# 硕 士 学 位 论 文

# 面向中文评论文本的情感倾向性研究

**Research on Opinion Mining for Chinese Review Text**


作 者 姓 名：吕韶华

学科、 专业：计算机应用技术

学　　号：20809348

指 导 教 师：林鸿飞教授


完 成 日 期：2010 年 11 月


大连理工大学

Dalian University of Technology

# 大连理工大学学位论文独创性声明

作者郑重声明：所呈交的学位论文，是本人在导师的指导下进行研究工作所取得的成果。尽我所知，除文中已经注明引用内容和致谢的地方外，本论文不包含其他个人或集体已经发表的研究成果，也不包含其他已申请学位或其他用途使用过的成果。与我一同工作的同志对本研究所做的贡献均已在论文中做了明确的说明并表示了谢意。

若有不实之处，本人愿意承担相关法律责任。

学位论文题目：面向中文评论文本的情感倾向性研究

作 者 签 名 ：_____ 日期：_____年____月____日

# 摘　　要

伴随着 Web2.0 的快速发展，互联网上每时每刻都在产生数量庞大的用户生成内容（UGC），如餐馆评论，博客评论等。这些主观文本具有很大的潜在商业价值，因此，引起了工业界和学术界的广泛重视。情感倾向性分析研究就是分析主观文本的论断，得到其支持或反对的态度，从而为决策者服务。如果仅仅依靠人们逐一浏览文本内容，然后加以总结得到自己的判断，那么将是一件十分耗时费力的事情。为此，本文就中文评论文本的倾向性判定方面的工作展开研究，提出了一种处理跨领域情感倾向性分析的问题的算法，该算法解决如何利用已标注倾向性的评论文本来判断不同领域的未知倾向性的文本的倾向性的问题。同时，本文对情感倾向性判断的应用也进行了研究，提出了一种基于用户评论的餐馆排序算法，该算法对用户发表的评论进行若干主题抽取并计算各个主题的得分，然后使用逻辑回归模型对其进行训练得到最后的餐馆排序。

大量研究关注在特定领域的情感倾向性问题上，研究结果表明使用有监督的方法可以提高情感倾向性判断的准确性，但是有监督的方法需要提前标注好大量训练语料，此项工作耗费大量时间和金钱，特别是面对不同领域的评论文本需要重新标注，如果直接应用上述有监督的方法，结果也不理想，因为不同领域的评论文本的词分布并不相同。本文提出一种基于 SimRank 算法来解决跨领域的评论文本倾向性判断问题，该算法充分考虑了源领域和目标领域的关系，把源领域和目标领域多次共现的词作为连接这两个领域的桥梁，利用情感词典和 SimRank 算法找到源领域和目标领域潜在情感空间，把情感空间中的词语作为情感分类特征，然后使用支持向量机模型对已标注的源领域进行训练进而得到情感分类模型，最后利用此模型预测目标领域的情感倾向性。实验结果表明该方法取得较好的实验效果。

餐馆评论真实地反映了用户对餐馆的评价和态度，挖掘其中隐藏的情感倾向性将给用户带来巨大的方便。但是当前的评论系统只是给出一个整体的分数，未对餐馆的各个方面进行排序，难以满足不同用户的需求，并且用户评论内容可能与评论等级不一致，可能导致对结论的误导，为此本文提出一种依据评论内容对餐馆进行排序的算法，该算法同时考虑用户评论文本和评论等级，利用 LDA 模型对评论文本进行服务、环境、价格、口味等方面进行主题抽取并计算在这些方面的得分，最后利用逻辑回归进行训练得到餐馆评论排序模型，使用此模型可以依据餐馆评论对餐馆进行排序。

**关键词：**情感倾向性；SimRank；LDA;跨领域

# The Research on Opinion Mining for Chinese Review Text

## Abstract

With the rapid development of Web2.0, huge number of UGC, such as hotel reviews, blog reviews, microblogs and so on, are generated all the time on the Internet. Since these reviews are great of importance to people, it draws the attentions of industry and academia. Opinion mining aims to obtain the attitude of the reviews by analyzing the subjectivity of them, which can help the administrators to make a decision. It is very tedious and time-consuming that just wading through the whole reviews to get the conclusion. In order to solve the difficult problem, this paper proposes a cross-domain opinion mining algorithm, which can be used to transfer the polarity of the labeled training corpus to the unlabeled test corpus. At the same time, this paper also does the research on the applications of opinion mining and introduces a LDA-based rank algorithm for hotels using those reviews. This algorithm extracts the aspects of the reviews and respectively computes those scores， then trains them with logistic regression to get the ranks.

