信 息 存 储 技 术 丛 书

方面级观点挖掘
理论与方法

吕 品 钟 忺 钟 珞 黄文心 著

科学出版社

信息存储技术丛书

方面级观点挖掘理论与方法

吕 品 钟 忺 钟 珞 黄文心 著

科学出版社
北京

内 容 简 介

随着信息科技的发展，人类进入了大数据时代，挖掘互联网上海量主观性文本已成为决策支持的重要手段。从产品消费、医疗保健、金融服务到社会事件和政治选举，观点挖掘几乎渗透到现实生活中每一个领域。这些实际的应用为观点挖掘的研究提供了强烈的动机。本书在分析观点挖掘相关概念和相关技术研究的基础上，阐述了方面级观点挖掘方法的分类、如何利用 CRF 方法及主题模型进行方面级观点挖掘、以及在观点挖掘环境下实体和方面的指代消解方法。实现了在线评论的智能化观点挖掘，并进一步研究了将观点挖掘的结果应用于用户满意度评价及产品属性绩效类型界定等方面。

本书可供高等学校及研究院所信息处理与应用硕士、博士研究生及研究工作者参考，也可作为信息应用领域科技工作者的高级指南。

图书在版编目(CIP)数据

方面级观点挖掘理论与方法/吕品等著. —北京：科学出版社，2015.11
(信息存储技术丛书)
ISBN 978-7-03-046185-8

Ⅰ. ①方… Ⅱ. ①吕… Ⅲ. ①计算机网络—情报检索—研究 Ⅳ. ①G354.4

中国版本图书馆 CIP 数据核字(2015)第 259341 号

责任编辑：张颖兵 闫 陶/责任校对：董艳辉

责任印制：高 嵘/封面设计：苏 波

科学出版社 出版

北京东黄城根北街 16 号

邮政编码：100717

http://www.sciencep.com

湖北卓冠印务有限公司印刷

科学出版社发行 各地新华书店经销

*

开本：B5(720×1000)

2015 年 11 月第 一 版 印张：10 1/4

2015 年 11 月第一次印刷 字数：200 000

定价：45.00 元

(如有印装质量问题，我社负责调换)

前　言

观点挖掘是一种分析人们对被评价实体的观点、情感、评估、评价、态度和情绪的方法，其主要任务包括被评价实体及对其评价的抽取、分类、分析等。随着信息技术的发展，互联网上的在线评论大数据已成为人们决策支持的有价值资源。一方面，个人可以参考这些评论内容，提前了解自己关注的对象，以做出更符合自身利益的决定；另一方面，行业组织诸如公司或一些商业组织可以参考消费者或相关对象反馈的评论内容，适时调整应对措施，提高自身的服务质量或行业的经营效益。

本书在分析观点挖掘的相关理论和方法研究现状的基础上，阐述了围绕观点挖掘方法和其在在线评论领域应用的研究工作，受到国家自然科学基金、支撑计划子项、科技创新团队等项目支持，研究工作包括基于 CRF 的方面级观点挖掘和指代消解方法，利用主题模型进行方面级观点挖掘的方法，以及基于观点挖掘的结果对用户满意度进行了评价和对产品属性绩效类型进行了界定的方法。书中详细阐述了每项研究的解决思路、算法、实验结果、结论等。

本书由吕品博士、钟忺博士、钟珞博士及黄文心博士撰写。本书撰写过程中得到了许多专家和学者的指导，还借鉴和引用了大量国内外有关文献及研究成果。在此向他们表示衷心的感谢！

研究工作是无止境的，加之研究时间有限，书中会存疏漏之处，恳请专家学者提出宝贵意见。

著　者

2015 年 8 月

目　录

第 1 章　观点挖掘 …… 1
1.1　从数据挖掘到观点挖掘 …… 1
1.2　观点挖掘的含义及应用 …… 2
1.3　观点挖掘研究 …… 4
1.3.1　观点挖掘的不同级别 …… 4
1.3.2　情感词典及相关问题 …… 5
1.3.3　自然语言处理问题 …… 6
1.4　垃圾观点检测 …… 6

