# 网络评论文本的细粒度情感分析研究 [>>](marginnote3app://note/522E680D-C15A-4EA5-ACE7-1D618142DBC4)

## 摘要 [>>](marginnote3app://note/1845BE21-87B2-49D5-BE82-56A90031394E)

### （1）针对文本预处理问题，基于构建的评论特征分类来识别垃圾评论，并构建用户词典改善中文分词 [>>](marginnote3app://note/D8AF3EE6-691B-41C6-AC08-13D11FFBDC07)

### （2）基于 CRFs 模型抽取情感要素，将情感对象、情感词及情感修饰词的联合识别任务转化为结构化序列标注任务 [>>](marginnote3app://note/9224369C-4063-42D5-BF86-DC9BBFB0C653)

### （3）提出了基于语境情感消歧的对立观点情感强度分析算 [>>](marginnote3app://note/E520596D-A665-4E91-A23A-3E4C90E85313)

### （4）设计并实现了评论文本细粒度情感分析系统 [>>](marginnote3app://note/4F6F2558-D131-403E-BEA7-BBA178984005)

## 第一章 绪论 [>>](marginnote3app://note/449C2278-609B-4870-9628-2367128013C5)

### 网络评论文本的情感分析 [>>](marginnote3app://note/D0A31B37-F6C7-407D-AC95-DBBFDE9ECFFD) 网络评论文本的情感分析[2-6]（sentiment analysis） ，是在自然语言处理技术帮助下快速整理和分析相关的评论信息，并且对主观倾向的文本进行处理、分析、归纳和推理的过程。

#### 情感分析主要任务是情感要素的抽取和分析。 [>>](marginnote3app://note/D8EBF5DB-E382-43A8-A5B3-5C5B7F42E57C)

### 篇章级的情感分析 [>>](marginnote3app://note/E9B2B97A-845B-42C9-93A0-9251BC7C207D) 篇章级的情感分析主要分析类似于新闻报道类的整篇文章

#### 包括有利用监督学习方法分类电影评论[7]， 分析对于某社会热点话题的新闻评论，挖掘酒店领域评论，最终返回是属于正向的还是负向的情感。 [>>](marginnote3app://note/13C95E39-E93F-4DD8-9C23-B72D8B30389B)

### 细粒度情感分析 [>>](marginnote3app://note/ECC7A4F4-8DF2-4C87-82E3-A0287ACA35D0) 细粒度情感分析也称为基于特征的意见挖掘，对电子商务和产品推荐等实际应用提供更具体的参考信息，近几年受到了追捧，主要任务与粗粒度的情感分析不同，面向产品特征级别抽取情感要素。

#### 细粒度的情感分析研究最早采用数据挖掘算法基于统计方法抽取评价要素，然后根据规则进行过滤，比如情感对象及产品属性的抽取，基于词性及词频，利用语法规则过滤掉高频的词语，并提取情感对象和情感词，以此实现细粒度要素的抽取[20]。 [>>](marginnote3app://note/5EB9E3E2-E7C6-4E3F-A050-3AC4DE54FF36)

### 情感词典的构建 [>>](marginnote3app://note/A0B0F23F-AF9C-4602-8B74-872192A8A3EA) 不管粗粒度还是细粒度的情感分析都依赖于情感词，词是最小的语义单元，而作为判断情感词极性的基础来源，情感词典的构建一直以来都是观点挖掘的重要研究方向。

#### 对英文和中文情感词典的构建做了很多工作，主要包括基于知识库的方法[36-39]、 基于语料库的方法[40-43]和基于语料库与知识库相结合的方法[44-46]。 [>>](marginnote3app://note/B9B72036-A8B9-4353-8004-334D464C57B3)

##### 基于知识库 [>>](marginnote3app://note/3000261A-6725-4550-8BE7-DD9023080BFF) 一种是基于知识库的方法，适合于具有开放完备语义知识库的语言，通过挖掘词与词之间的关系构建通用性比较强的情感词典[36]，情感知识库存在领域适应性问题，对于在不同领域或主题下具有不同情感倾向性的词语不适用，缺乏领域情感词典，各不同领域语料集中的词语情感极性不容易判断。

##### 基于语料库 [>>](marginnote3app://note/4E16F110-2774-4045-A82F-8754F2F3334E) 一种主要通过词语共现[40]和连词关系[41]基于语料库构建领域情感词典，比如网络评论文本的主观性较强，具有明显的针对商品属性的观点评价，最适宜用连词关系法构建领域情感词典。