Lots of attentions have been paid on the issue of opinion mining in special domain and the results show that supervised method can improve the precision. However, this method needs a lot of labeled training corpus in advance, which costs much time and money. Besides, when directly applying it to the heterogeneous domain text, the performance is really poor because of the different distributions of words in those domains. With the purpose of meeting the challenge, this paper gives a SimRank-based algorithm to crack this problem. It makes use of the relation between source and target domains, which connects them by the common words between them, then builds the latent emotional space with the help of SentiDic and SimRank, at last it trains the labeled source domain with SVM to get emotional classification model, which can be used to predict the opinion of target domain. The experimental results show the effectiveness of it.

The hotel reviews are the mirror of people's attitude on them, so it is convenient for people to mining the deep opinion in it. But the current review system just shows an overall score lacking of ranks on every aspect, which hardly meets different people's needs. On the other hand, the review text may not be consistent with the overall score given by customers, which is able to mislead people's decisions. As a result, this paper introduces a rank algorithm for hotels based on the reviews, which combines the review text and overall score to extract and aspects of service, environment, price and taste, then compute the scores for all of them,

at last it trains them with logistic regression to get the rank model for hotels which can be used to get ranks based on hotel reviews.

**Key Words**：Opining Mining; SimRank; LDA; Cross-domain

# 目　　录

# 1. 绪论

## 1.1 研究背景

随着 Web2.0 的飞速发展，互联网给人们生活带来巨大便利，其已经成为人们生活中不可或缺的一部分，与此同时，互联网信息以指数级的速度膨胀，传统的导航式信息检索已经不能满足人们获得信息的需求，搜索引擎技术应运而生。搜索引擎采用多种自然语言处理方法对用户提交的查询请求进行处理，最后返回一个排序后的结果供用户选择，极大地提高了人们获取信息的速度。但是传统的搜索引擎技术处理的是客观性的文本，极少考虑文本的主观性问题，所以面对现今网络上不断涌现的评论等主观性文本，需要在传统的搜索引擎技术的基础上进行主观文本的研究。

当前互联网上出现的海量的主观性评论内容，如餐馆评论，电子评论，博客（Blog），微博（MicroBlog）等具有巨大的潜在价值，一方面用户可以参考这些评论内容，提前了解自己关心的对象，以做出更符合自身利益的决定，另一方面，被评论对象也可以利用这些文本所反馈的信息，适时做出应对措施，提高自身的服务质量。但是面对规模如此庞大的评论文本，如果让用户逐一浏览然后加以总结再做出自己的判断，那么将是一件十分耗时费力的事情，并且在当今这个惜时如金，讲求效率时代也是不现实的。所以，围绕这些海量的主观性文本的研究工作便成为现今研究领域的一个热点问题。国内外顶级学术会议都设有专门的主观文本研究的议程，如 WSDM，SIGIR，KDD，CIKM，WWW，SEWM，CCIR 等。

## 1.2 研究现状

情感倾向性问题由 Hatzivassiloglou 等人率先提出。Hatzivassiloglou and McKeown[1] 通过分析从大量的未标注文档集中抽取的形容词对来判定形容词的倾向性，这些抽取出的形容词对都共现于某些连接词，如 or，and，neither-or，either-or 等。其主要处理思想是：从语言学的角度，形容词的情感倾向性受这些连接词影响，例如，and 连接的两个形容词的倾向性一致，而 but 连接的两个形容词的倾向性相反。作者依据上述思想建立了一个由所有词组成的图，图上的边有“equal-orientation”或“opposite-orientation”两种，计算图上各条边之间的相似度之后，使用聚类算法，得到 Positive 和 Negative 两个类。Turney[2]等人借助 bootstrap 算法对小规模的正负向种子词集合进行计算（正向种子词集合包括 good，nice 等，负向种子词集合包括 bad，nasty 等），计算方法用到了所谓的互信息（PMI），即对于待判断倾向性的词语 t，计算其与所有正向种子词的权重之减去其与负向种子词的权重之和，如果得到的值为正那么把 t 加入到正向种子词集合中，