第 2 章　观点挖掘目标 …… 8
2.1　实体与观点 …… 8
2.2　观点挖掘的目标 …… 10
2.3　观点文摘 …… 13
2.3.1　方面级观点文摘的特点 …… 13
2.3.2　方面级观点文摘的改进 …… 15
2.3.3　方面级对比观点文摘 …… 18
2.3.4　传统短文本文摘 …… 18

第 3 章　方面级观点挖掘方法 …… 20
3.1　方面抽取 …… 20
3.1.1　基于频率的方法 …… 20
3.1.2　基于关系的方法 …… 21
3.1.3　混合方法 …… 23
3.1.4　监督学习方法 …… 24
3.1.5　主题模型方法 …… 26
3.2　方面分类 …… 29
3.3　方面情感分类 …… 31

第 4 章　基于 CRF 的方面级观点挖掘 …… 34
4.1　挖掘方法 …… 34
4.1.1　准备符号标记与训练集 …… 34
4.1.2　定义学习函数 …… 36
4.1.3　训练 CRF 模型 …… 36
4.1.4　CRF 模型的使用 …… 37
4.2　实验 …… 38
4.2.1　数据集 …… 38
4.2.2　性能评估方法 …… 38
4.2.3　用 CRF 挖掘中文评论的实验结果 …… 39
4.2.4　差异显著性检验 …… 40
4.2.5　CRF 方法与其他方法挖掘中文评论的比较 …… 42

第 5 章　主题模型的设计决策 …… 44
5.1　主题因素 …… 44
5.1.1　全局主题与局部主题 …… 44
5.1.2　主题与方面的映射 …… 45
5.1.3　基于约束的方面分类 …… 45
5.1.4　基于需求的方面分类 …… 45
5.2　情感因素 …… 46
5.2.1　情感层 …… 46
5.2.2　情感之间的依赖 …… 47
5.2.3　情感与方面的分离 …… 47
5.2.4　情感先验信息 …… 48
5.3　相关因素的实现技术 …… 48
5.3.1　主题因素的实现 …… 48
5.3.2　情感因素的实现 …… 52
5.4　相关因素的选取策略 …… 55

第 6 章　基于 TMDP 的方面级观点挖掘 …… 58
6.1　TMDP 模型 …… 58
6.1.1　模型参数的物理含义 …… 58
6.1.2　模型表示 …… 59
6.1.3　模型生成过程 GP-TMDP …… 61
6.1.4　参数 τ 的设置 …… 62

6.2　模型参数估计 …… 62
6.2.1　基于 collapsed 吉布斯抽样的主题参数估计算法 Gibbs-TE …… 62
6.2.2　基于情感预测的主题-方面映射算法 TAMSP …… 65
6.2.3　基于 collapsed 吉布斯抽样的方面-情感分离参数估计算法 Gibbs-ASS … 66
6.3　实验结果及分析 …… 67
6.3.1　数据集 …… 67
6.3.2　主题与方面对应关系的定性分析 …… 68
6.3.3　主题与方面对应关系的定量分析 …… 70
6.3.4　局部主题对方面识别的贡献 …… 72
6.3.5　观点词识别的性能分析 …… 74
6.3.6　观点词和方面词之间关联的评估 …… 76
6.3.7　特征选择对方面与情感分离的影响 …… 77
6.3.8　标签数据尺寸对方面与情感分离的影响 …… 78
6.3.9　不同领域数据集对方面与情感分离的影响 …… 79

第 7 章　基于 TMPP 的方面级观点挖掘 …… 81
7.1　TMPP　模　型 …… 81
7.1.1　基于短语级参数学习的主题模型演化 …… 81
7.1.2　基于短语级参数学习的主题模型中的参数 …… 83
7.1.3　模型生成过程 …… 85
7.2　模型参数估计 …… 85
7.2.1　AR-PLSI 模型参数估计 …… 85
7.2.2　AR-LDA 模型的参数估计 …… 86
7.2.3　ARI 模型的参数估计 …… 88
7.3　实验结果及分析 …… 90
7.3.1　数据集 …… 90
7.3.2　方面识别分析 …… 91
7.3.3　等级预测分析 …… 92
7.3.4　模型的适应性分析 …… 95