##### 结合两种方法 [>>](marginnote3app://note/A2D61497-671A-42DD-8F00-D0DE60558E66) 以上两种方法在准确率和通用性上仍有局限，目前大都采用结合两种方法互补的方法，根据知识库先验知识构建种子词集，再基于语料库构建共现位置情感等信息，得到情感词的情感极性，构建更为完善的领域情感知识库[46]。

### 网络评论文本的细粒度情感分析处理流程图 [>>](marginnote3app://note/36C98123-7354-4086-A19A-929E57768DB5)

## 第二章基于评论特征分类的垃圾评论过滤 [>>](marginnote3app://note/130767E5-56B4-4DF2-9A55-590E7E35D0DE)

### 垃圾评论过滤 [>>](marginnote3app://note/DC97A769-27F3-4A6C-86DD-D7CC1F3F73A9)

#### 垃圾评论过滤 [>>](marginnote3app://note/7FCDAF42-EB67-4272-8F5B-FE07EBFA54D1) 垃圾评论过滤的目的是保留出真实有效的商品评论，其首要任务是要先识别出评论中哪些属于垃圾评论，目前关于垃圾评论识别方面的研究大多集中在评论用户行为特征、垃圾主题、产品特征、情感特征和用户评论可信度等方面

#### 垃圾评论 [>>](marginnote3app://note/648E9E16-B632-4F44-A753-023C4E13BF5C) 垃圾评论（Review Spam）也称为 Opinion Spam，根据不同视角界定概念可分为多种类型，垃圾评论制造者国外称为 Spammer，国内称为“网络水军”，狭义的垃圾评论指一些用户发表的虚假不真实或欺骗性的评论，广义的垃圾评论也指非观点型的不包含有价值情感信息的评论，本文文本预处理过程中过滤的垃圾评论即包含虚假评论，也包括无关评论。

#### 论文将垃圾评论识别转换为二元分类问题，根据构建的评论特征构建训练语料集，然后采用朴素贝叶斯分类器对评论文本进行文本分类，对应于 Spam 标记， 垃圾评论标记为 0，真实价值评论标记为为 1。 [>>](marginnote3app://note/E7AC1740-6C9B-4B15-AF87-39605139CE81)

#### 构建评论可信度 Credibility 评估函数 [>>](marginnote3app://note/CDBD9B50-41EC-40DC-9BB9-9CB87EF2159E)

##### [>>](marginnote3app://note/00CD40D2-932C-4CF0-B0B8-DD384767DF7C)

### 中文自动分词 [>>](marginnote3app://note/89A821CA-A42C-41E9-8414-BAF32127F0EF) 中文自动分词是中文信息处理中的前提基础，与英文文本不同，词是最小的语义单元，且中文句子结构表达都很复杂，各不同词性语气标点等都包含信息，由词组成句再连成篇，没有英文明显的空格分隔，所以，分词成为了处理中文信息的首要工作和基础。

### 用户词典功能 [>>](marginnote3app://note/46B5B391-AAFE-4E3E-8C11-49129405955A) 本文利用 NLPIR 的用户词典功能，将识别出的新词添加到用户词典，并将领域术语类的关键词添加到用户词典，容易切分错误的网络新词也添加入用户词典，在分词过程中出现的切分错误的词语也可手动更改添加到用户词典。

## 第三章基于 CRFs 的情感要素抽取模型 [>>](marginnote3app://note/AF194E9A-8A04-4AA5-AA54-B65BEC80DB5C)

#important

### 网络评论文本的细粒度的情感分析也称为基于特征的情感分析，关键任务在于从评论文本中抽取情感要素，即准确提取所要分析的情感对象 [>>](marginnote3app://note/98E220B4-CFF4-4E4F-BDA2-911A57398821)

#### 电商评论的情感对象对应产品属性词，并抽取情感对象对应的带有情感色彩的评价词，也称情感词，还要抽取对情感分析有反转增强等影响的情感影响因子，抽取出文本中更具体的情感要素，为一些实际应用提供必要的细节信息 [>>](marginnote3app://note/93096CD2-D3F8-4DB0-9B7D-7006CBC5292D)

### 条件随机场（CRFs） [>>](marginnote3app://note/A1DB12D5-751A-481F-BC11-8E068A29DF9D) 条件随机场（CRFs）是 Lafferty 等人[62]在 2001 年基于最大熵模型（MEM）和隐马尔可夫模型 （HMM） 提出的无向图模型。