反之加入到负向种子词集合中。Kamps[3]等人利用 WordNet[4]中词语的同义关系构建一个图，对于待判断倾向性的词语 t，计算 t 到图上的词 good 和 bad 的距离，如果前者大于后者那么 t 的倾向性为正向，否则为负向。随后，许多研究者使用有监督的方法判断主观文本的倾向性，Pang[5]等人分别使用 Naive Bayes，Support Vector Machines 和 Maximum-Entropy 分类方法在电影评论语料上做了有关情感倾向性分析问题的实验。Turney[6]等人介绍了一种无监督的情感倾向性分类算法，它分为三个处理部分：第一步，从文本中抽取出形容词和副词短语；第二步，预测它们的情感倾向性，第三步，根据每个短语的平均情感倾向性值对其的最终进行判断。其中核心的第二步用到了 PMI-IR 算法。

近两年来，为了满足用户强烈的个性化要求和市场的需要，情感倾向性分析的研究工作的重点逐步向更细粒度的倾向性判定和跨领域情感倾向性判断方向转变。前者的研究是倾向性判定从以往只在篇章级别上进行深入到篇章所包含的若干个评价方面，以挖掘评论者的深层次意图，为用户提供更详细的倾向性分析结果。后者是利用已标注情感倾向性的领域文本，使用某种算法来判定未知情感的另一个领域文本的情感倾向性。Snyder [7]等人处理了评论语料中多个aspect的排序问题，Zhu [8]等人提出了一种无监督方法来抽取评论语料中不同的Aspect的方法。Wu[9]等人利用图排序算法处理跨领域情感倾向性分析问题，文中同时考虑了新旧领域之间文档的相似度从而对每个文档相似度进行赋值。首先，利用旧领域和新领域文本之间建立内容相似矩阵，然后对该矩阵标准化后得到旧领域中与新领域每个文档的相似度最大的前K个文档，然后，使用同样的方法找到新领域中文档内容之间的相似文档。最后，依据标注文本的倾向性和得到的K个相似文档，计算各个文档的情感分数，对上述两个情感值进行线性加和后得到文档的最后情感分。Pan[10]等人借助与领域独立的词语作为连接源领域和目标领域的桥梁，提出SFA算法，把不同领域的词语映射到统一的潜在空间，从而对目标领域的文本进行情感倾向性判断。

同时，国内外举行的信息检索会议上都有关于主观性文本研究的主题论坛和评测，其中影响最大的是TREC，NTCIR和COAE。

（1）文本检索会议TREC（Text Retrieval Conference）。它是信息检索领域最有影响的评测活动，从1992年开始，每年举办一次，它是由美国国家标准技术研究所（NIST：National Institute of Standards and Technology ）和美国国防部（U.S. Department of Defense）主办。它针对不同的任务设立多种Track，其中Blog Track[11]是处理主观文本中blog中的行为，它的track从2006年开始先后设立观点检索（Opinion Retrieval Task）、

细粒度的博客检索（Faceted blog distillation）和主要故事识别（Top stories identification）等任务，主要处理的是与Blog有关的各项研究工作，至今已经吸引了国内外许多优秀的大学参加评测，有利地推动了Blog检索和情感倾向性的研究发展。

（2）日本国家科学信息系统中心信息检索系统测试集会议（NTCIR：NACSIS Test Collections for IR）[12]，该会议是由日本国立信息研究所主办，主要负责中、日、韩等亚洲语种的评测，该评测设立了多项评测任务，既包括传统的信息检索（Information Retrieval）和文本摘要（Text Summarization），也有关于主观性文本的评测，如问答系统（Question Answer）、多语言观点检索（Multilingual Opinion Analysis）、观点持有者抽取（Opinion Holder Extraction）和评论对象识别（Opinion Target Identification）等。自1997年设立以来，每一年半举行一次，每届的评测任务会根据上届的评测项目和当时相关领域研究的热点问题进行设置，参赛人员主要来自美国、日本、韩国、中国大陆等研究机构，参数单位的规模不断壮大，该评测的影响力也随之逐步增强。