第 8 章　基于 CRF 的实体和方面的指代消解 …… 99
8.1　已有的指代消解方法 …… 99
8.2　针对观点挖掘环境提出的新特征 …… 100
8.2.1　观点极性一致性特征 …… 101
8.2.2　实体和方面与观点词的关系特征 …… 101

8.3　新特征的实验分析 …… 102
8.3.1　数据集与特征集 …… 102
8.3.2　基线方法 …… 103
8.3.3　新特征的实验结果分析 …… 104
8.4　基于 CRF 的指代消解 …… 105
8.4.1　约束化局部训练 CLT …… 105
8.4.2　基于 CLT 的指代消解 …… 106
8.4.3　基于 CLT 的指代消解实验验证 …… 109
8.4.4　CLT 的优化 …… 110
8.4.5　基于 CLT 的指代消解方法与分类方法的比较 …… 114

第 9 章　用户满意度评价及产品属性绩效类型界定 …… 116
9.1　情感词相似性判定选择 …… 117
9.2　基于协同过滤推荐算法的用户对产品情感满意度排序 …… 118
9.2.1　用户情感满意度评价指标体系 …… 118
9.2.2　构建用户对产品的评分矩阵 …… 119
9.2.3　构建用户相似集 …… 120
9.2.4　求解用户情感满意度推荐排序分值及结论 …… 120
9.3　基于雷达图法的产品属性绩效类型界定 …… 121
9.3.1　产品绩效评价指标体系 …… 121
9.3.2　产品绩效评价指标的指标权重 …… 123
9.3.3　基于雷达图法的产品属性绩效类型的判据界定 …… 126
9.3.4　基于雷达图指标赋权值的分数值求法 …… 127
9.3.5　产品绩效特征类型界定雷达图 …… 129
9.4　基于灰色评估法的用户满意度评价 …… 129
9.4.1　用户满意度评测指标体系 …… 130
9.4.2　用户满意度的灰色评估模型 …… 131
9.4.3　用户满意度的灰色评价仿真运算 …… 133
9.4.4　软件实现的程序流程 …… 136
9.4.5　求解用户满意度评价因素用户等级结论值 …… 136

参考文献 …… 139

第1章 观点挖掘

观点的重要性在于它是人类所有活动的中心，对人类的行为有着重要的影响，而互联网上在线评论的观点更为重要。随着 Web 2.0 应用的发展，互联网上出现了海量的包含观点的文本（主观性评论内容），如在线评论、论坛讨论、博客、微博等，它们具有巨大的潜在价值。因而，利用计算机进行自动的在线评论观点挖掘是 Web 应用的重要组成部分和 Web 应用的发展趋势。

观点挖掘是一个研究和分析人类对事物的观点、情感、评估、评价、态度和情绪的领域，其中事物通常用实体来描述。实体可以是产品、服务、组织、个人、问题、事件或主题等。在产业界，常用“情感分析”表示观点挖掘。观点挖掘应用于消费产品、服务、医疗、保健、金融服务、社会事件、政治选举，还涉及指代消解，否定形式的处理和词义消歧的自然语言处理问题。

从主观性文本中挖掘观点，有篇章级观点挖掘、句子级观点挖掘和实体与方面级观点挖掘三种不同的级别。为保证观点的可利用性要监测且消除垃圾观点。

1.1 从数据挖掘到观点挖掘

数据挖掘成为热点研究领域已有相当长的时间[1]，与数据挖掘相关的基础学科的理论与技术也取得了许多阶段性的成果。数据挖掘应用非常广泛，包括文本挖掘[2]、Web 挖掘[1]、图像与视频的挖掘[3-6]等。