#### CRFs 模型的三个关键步骤是特征选取，参数评估和模型推理。 [>>](marginnote3app://note/897F22C7-C44A-4E49-B6B2-37BE15A2DE1B)

#### 大多数传统方法使用 CRFs 来提取文本的情感对象， 而忽略了句中的情感词特征对句子的情感性影响。 [>>](marginnote3app://note/C04C2973-7239-4BEF-A4BF-AA390BBEB97F)

#### 评论语句可看成由词组成的序列， 采用 CRFs 序列标注情感要素[63]， 为实现情感要素的有效识别，将标注集设定为四种简单的标记来防止特征稀疏。 [>>](marginnote3app://note/589D40AF-8D8C-4269-AAFF-D6D86505A490)

##### [>>](marginnote3app://note/AFD48486-65BA-4644-BBDC-CF0CBE9F36E2)

#### [>>](marginnote3app://note/3247356D-F976-4F2D-A681-D8EE7F419561)

### 实验 [>>](marginnote3app://note/C8F66FE3-8740-4682-9D1C-09F8D34564C2) 对系统数据采集模块采集到的评论数据进行预处理，即第二章中介绍的垃圾评论过滤、NLPIR 分词、用户词典构建和意群划分。本章实验输入数据为已经预处理后的意群集，通过 CRFs 进行情感要素序列化标注。

### 隐式情感对象 [>>](marginnote3app://note/695E6498-122D-4E31-9B06-A30DFFD054AB) 另一种是在评论语句中没有直接的对产品的性能进行描述，用户需要对评论语句进行语义理解，根据评价词和上下文自主推断才能获取的隐式情感对象，如：“手机太贵了”，可以从“贵”这个情感词推断出评论句所评论的隐式情感对象为“价格”[64]。

### 用朴素贝叶斯分类器识别隐式情感对象 [>>](marginnote3app://note/81541CA5-5CCA-44BF-9495-54BE9EE02CBF) 本节采用朴素贝叶斯分类器识别隐式情感对象，充分利用评论信息，提高了细粒度情感分析的准确率。

#### 分类识别隐式情感对象 [>>](marginnote3app://note/87568876-EFCC-461B-9EDB-2CA704570EAD) 用 F 代表评论句中的情感对象，O 代表评论句中的情感词，组成特征观点对<F,O>，构建训练文档，训练朴素贝叶斯文本分类器，分类识别隐式情感对象，然后采用训练好的分类器完成所有预处理后语料集中评论句的情感要素识别。

#### [>>](marginnote3app://note/C4B52131-E76B-44BF-9E7E-D17CCDE54F6F)

## 第四章基于语境情感消歧的对立观点情感强度分析 [>>](marginnote3app://note/27C2A03F-26FA-4E5F-AC0E-294D42998E0D)

#important

### 属性的对立观点的情感强度值 [>>](marginnote3app://note/7DCB7862-142B-42DC-93EF-BA85E1E6BA01) 目前，情感分析中虽然有对评价对象的多个方面进行分析，但是反馈结果是各属性整体的情感倾向性，仍旧不能让用户快速直观的了解到属性的对立观点的情感强度值，如<正向 50.1%，负向 49.9%，正向>与<正向 99.9%，负向 0.01%，正向>，均为对同一属性的正向分析结果，但由于正负向情感强度相差悬殊，真实评价也会相差很大，迫切需要包含对立观点强度信息在内的情感分析。

### 情感词 [>>](marginnote3app://note/DBE4CD1F-AF4D-4682-A07F-CCD6F66031B0) 情感词根据语境信息可分为两种

#### [>>](marginnote3app://note/95C2A5F0-C78E-461C-B333-03EE54A41C27)

第一种是上下文无关型，具有明显的褒贬倾向，如喜欢、讨厌等；  
第一种情感词依据基础情感词典可得到确定的情感极性

#### [>>](marginnote3app://note/E3EAD4B5-0678-4F0E-8CAE-DD379A9F1AF4)

第二种是上下文相关型，如高、大、长、快等。  
第二种则需要结合情感词所搭配的上下文语境信息，根据不同词语的搭配动态的选择情感极性，本文定义此类情感词为情感歧义词。

### 挖掘词语间搭配关系强度 [>>](marginnote3app://note/A1C7F648-DECA-4D2B-918B-8F3EA0F21A7F) 通过点互信息 PMI (PointWise Mutual Information)来进一步挖掘词语间搭配关系强度。

#### PMI 的值越大表示词语1 word与2 word 间的搭配关系越强。设定一个阈值δ，过滤掉词语互信息满足的弱关联搭配对。 [>>](marginnote3app://note/7F0C71BF-942C-4BB1-B0B5-8BD397C57775)

