文章编号：1007-1423(2017)05-0048-04 DOI：10.3969/j.issn.1007-1423.2017.05.012

基于微博电影评论的情感分析研究

钱慎一，杨铁松

（郑州轻工业学院计算机与通信工程学院，郑州 450001）

摘要：

近几年，数据挖掘分析成为一个热点研究的课题，其中的文本研究分析更成为热中之热，而微博电影评论成为一种新的电影设计模式，也就自然成为研究对象。主要从数据采集、特征提取、情感词典构建及情感计算几个方面进行研究，提出基于句法分析算法，并进行必要研究，进一步提高微博电影评论情感倾向分析的正确率。

关键词：

数据挖掘；情感分析；特征提取

0 引言

近些年，随着互联网的飞速发展，网络技术日新月异的变革，人们的各种思想也就充斥在各种网络论坛之上。微博作为一种新兴的社交平台，凭借着快、短、灵活的特点，成为了最火热的用户发布、传播、共享信息的平台。随着用户量的增涨，微博对社会舆论的影响日益增加，并潜移默化的改变着人们的生活方式。微博里海量的文本信息，很多都有用户的参与，存在着大量的有价值信息。微博电影评论就是其中一类，用户借助微博平台，表达着自己的观点，成为了一种新的电影社交模式。

微博电影评论与传统的网络电影评论相比，信息量更大，及时性更强，获得人们的关注度更高。因此对微博电影评论的情感分析研究意义重大，不仅可以引导观众的观影决策，而且可以使制片商调整他们的营销策略。微博电影评论挖掘是在一个特定的领域，所有它更有针对性，并且特征丰富，除了要关注电影品质本身，还要关注演员、编剧、导演、制作人、出品公司等。这些都是电影评论的特征，相比其他产品可能更具挑战性。目前，国内外对于电影评论的研究相对较少，Chaovalit P 等人[1]分别采用基于机器学习和语义倾向两种方法进行研究，Zhuang L 等人[2]通过提取特征词对的方法等。本文主要是基于依存句法规则方法对微博电影品论情感分析进行研究。

1 微博电影情感分析框架

首先我们先来明确情感的定义[3]，情感就是人们情绪上的变化，例如喜怒哀乐。这样我们就可以把情感划分成正倾向、负倾向和中立态度几类。正倾向的态度就是积极的、乐观的、使人向上的态度。负倾向就正好相反，使人悲观、愤怒。像生气，郁闷等是属于负倾向态度这类的。中立态度是指客观的去分析，并没有一自己的好恶去评判。

情感分析倾向的计算可分为以下几步：(1)数据预处理；(2)聚类分析(特征提取)；(3)情感词典的建立；(4)情感计算。如图 1 所示：

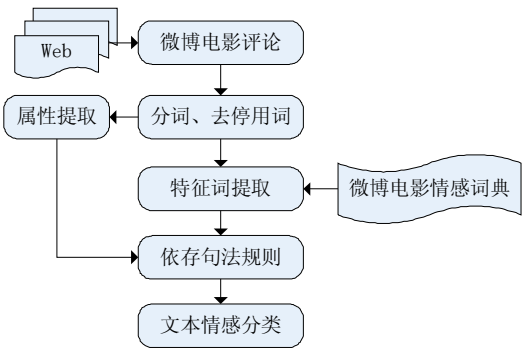

图 1 微博电影情感分析框架

首先就是行进数据预处理阶段，在这个阶段主要的工作是对需要分析的文本进行爬取收集，再就是进

行简单的分词处理、去除停用词、词频计算[4]等操作，把文本储存到准备好的数据库中，以备后续使用。接着是特征提取，情感词的抽取是短文本情感分析的重要部分。在聚类分类的过程中，词语是基本特征，电脑若是想要理解人类的语言时，一般经过两步的加工量化，第一步是特征提取，确保主要的部分被筛选出来；第二步是特征权重计算，将文本量化，方便理解并计算。其中的特征提取，作用是为了降维，降低复杂度，去除噪声，从而增加分类精度。再来就是情感词典的建立，虽说眼下的情感词典很多，却还没有一部完整且通用的情感词典。在国外，目前较为流行且成熟的情感词典资源有 GI 词典[5]。该词典给出的每个词条都相当全面。如褒义词、贬义词、反义词等。还有 LIWC 词典[6]，该词典的类别体系和 CI 词典大致相同 SentiWordNet 词典[7]，该词典是基于 WordNet 中的词条进行情感分类的，国内的情感分析研究起步不久，当前能应用的词典资源自然有限。大概有知网的情感词典[8]、台湾大学的情感极性词典、还有大连理工大学信息检索样就是整理标注的情感词汇本题库等。目前比较常用的方法是，先对大规模的词典库进行分析研究，对常用的词语进行标注，选为基础词。然后具有针对性的获得新的情感词，从而扩展情感词典。如下图：