早期的文本挖掘主要挖掘和检索文本中的事实信息。然而，随着 Web 2.0 的迅速发展和 Web 3.0 的兴起，人们可以通过互联网上的社会媒体对自己感兴趣的事件、在网络上消费的产品或服务等发表自己的看法与观点，这就使得文本中包含了大量的主观信息。所谓主观信息即主观性观点，它描述了人们对实体，事件或它们的属性的一种情绪，评估或情感的主观表达。一般，把包含有主观信息的文本称之为主观性文本。

互联网作为一个聚集了海量主观性文本的平台，已成为决策支持的一个有价值资源。企业、普通用户和政府对这种主观性文本挖掘都有需求。对企业而言，通过了解消费者的观点可获取市场情报、消费者的消费趋势；寻找一些商业机会；对产品声誉进行有效管理、针对特定人群的广告精确投放等。对普通用户而言，了解

他人的消费评价，能帮助自己进行合理消费。对关心政治或社会事件的普通用户，可让他们快速了解热点的政治话题或新闻事件。对政府而言，了解公众的观点，可以控制舆情、监测社会事件。2014 年“两会”期间，中国政府首次增加了两会大数据版块，运用大数据分析技术分析用户的搜索行为，及时感知网民情绪、洞察两会的社会影响。时任国务院总理李克强在 3 月 5 日的政府工作报告中提出进一步加强互联网金融、扩大跨境电子商务试点。大数据分析在两会中的应用，政府工作报告中的互联网金融和电子商务信号充分说明了对以中文语料为研究对象的主观性文本挖掘应用前景美好。

1.2 观点挖掘的含义及应用

“观点挖掘”又称“情感分析”，最早出现于 2003 年[7-8]。然而，关于“情感”或“观点”的研究更早[9—14]。在产业界，常用“情感分析”表示观点挖掘。在学术界，“情感分析”与“观点挖掘”频繁交替使用。为了简化表达，本书用术语“观点”表示人类对实体的观点、情感、评估、评价、态度和情绪，实际上这些概念并不等价。再加上“观点”本身的含义非常广泛，因而，本书的内容主要涉及具有肯定情感极性的观点和否定情感极性的观点。

观点挖掘逐渐成为一个活跃的研究领域的主要原因在于：

(1) 应用领域广泛。由于商业应用的增殖，围绕着情感分析的产业得到了蓬勃发展。这为科学研究提供了一个强有力的动机。

(2) 应用过程中出现了许多挑战性的研究问题，这些问题以前从未研究过。

(3) Web 上的社会媒体含有海量的主观性信息(包含观点的数据)。

毋庸置疑，观点挖掘是社会媒体研究的中心。因此，观点挖掘的研究不仅对自然语言处理有重要的影响，而且对被人类观点所影响的管理科学、政治科学、经济学和社会学也有重要的影响。

观点挖掘是一个研究和分析人类对事物的观点、情感、评估、评价、态度和情绪的领域。其中，事物通常用实体来描述。实体可以是产品、服务、组织、个人、问题、事件或主题等。观点的重要性在于它是人类所有活动的中心，对人类的行为有重要的影响。在现实世界中，每当我们需要做一个决定之前，都希望能了解他人的看法；为了改进产品或改善服务以便产生更多的利润，企业或公司也希望能了解消费者对他们的产品或服务有何意见。再例如，个体消费者在决定是否购买某一产品时，也想了解已购买过该产品的用户对这个产品的评价；为了维护自己的选举权，公民做投票决定之前，常常想了解其他公民对他想推荐的候选人的态度。在信息

不发达的时代，人们常使用传统的观点获取方式，如询问朋友或家庭、调研、民意调查或跟踪目标人群等。通过这种方式获取的观点不仅耗时，而且数量有限。随着Web应用的迅猛发展，人们没必要使用传统的观点获取方式，因为社会媒体（网络站点上的评论、论坛讨论、博客、微博）上已有大量的包含观点的文本，人们可以直接利用这些内容作决策。例如，如果你想购买一个产品，不需要再询问家人或朋友了，因为可在公共论坛上找到关于这个产品的用户评论和讨论；为了改进产品或改善服务，企业或公司不再需要执行调研、民意调查和跟踪目标人群，因为网络上有大量关于该产品或服务的评论或讨论的公共信息。