#### [>>](marginnote3app://note/AF645CCB-931F-41A2-9D74-0D5D6EDB2C06)

### 情感歧义词搭配词典 [>>](marginnote3app://note/B5030CF2-FA5B-4C22-A4F5-3881B48DD26D) 经过 PMI 过滤后构成情感歧义词候选搭配集，利用情感词典标注搭配集中评价词语的情感极性， 进而构建成情感歧义词搭配词典， 词条存储形式： <情感对象，情感词，情感倾向性>。在情感倾向性分析时情感歧义词动态极性值的确定需同时满足词语搭配对，解决同一情感词修饰不同情感属性时不同情感倾向的问题。

### 正向和负向的情感强度计算 [>>](marginnote3app://note/216AFDA6-8E56-43FD-B245-8085016FE4DB) 对具体的每类聚合后的特征观点对进行正向和负向的情感强度计算，计算算法整合了基础情感词典、本章首先构建的情感歧义搭配词典、预处理过程构建的新词及网络、否定、连词、程度副词等词典，并提取各特征正向和负向的前十条特征观点对组成细粒度情感摘要。

### 情感词典 [>>](marginnote3app://note/53CBB69E-6AFF-4021-8C1D-9736F1A75820) 本文针对网络评论文本的特点添加了网络词汇词典、否定副词词典、程度副词词典和情感歧义词搭配词典进行特征级别的情感分析

#### 加入了句法分析对文本中出现的并列评价句类的连接词、代表情感强度的程度副词和倾向性逆转的否定词进行分析处理 [>>](marginnote3app://note/6B1BCF99-5A39-46B1-A1EE-79CB5717887B)

#### 否定词词典 [>>](marginnote3app://note/8CDBD154-00CE-4815-8EAA-00C3D756B8D3)

#### 程度副词词典 [>>](marginnote3app://note/242E26A8-AD0F-4D20-BEBB-35DF0B89D21F)

### 情感倾向计算算法 [>>](marginnote3app://note/EED9F818-8E3F-47EB-B403-3C0FD94931FC) 情感倾向计算算法的基本思想：利用情感词和影响情感的情感修饰词计算情感对象情感。

#### （1）根据否定词的情感值和程度副词的情感值，计算情感修饰词的情感影响因子 [>>](marginnote3app://note/90F73DB3-7DF5-49B3-8F0F-28BD9495319E)

##### [>>](marginnote3app://note/5A6E478E-19B4-4171-96E9-F9A2DB8FFF72)

#### （2）结合情感词的极性计算情感要素组成的属性观点对的情感极性值 [>>](marginnote3app://note/4B4F0880-1A83-40C7-8FCC-A34FEBDDC8B7)

##### [>>](marginnote3app://note/1FA1E841-454C-4353-8701-F1E639FA5A35)

#### （3）计算评论语料中产品属性的正向情感强度和负向情感强度 [>>](marginnote3app://note/C61E7A24-F00A-42D0-BEA3-EA47B022DFA7)

##### [>>](marginnote3app://note/6BA46308-51CD-4847-8A01-72F0D4ABF6BC)

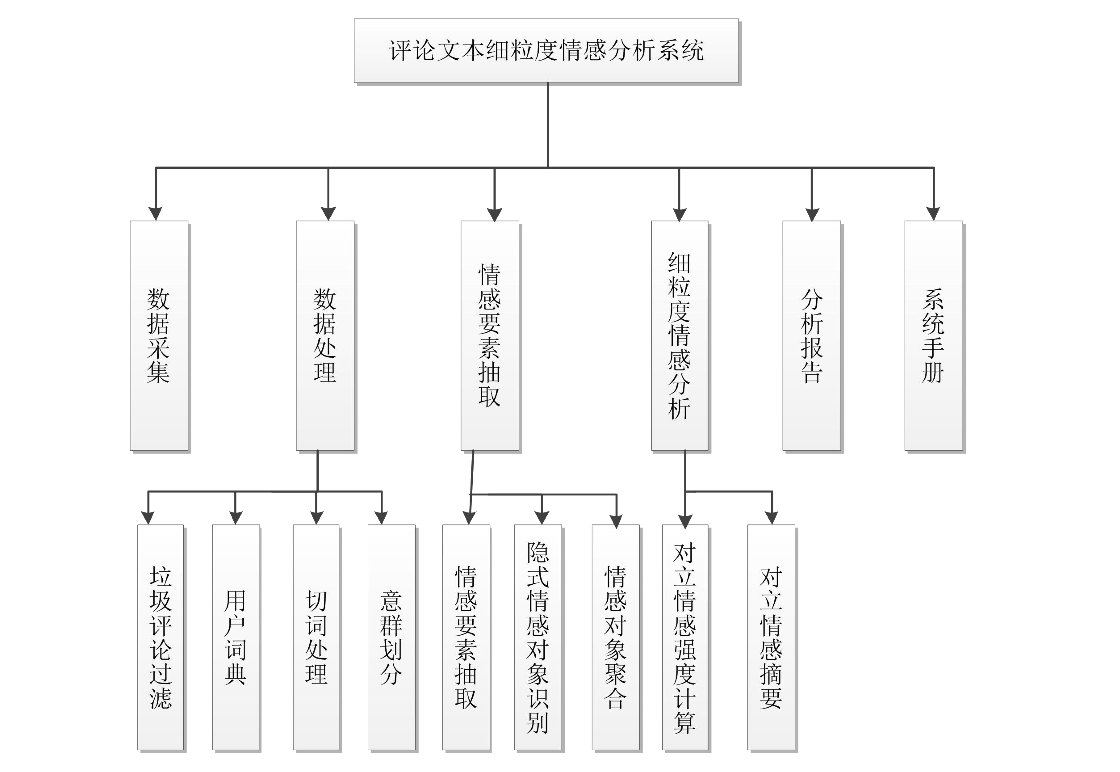
### 对立观点情感摘要 [>>](marginnote3app://note/79033E1F-89FD-4C29-8DF3-C2423113E913)