（3）中文倾向性分析评测（COAE：Chinese Opinion Analysis Evaluation）[13]，由中文信息学会信息检索专业委员授权成立的中文倾向性分析评测委员会具体负责组织实施，吸引了国内外多家评测单位的参与，会议旨在探索中文倾向性分析的新技术、新方法，建立中文倾向性分析研究的基础数据集和评测标准，进一步推动中文倾向性分析技术的发展和应用。迄今为止，该评测已经成功举行两次，其中倾向性分析任务共5个，总体上分为词语级、句子级和篇章级三组，词语级的任务包括情感词识别及分类，句子级的任务有情感句识别及分类和观点句抽取，篇章级的任务要求实现观点评价对象抽取和观点检索，该评测共吸引国内外十余家知名科研机构的20多个研究团队积极参与，取得了良好的效果，有利地推动了中文倾向性研究的发展。

## 1.3 本文的工作

本文主要研究针对评论文本的情感倾向性判定及其应用问题的方法。跨领域情感倾向性分析问题是指利用已标注情感倾向性的源领域文本，判定未知情感倾向性的目的领域文本的倾向性。本文提出一种基于 SimRank 算法的跨领域情感倾向性判定的算法，它利用领域无关词作为连接两个领域的桥梁，借助情感词汇本体和 HowNet 等外部资源实现跨领域情感倾向性的判断，取得较好效果。同时，本文利用情感倾向性分析的方法，结合 LDA 模型对餐馆评论内容抽取出四个主题，然后结合用户评论和用户给出的评论等级计算餐馆评论在这四个方面的得分，使用逻辑回归进行训练，得到餐馆评论排序模型，利用此模型可以实现依据餐馆评论对餐馆进行排序。

## 1.4 本文的结构

全文分四章介绍情感倾向性分析及其应用：

第一章，综述了论文研究内容的背景和研究现状，介绍了本文的主要工作和论文的结构安排。

第二章，主要介绍情感倾向性分析的相关知识和基本技术，包括分词，词性标注，去停用词等技术，情感词汇本体等外部资源。

第三章，详细介绍本文提出的解决跨领域情感倾向性分析算法及其实验分析。

第四章，详细介绍利用餐馆评论的情感倾向性分析实现对餐馆排序的方法及实验分析。

论文的最后一部分总结本文的内容，以及下一步的工作。

# 2. 相关知识和外部资源

## 2.1 分词与去停用词

分词技术是处理中文文本所特有的，也是处理中文文本最基础、最重要的技术。所谓分词就是将连续的字序列按照一定的规范切分组合成词序列的过程。众所周知，在英文的行文中，空格把单词之间清楚地分割开来，而中文在书面表达或计算机内部存储时候，字词之间没有明显的切分标志。所以在自然语言处理研究的预处理阶段必须要进行分词，分词质量的好坏在很大程度上决定后续处理的质量。英文中固然也同样存在短语的划分问题，不过在词这一层上，处理中文要比处理英文要复杂得多、困难得多，分词处理技术甚至是当今研究的一个热点领域。

为了实现机器的自动分词，首先高效准确的分词词典必不可少，同时需要有快速的字符串匹配算法。由于汉语中广泛存在的歧义性，消歧算法的研究显得尤为突出，最后还要解决未登录词的识别问题。这些所有问题也就造成了自动分词是一项艰难的工作，需要综合利用各种资源和算法。

由于中文在基本文法上有其特殊性才出现了中文分词技术，其特殊性表现在：

（1）与英文为代表的拉丁语系语言相比，英文以空格作为天然的分隔符，而中文由于词语之间没有形式上的分隔符。

例如英语：“Knowledge is power”， 可自然分割为 Knowledge/ is/ power 三个词。

而汉语里：“知识就是力量”， 由于没有词语之间的分隔符，书写时无法切分成：知识/ 就是/ 力量。

（2）在中文里，“词”和“词组”边界模糊

现代汉语的基本表达单元虽然为“词”，且以双字或者多字词居多，但由于人们认识水平的不同，对词和短语的边界很难去区分。例如：“对煽风点火者给予处罚”，“煽风点火者”本身是一个词还是一个短语，不同的人会有不同的标准。