然而随着网站数量的激增，并且每一个网站通常包含大量的主观性文本，再加上观点在内容较长的博客中或论坛贴子中不易破译，寻找和监测Web上的观点网站并提取包含在它们中的信息是一项艰巨的任务。因此，研究一些观点挖掘方法来开发自动的观点挖掘系统将成为趋势。

近年来，从社会媒体中挖掘出的观点对商业、对公众的情绪以及对社会和政治制度产生了深刻影响。因此，有必要收集和研究Web上的观点。当然，主观性文本不仅存在于Web上（外部数据），许多机构也有它们自己的内部数据，例如，从E-mail、呼叫中心或调研中心收集的消费者的反馈意见。

从消费产品、服务、医疗保健和金融服务到社会事件和政治选举，观点挖掘几乎已渗透到每一个可能的领域。研究者Liu已实现了一个观点挖掘系统Opinion Parser[15]，并创办公司研究以上领域中的项目。仅在美国，研究开发观点挖掘系统的公司至少有40～60家。此外，许多大公司还在自己的产品中构建了观点挖掘的模块，例如，Microrsoft，Google，Hewlett-Packard，SAP和SAS。由此可见，这些实际的应用和产业兴趣为观点挖掘的研究提供了强烈的动机。

除了实际的观点挖掘应用之外，研究者们也发表了许多面向应用的研究论文。例如，刘晶晶等学者于2007年提出了一个情感模型用于预测销售性能[16]。Mary等用评论排序产品和商人[17]。洪炎诚研究了美国国家足球联盟博彩市场与博客和Twitter中公众的观点之间的关系[18]。Brendan等研究了Twitter情感与民意调查的联系[19]。Andranik等利用Twitter情感预测竞选结果[20]。陈毕研究了政治立场[21]。Ten报道了一种预测政治博客评论的方法[22]。Sitaram等学者利用了Twitter数据、电影评论和博客预测电影票房的收入[23-25]。Mahalia等研究了社会网络中的情感流[26]。Saif等学者使用E-mail中的情感寻找性别对情感的影响[27]，跟踪了小说和童话故事中的情感[28]。Johan等使用了Twitter情绪预测股票市场[29]。Ronen等识别了微博中的专家投资者，并且执行了股票的情感分析[30—31]。张文斌使用了博客和新闻中的情感研究了贸易策略[32]。Sitaram等学者研究了在线书籍评论的社会影响[33]。Georg等使用情感分析特

征化社会关系[34]。Malu 等研究了一个综合的情感分析系统和一些案例研究[35]。Liu 等人也跟踪了 Twitter 的电影观点，并利用这些观点，以非常精确的结果预测了电影的票房收入。在分析每一部电影中的肯定观点和否定观点时，只利用了 Opinion Parser 系统，而没有使用其他算法。

1.3 观点挖掘研究

无所不在的实际应用只是观点挖掘成为热点研究领域的原因之一。实际上，观点挖掘还是自然语言处理的一个研究主题，涵盖了许多新颖的问题，具有高度挑战性。2000 年以前，由于可利用的主观性文本很少，无论是自然语言处理还是语言学，都没有关于观点挖掘的研究。但到 2000 年以后，由于 Web 应用的激剧增长，人们很容易获得海量的主观性文本，所以观点挖掘迅速发展成为自然语言处理的一个最活跃的研究领域。另外，观点挖掘在数据挖掘、Web 挖掘和信息检索中也得到了广泛研究。由此可见，观点挖掘已从计算机科学发展到管理科学[36-42]。

1.3.1 观点挖掘的不同级别

一般地，从主观性文本中挖掘观点，有三种不同的级别：篇章级观点挖掘；句子级观点挖掘；实体与方面级观点挖掘。

1. 篇章级观点挖掘

篇章级观点挖掘的任务就是按肯定的情感极性或否定的情感极性分类评论。其特点是将整篇评论作为一个基本的信息单元。例如，已知一篇产品评论，系统决定这篇评论对于该产品总体上是表达了一个的肯定观点或否定观点。由于篇章级观点挖掘只给出一个总体观点，篇章级观点挖掘又称为篇章级情感分类。这个级别的观点挖掘总是假定每一篇评论只涉及一个单一的被评价实体。因此，篇章级观点挖掘不适合评估和比较包含有多个被评价实体的评论。