对立观点情感摘要：得到各聚合后特征观点对的正向和负向的情感强度之后，可直接由此生成情感摘要，提取前十个热门评价属性，组成<正向情感强度+“的用  
户认为”+情感对象+情感词>、 <负向情感强度值+“的用户认为”+情感对象+情感词> 形式的包含对立观点的情感摘要，以方便用户直观的了解评论情感强度。

### 对立观点情感强度分析结果示例 [>>](marginnote3app://note/BF163757-1C3D-40FC-A3BC-486C3D3D1F53)

## 第五章 评论文本细粒度情感分析系统的设计与实现 [>>](marginnote3app://note/885EB310-CD2C-4D4B-A5B4-D37A9C524BCB)

### 系统设计方案 [>>](marginnote3app://note/9F59E12B-A6D7-4434-869C-22F3BC996EC8)

系统框架图  


#### 数据采集 [>>](marginnote3app://note/0156AE4F-7E87-4C2F-B05B-5DC0F174F13C) 数据采集实现从电商平台采集评论数据，并将其保存到本地数据库中。

#### 数据处理模块 [>>](marginnote3app://note/6B2B2D77-0AD9-49C3-9773-059ACA254A4C) 数据处理模块的主要工作是对采集到的评论语料进行预处理，包括垃圾评论过滤，中文分词和词性标注， 提取新词构建用户词典。

#### 情感要素抽取 [>>](marginnote3app://note/67B00265-4CBD-43B7-BBD8-1B7D7633CA3C) 情感要素抽取模块基于 CRFs 对情感对象和情感词及情感影响因子联合抽取，隐式情感对象分类识别，最后聚合情感对象。

#### 细粒度情感分析 [>>](marginnote3app://note/F3F32B30-4118-4171-9EE5-C29F09BA1D75) 细粒度情感分析模块进行正向和负向的情感强度计算并可视化展示。图 5-1 是该系统的框架布局。

## 第六章 总结与展望 [>>](marginnote3app://note/CBF3C0A4-920B-47E4-B042-146BF6822A5F)

### 全文首先介绍了本课题的研究背景和现状，并对垃圾评论过滤和中文分词两种文本预处理任务进行了分析； 其次基于 CRFs 模型对情感要素进行抽取， 并通过分类实现隐式情感对象的抽取和情感对象的聚合处理；然后提出一种通过构建情感歧义搭配词典解决情感词动态极性变化问题的方法，进而对聚合后特征类的对立观点进行情感强度计算；最后设计了网络评论文本的细粒度情感分析系统，该系统实现了评论文本采集、分析与可视化等功能模块。 [>>](marginnote3app://note/35B596CB-0C91-4875-8665-92622A916821)

### 创新点 [>>](marginnote3app://note/A9FA9FE5-8D41-4BA3-A993-A2F1ABED245D) 采用 CRFs 联合抽取情感对象和情感词，通过朴素贝叶斯分类抽取隐式情感对象，预处理中文分词时添加用户词典改善了分词准确率，情感计算时又添加了网络词典和情感影响因子的计算，同时分析出产品特征的正向和负向对立的情感，组成细粒度情感摘要