现今的分词算法大致可分为基于字符串匹配的分词方法、基于理解的分词方法和基于统计的分词方法等[14]。

正因为中文分词技术在中文文本处理中占据重要的地位，现今出现多种中文分词项目，ICTCLAS[15]，Paoding[16]，盘古分词[17]，MMSEG4J[18]等。ICTCLAS（Institute of Computing Technology， Chinese Lexical Analysis System）是中国科学院计算技术研究所在多年研究工作积累的基础上研制出的汉语词法分析系统，也是最早的中文开源分词项目之一，该系统曾在国内 973 专家组组织的评测中活动拔得头筹，在第一届国际中文

处理研究机构 SigHan 组织的评测中也都获得过多项第一名。ICTCLAS3.0 的性能优越，分词速度单机 996KB/s，分词精度 98.45%，API 不超过 200KB，各种词典数据压缩后不到 3M。ICTCLAS 全部采用 C/C++编写，支持 Linux、FreeBSD 及 Windows 系列操作系统，支持 C/C++、C#、Delphi、Java 等主流的开发语言。本文所有中文文本的分词处理都用该系统实现。

停用词是指文本中出现频率很高，但实际意义又不大的词。这一类主要包括了语气助词、副词、介词、连词等，通常自身并无明确意义，只有将其放入一个完整的句子中才有一定作用的词语。如常见的“的”、“在”、“和”、“接着”之类，比如“我是大连理工大学的一名硕士研究生。”，这句话中的“是”、“的”就是两个停用词，在进一步处理文本之前可以把它们去掉，因为它们对文本处理无有效作用，并且会影响算法的复杂度。本文在预处理阶段对所有的文本都进行了去停用词的操作。

## 2.2 词性标注

自然语言处理研究的最终目标是分析和理解语言，经过分词处理后只是将字串划分为词，即便如此仍然不能够达到目标，故而许多自然语言处理研究者进行了一些中间任务的探讨，即在不需要完全理解语言的情况下如何理解语言的内在结构，词性标注便是其中之一。在我们的情感倾向性分析中同样也要用到词性的相关信息，因为具有某些词性的词语，比如形容词等具有较强的情感，可作为判断文本情感的依据。

在指定的语境中确定文本句子中各词的词性和概念的工作就是词性标注。标注的任务是为句子中的每个词语都标上一个合适的词性，也就是确定每一个词是名词、形容词、副词或其他词性。例如下面的句子是标注词性后的效果：

汉语/n 学习/n 十分/adv 重要/adj。

注意，在此“学习”被标注为名词，但是在很多情况下，“学习”是以动词形式出现的，比如：

我/pron 要/v 好好/adv 学习/v。

由此可知，自动分词后的文本是一个词串：$w_1w_2\cdots\cdots w_n$，对其中每个词 $w_i$ 孤立而言它可能有多种词性的可能，但是，一般而言，在上述文本的特定语境下，每个词的词性都是唯一确定的，这也是人们能正确理解给定文本的基础。如何对给定文本中的词，根据它在文本中的语境确定它的词性就是词性标注的任务。

大多数的标注算法可以归纳为三类：一类是基于规则的标注算法（rule-based tagger），一类是随机标注算法（stochastic tagger），最后一类是混合型的标注算法。基

于规则的标注算法一般都包括一个手工制作的歧义消解规则库；随机标注算法一般会使用一个训练语料库来计算在给定的上下文中某一给定单词具有某一给定标记的概率，如基于 HMM 的标注算法；而混合型标注算法具有上述两种算法的特点，如 TBL 标注算法。

本文针对评论文本所用的词性标注由上节所介绍的 ICTCLAS3.0 实现。它包括的汉语词性标记集共计 99 个，22 个一类，66 个二类，11 个三类，具体的标记如下表所示