2. 句子级观点挖掘

句子级评论挖掘的任务就是确定每一个句子是否表达了一个肯定的观点或否定的观点。句子级观点挖掘与主观分类紧密相关。主观分类就是将评论中表达观点的主观句与表达事实的客观句区分开来。然而，并不只有主观句才表达情感，实际上许多客观句也隐式表达了观点。篇章级观点挖掘和句子级观点挖掘都不能准确发现人们喜欢什么或不喜欢什么。

3. 实体和方面级观点挖掘

与篇章级观点挖掘和句子级观点挖掘不同的是，实体和方面级观点挖掘并不研究语言本身的结构，即篇章、段落、句子、从句或短语，它主要研究观点本身。基本思想是，一个观点由一个情感（肯定极性或否定极性）和一个目标（观点的目标）构成。一个观点如果没有对应的目标，那么这个观点的使用就会受到限制。观点目标能帮助研究者更好理解观点挖掘问题。例如，“虽然这家酒店的服务不是那么好，但我仍喜欢这家酒店”明显有一个肯定的语气，但这并不能断定该句只表达了肯定的观点。事实上，句子强调对酒店而言，评论者持肯定的态度，但对酒店的服务，评论者持否定的态度。

在许多应用中，观点目标被描述为实体或实体的不同方面。因此，实体和方面级观点挖掘的任务就是识别实体和它的方面以及与之相关联的情感。例如，句子“这部××手机的通话质量好，但电池寿命短”评价了实体该款手机的两个方面：“通话质量”和“电池寿命”。方面“通话质量”的情感是肯定的，而方面“电池寿命”的情感是否定的。基于这个分析的级别，可以获得一个关于实体和它的方面的一个结构化观点文摘，这样就可把无结构化的文本转变成结构化数据。这种结构化观点文摘可用于各种定性分析和定量分析。

在观点挖掘的应用中，常把观点为分为规则观点和比较观点。规则观点只对一个特定的实体或实体的一个方面表达情感。例如，“××可乐味道非常好”表达了对实体“可乐”的“味道”这个方面有一个肯定的观点。比较观点则基于实体共享的某些方面对多个实体进行比较。例如，“××可乐的味道比××可乐好”基于“味道”比较了两实体，并且表示了对前者的一种偏好。

1.3.2 情感词典及相关问题

毫无疑问，表达情感的重要标志是情感词，又称观点词。它通常用于表达肯定的情感极性或否定的情感极性。例如，“好”“妙”是肯定的观点词。“坏”“差”是否定的观点词。除了单个词以外，还有一些短词和成语也能表达情感。观点挖掘离不开情感词和情感短语。情感词和情感短语的集合称之为情感词典（观点词典）。近年来，研究者们设计了大量算法编制情感词典。

尽管情感词和情感短语对于观点挖掘很重要，但仅仅使用它们进行观点挖掘还远远不够。这主要表现在如下几个方面：

（1）一个肯定的情感词或一个否定的情感词在不同的应用领域可能有相反的极性。例如，句子“这部相机真烂。”中词“烂”通常暗示着否定的情感。但是，句子

“这道菜中的牛肉很烂，味道不错。”中词“烂”表达了肯定的情感。

(2) 一个包含情感词的句子可能没有表达任何观点。这种现象在疑问句和条件句中很常见。例如，“能告诉我哪款相机好吗?”和“如果我在这家商店找到了一款好的相机，就直接买了。”这两个句子都包含了词“好”。但是，这两个句子对任何一部相机都没有表达肯定的观点或否定的观点。但并不是所有的条件句或疑问句都不表达观点。例如，“有人知道怎么修这台糟糕的打印机吗?”和“如果你在找一部好车，那就买……”。

