negativ	中止
abate	Negativ	减弱
abate	Negativ	降低
abatement	Negativ	减弱
abatement	Negativ	降低
abdicate	Negativ	让位
abdicate	Negativ	逊位
abhor	Negativ	憎恨
abhor	Negativ	痛恨
abhor	Negativ	厌恶
abide	Positiv	逗留
abide	Positiv	居留
abide	Positiv	讨厌
ability	Positiv	能力
ability	Positiv	才能
abject	Negativ	凄惨的
abject	Negativ	绝望的
abject	Negativ	下贱的

附录 III HOWNET 情感分析用词语集示例

中文正面情感词语	中文负面情感词语
嗆	傒倖
媚	恤
媚嫉	恟
忺	懆懆
爱	捶胸顿足
爱不忍释	哀
爱不释手	哀哀切切
爱宠	哀愁
爱戴	哀怜
爱抚	哀悯
爱好	哀戚
爱护	哀凄
爱怜	哀切
爱恋	哀伤
爱慕	哀痛
爱上	哀痛欲绝
爱屋及乌	哀怨
爱惜	哀恸
安	哀矜
昂扬	傲视
巴	懊
巴不得	懊恨
巴望	懊悔
摆好	懊恼
拜	懊丧
拜拜	百无聊赖
拜服	败兴
拜贺	板
拜谒	板脸
褒	板面孔
褒奖	板起脸
褒扬	板着脸
睎	板着面孔
表扬	半信半疑

发表文章及参加项目

攻读硕士期间完成的论文：

[1] Wang Suge, Wei Yingjie, Zhang Wu, Li Deyu, Li Wei. A hybrid method of feature selection for Chinese text sentiment classification. The 4th International Conference on Fuzzy Systems and Knowledge Discovery. Vol 3, Proceedings, Lei JS, Yu J, Zhou SG, Editors. 2007. IEEE Coputer Soc:Los Alamitor. 435-439 （EI、ISTP:000252460600087）.

[2] 王素格,魏英杰.停用词表对中文文本情感分类的影响.情报学报,2008,27（2）：176-180.

攻读硕士期间参加的项目：

[1] 基于 Web 的文本情感分类关键技术研究（2007011042）. 山西省自然科学基金，2007.4-2009.6.

[2] 文本流派分类及其在信息安全中的应用研究（200611002）. 山西高校科技研究开发项目.2006.2-2008.7.

致 谢

毕业在即，首先应该感谢父母家人对我学业一如既往的理解支持，没有他们的无私奉献和亲情关爱，就没有今天的一切。

论文的研究工作是在我的导师王素格教授的悉心教导下完成的。在硕士研究生的三年时间里，您不断以严谨求实的工作作风，广博坚实的学术修养以及敏锐开阔的科研视野引导着我在专业方向上稳步前行，其中一点一滴的成长，都浸透着您的辛劳，您在生活中的关照以及在学业方面的热忱帮助也将使我铭记于心。借此送上学生最真诚的祝福，向恩师表示诚挚的敬意和深深的谢意！

感谢学院贾新春教授，李晓明教授，郭春梅院长和李世惠老师以及学院的各位领导老师这三年来在学业和生活上给予我的支持和帮助。

感谢李伟，三年来我们一同学习共同进步，感谢给予我支持和帮助的同学和朋友们。

时光飞逝，在山西大学求学的七年间，一个懵懂的少年逐步成熟，成长的道路上，从曾经的迷惘失落到现在对未来的奋发憧憬，亲爱的母校，各位老师和同学们给予我很多，而我却无以为报，唯有默默感激。

最后，再次感谢所有给予我帮助和支持的良师益友，请接受我最诚挚的祝福！

承 诺 书

本人郑重声明：所呈交的学位论文，是在导师指导下独立完成的，学位论文的知识产权属于山西大学。如果今后以其他单位名义发表与在读期间学位论文相关的内容，将承担法律责任。除文中已经注明引用的文献资料外，本学位论文不包括任何其他个人或集体已经发表或撰写过的成果。

学位论文作者（签章）：

2008 年　月　日