表 2.1　词性标记

Tab. 2.1　POS

| 词性 | 标记 | 词性 | 标记 |
|---|---|---|---|
| 名词 | n | 动词 | v |
| 时间词 | t | 形容词 | a |
| 处所词 | s | 区别词 | b |
| 方位词 | f | 状态词 | z |
| 代词 | r | 介词 | p |
| 数词 | m | 连词 | c |
| 量词 | q | 助词 | u |
| 副词 | d | 叹词 | e |
| 语气词 | y | 拟声词 | o |
| 前缀 | h | 后缀 | k |
| 字符串 | x | 标点符号 | w |

## 2.3　相关资源

情感倾向性是指通过分析文本的论断，得到文本的倾向性，即支持或反对的态度。情感倾向性分析和其他自然语言处理的技术类似，同样少不了外部资源的辅助。除了依靠传统的语义资源的支撑，还要结合情感倾向性分析的特殊之处，考虑情感判定方面的资源。本节主要介绍后文用到的情感词汇本体[19]和 HowNet[20]，这些资源在情感倾向性判定中起着举足轻重的作用。

### 2.3.1 情感词汇本体

文本是由若干个句子组成，而词汇又是构成句子的基本元素，因此判定文本的情感倾向性首先要得到词汇的倾向性，为此我们利用大连理工大学信息检索实验室的情感词汇本体进行词汇级的情感倾向性判定。

到目前为止，心理学界对情感的划分还没有一个公认的标准，情感的分类可以从不同的角度进行分类，这主要是因为人类的情感复杂多变，并且人们对情感的认识还不够深入和全面导致的。

参照国内外比较有影响的情感分类方法，综合现有的情感词汇资源，该情感词汇本体将情感分为 7 大类，20 小类。具体划分结构如表 2.2：

表 2.2 情感分类

Tab. 2.2 Classification of words

| 编号 | 情感大类 | 情感类 | 例词 |
|---|---|---|---|
| 1 | 乐 | 快乐 | 热切、开心、笑咪咪、喜上眉梢 |
| 2 | | 安心 | 悠闲、舒服、定心丸、问心无愧 |
| 3 | 哀 | 悲伤 | 悲哀、哀伤、心如刀割，悲痛欲绝 |
| 4 | | 失望 | 憾事、绝望、灰心丧气、心灰意冷 |
| 5 | | 疚 | 内疚、忏悔、过意不去、问心有愧 |
| 6 | | 思 | 相思、思念、脉脉含情、朝思暮想 |
| 7 | 惊 | 惊奇 | 奇怪、奇迹、大吃一惊、日升月恒 |
| 8 | 好 | 尊敬 | 景仰、敬重、毕恭毕敬、肃然起敬 |
| 9 | | 赞扬 | 英俊、优秀、别开生面、独具特色 |
| 10 | | 相信 | 信任、信赖、保证、毋庸置疑、 |
| 11 | | 喜爱 | 倾慕、宝贝、一见钟情、爱不释手 |
| 12 | 惧 | 慌 | 慌张、心慌、不知所措、手忙脚乱 |
| 13 | | 恐惧 | 胆怯、害怕、担惊受怕、胆颤心惊 |
| 14 | | 羞 | 害羞、害臊、面红耳赤、无地自容 |
| 15 | 恶 | 烦闷 | 憋闷、烦躁、心烦意乱、自寻烦恼 |
| 16 | | 憎恶 | 反感、可耻、恨之入骨、深恶痛绝 |
| 17 | | 贬责 | 呆板、虚荣、杂乱无章、心狠手辣 |
| 18 | | 妒忌 | 眼红、吃醋、醋坛子、嫉贤妒能 |
| 19 | | 怀疑 | 多心、生疑、将信将疑、疑神疑鬼 |
| 20 | 怒 | 愤怒 | 叱喝、恼火、大发雷霆、气急败坏 |

该词汇本体可以用下面所示的三元组来描述：

$$Lexicon = （B， R， E）\quad (2.1)$$

其中：B 表示词汇的基本信息，主要包括编号，词条，对应英文，词性，录入者和版本信息。R 代表词汇之间的同义关系，即表示该词汇与哪些词汇有同义的关系。该部