(3) 讽刺句问题。含有情感的讽刺句和不含有情感的讽刺句都很难处理。例如，“真是一辆好车！两天不能工作了”。讽刺句通常出现在政治讨论中，关于产品和服务的消费者评论中通常没有讽刺句。

(4) 许多表达事实信息的客观句暗示了情感。例如，句子“这部洗衣机用水量大”暗示了由于这个洗衣机使用了大量的资源(水)而表达了一个否定的情感。句子“这个床垫睡两天后，中间有一个凹处”对床垫表达了一个否定的情感。显然，这两个客观句中没有任何情感词。

1.3.3 自然语言处理问题

事实上，观点挖掘是一个涉及指代消解、否定处理和词义消歧等的自然语言处理问题。因为这些问题在自然语言处理领域还没有得到完全解决，所以造成了要实现精度更高的观点挖掘有更大的难度。然而另一方面，观点挖掘又是一个高度受限制的自然语言处理问题，因为观点挖掘系统不需要完全理解每一个句子或每一篇文档的语义，仅仅只需要理解句子或文档的一些方面，例如，肯定的情感或否定的情感，这些情感所对应的目标实体或主题。从这个意义上看，观点挖掘又为自然语言研究者们提供了一个平台。近 5 年来，观点挖掘在深度和广度上都取得了显著的进步，这使得观点挖掘的核心问题越来越明朗。观点挖掘领域的早期研究主要集中在文档级别或句子级别对情感或主观进行分类，但这不能满足绝大多数现实的应用，因为实际的应用要求实体和方面级的更细粒度的分析。

1.4 垃圾观点检测

允许世界上任何地方的任何人都能在网络上自由表达看法或观点而不揭露真实身份是社会媒体的一个重要特征。这种匿名身份有效保护了观点发表者的安全。然而，匿名也会带来一定的代价，即另有企图的人或具有恶意意图的人能轻易

在系统中进行伪装，把自己以一种独立成员的印象公布于众，从而达到在网络中发布虚假观点以提升或诋毁目标产品、服务、组织或个体。这种发表虚假观点而不揭露其真实的意图的人或组织称之为垃圾观点制造者(opinion spammer)。他们的活动称之为垃圾观点滥发(opinion spamming)[43-44]。

垃圾观点滥发已成为观点挖掘研究领域的一个重要问题。除了个人在评论或论坛讨论中写虚假观点外，也有一些商业公司给他们的客户提供虚假评论。美国已报道了一些虚假评论的案例。为保证 Web 上的观点具有可利用性，检测滥发垃圾观点的行为尤为重要。与肯定观点的抽取或否定观点的抽取不同的是，垃圾观点检测不仅是一个自然语言处理问题，它还包含了人们发布行为的分析。因而它也是一个数据挖掘问题。

第2章　观点挖掘目标

本章主要阐述方面级观点挖掘的研究目标，以评论中的句子为例描述了观点挖掘的抽象定义和一些关键概念。从研究的角度看，定义给出了观点挖掘需要解决的问题，把复杂的海量无结构化自然语言文本抽象为一个结构化问题。从应用的角度看，定义给出了实际的观点挖掘系统需要解决的问题，问题之间的关联以及系统应该输出何种形式的观点文摘。

方面级观点挖掘不是以语言结构作为基本单位信息，而是直接研究观点本身，将观点视为由实体、方面、情感、观点持有者和观点发表的时间5个基本信息组成的五元组。方面级观点挖掘结果既有定量分析，也有定性分析，所以很容易形成结构化的产品属性观点文摘，为潜在的用户和组织提供决策支持。

2.1　实体与观点

在线评论观点挖掘是评论文本中表达出的观点，情感和情绪的可计算研究。下面用产品××手机的评论片段(a)介绍在线评论观点挖掘问题中一些术语的相关定义。

评论片段(a)：

①前几天我买了一款××手机；②机子很好；③屏幕很酷；④通话的声音也很清晰；⑤电池时间不长，不过对我来说这不是问题；⑥由于买手机没告诉妈妈，所以她很生气；⑦她认为这款手机太贵了，想要我去换掉。

评论片段(a)中有多个观点。第②句，第③句和第④句表达了肯定的观点，第⑤句，第⑥句和第⑦句表达了否定的观点。与此同时，所有的观点都有它针对的目标或对象。第②句的观点目标是作为一个整体的产品××手机，第③句，第④句和第⑤句中的观点目标分别是产品××手机的“屏幕”“话音”和“电池”。第⑦句中的观点目标是产品××手机的“价格”，而第⑥句中的观点目标是“我”而不是产品××手机。由此可见，观点针对的被评价对象是非常重要的。因为在实际应用中，用户可能只对一些特定的观点目标很感兴趣。针对评论中的观点源(观点持有者)，第②句、第③句、第④句和第⑤句的观点源是写评论的作者“我”，但第⑥句和第⑦句的观点源却是“妈妈”。

通过上述分析可见，观点主要由两个关键部分构成：目标 g 和针对目标的情

感 s，例如，(g,s)，其中 g 是观点所针对的任何实体或实体的方面，s 是一个肯定的，否定的或中立的情感（肯定、否定或中立表示情感 s 的极性），或是一个用数字（1～5）表示情感强度的等级。

定义1　实体：实体就是观点的目标。一个实体 e 可以是一个产品、一项服务、一个主题、一个事件、某个人、某个组织或事件。可用一个偶对 (T,W) 表示实体，其中，T 是实体的组成部分的一个层次或子组成部分的一个层次，W 是实体 e 的属性的集合，实体 e 的每一个组成部分或子组成部分也有它自己的属性。

例如，“××手机”就是一个实体。它有一个属性集合如“屏幕”“大小”“重量”等，有一个组成部分的集合如“电池”“镜头”“取景器”等。组成部分“电池”也有它自己的属性，如“电池寿命”“电池重量”等。

实体的定义本质上描述了实体的基本组成部分的层次关系。在这个层次关系图中，根结点是实体本身，如评论中所指的“××手机”。其他的所有结点是实体的组成部分和子组成部分。评论者可以对层次结构中的任意一个结点或结点的任意一个属性给出自己的观点。

一个有任意层次结构的实体需要用一个嵌套的关系表达，但在实际应用中却十分复杂。其主要原因是利用自然语言处理技术在不同层次上识别一个实体的组成部分或它的属性极端困难。大部分应用并不需要进行复杂的分析，因此常常把实体的层次图的层数简化为两层，根结点表示实体本身，词“aspect”表示实体的组成部分或属性。在观点挖掘的实际应用中将使用这种典型的两层层次结构。

分析了观点目标后，下面给出观点的定义。

定义2　观点：观点是一个五元组 $(e_i,_{ij},s_{ijkl},h_k,t_l)$，其中 e_i 是实体名，$_{ij}$ 是实体 e_i 的一个方面，s_{ijkl} 是实体 e_i 的方面 $_{ij}$ 的情感，h_k 是观点持有者，t_l 是观点持有者表达观点的时间。

在上面的定义中，利用下标有目的地强调了五元组中五个信息之间的相互对应关系。例如，s_{ijkl} 表示观点持有者 h_k 在时间 t_l 对被评价实体 e_i 的 $_{ij}$ 方面发表的观点。五元组中的这五部分信息缺一不可。通常情况下，遗失任何一部分信息都有问题。例如，如果五元组中缺少了时间信息，就不能针对时间的变化情况来分析观点的变化情况。这种现象在实际中很常见，而且也十分重要。因为一个两年前的观点可能和昨天的观点是不相同的。如果五元组中缺少了观点持有者信息也会有问题。例如，句子“⑤电池时间不长，不过对我来说这不是问题。⑥由于买手机没告诉妈妈，所以她很生气。⑦她认为这款手机太贵了，想要我去换掉。”中有两个观点持有者“我”和“妈妈”。识别不同的观点持有者对于实际应用十分重要。

需要注意的是，由于观点的语义可能十分复杂，上述观点的定义并没有覆盖观