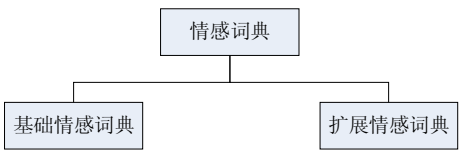

图 2　扩展词典

最后是情感倾向分析，本文从句子的结构角度出发，采用基于依存句法的情感分析方法，对句子中的短语进行识别抽取、从细粒度的角度对基础情感词和极性短语进行量化计算，再对句子进行特定句式识别消除它们对句子极性的影响，进而以量化的文本极性值完成句子级细粒度的情感计算。

2　依存句法分析

依存语法[9](Dependency Parsing，DP)是研究句子内各个成分之间的句法依存关系来揭示其句法结构。将汉语句子从一个线性序列转换成一棵完整的依存分析树。它的表达形式十分简洁，无需额外添加语法符号，所以相对来说容易理解。由于句法分析是深入语言内部结构进行分析的，其分析结果能够强有力的支持句子情感分析。依存句法分析的目的是构建输入句子的句法结构树。

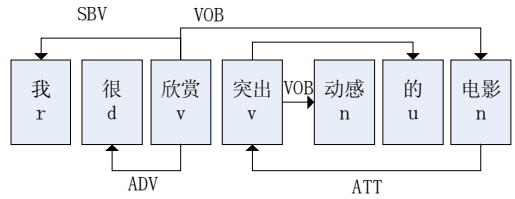

图 3　句法结构树实例

图中我们可以看出由“我”与“欣赏”、“很”与“欣赏”、“突出”与“的”、“欣赏”与“电影”、“突出”与“电影”等组成的短语。并且中间都有一条带有箭头的有标记的弧线。每条弧线清晰的给出了每个词语的依存关系。

(1)依存关系对的表示

微博句子的情感分析关键在于对情感词依存关系的选取上，对于依存关系树上存在的两个节点 x 和 y，x 为子节点，y 为父节点，通过分析依存关系树，我们可以找到两者在书中的节点 id，从而给出的依存关系对的表达方式为：

$$RelationPair=<id_x,x,Word_x,id_y,y,Word_y,relation,rawScore>$$

从上图所示的例子中，抽取“很欣赏”和“突出动感”的依存关系对。表示如下：

<1，很，d，2，欣赏，v，ADV，0.8>

<3，突出，v，4，动感，n，VOB，0.4>

(2)依存关系的距离

依存距离这里是指两个存在依存关系的词汇之间的线性距离，也就是两个节点次序之差的绝对值大小。例如下面的两个句子：a:“这部电影不太好看”。 B:“这部电影太不好看”。我们可以看出，虽然只有两个字的次序不一样，但这两句话的感情程度是有很大的差异的。对两句话进行句法分析，可得到的关系对如下：

a:<2，不，d，3，好看，a，ADV，0.8>

<1，太，d，3，好看，a，ADV，0.8>

b:<2，太，d，3，好看，a，ADV，0.8>

<1，不，d，3，好看，a，ADV，0.8>

可以看出 a 中否定词“不”与“好看”的依存距离是 2，b 中的否定词“不”与“好看”的依存距离是 1，由此可知，依存距离越小，感情极性越强。由依存关系定义，若

将依存距离看成是主导词和从属词在句子中距离的差，我们不分正负，只求句子距离上的差别，所以采用去计算的绝对值：

$$Distance(Word_x,Word_y)=|id_x-id_y|$$

其中 $Word_x$、$Word_y$ 表示遍历依存句法树得到的节点 id，也就是 $Word_x$、$Word_y$ 的词号。

(3)情感短语的计算

在进行句子级的情感计算时，主要对句子中出现的情感词构成的依存关系进行分析。先对文本分分句，再进行分词、词性标注；继而通过情感词典来判断是否有情感值，若有则将之添加到情感词类表，如果有否定词或程度副词，则根据扩展的情感词典进行相应的处理。最后用句子中情感词和情感短语的情感强度平均值作为整个文本的感情倾向值。

情感短语计算：

$$Value=degree(Word1)*polarity(Word2)/Distance(Word1,Word2)$$

其中 Word1、Word2 分别为副词和情感词，polarity 表示情感词，degree 表示情感程度词。

(4)句子级情感计算

有了依存关系对的情感极性，再加上句子中每个情感词，并将其情感倾向值归一求和。就得到了句子级的情感计算[10]公式：

$$sentenceValue=\frac{\sum_{1}^{n}\beta_i\cdot value_i}{|\sum_{1}^{n}\beta_i\cdot value_i|+05n}$$

其中 β_i 为情感词的权值，n 为情感词、情感短语总数。

这样就计算出了句子的情感极性，先给出依存句法的情感计算方法，进行深入的讨论，再分析了影响微博情感的词语及短语情感倾向，最终完成了句子级的情感计算。

3 实验

现如今，针对电影的情感分类方法有很多，其中基于协同训练的半监督情感分类方法相对高效。那就用本文的算法与之相比较。

(1)先进行数据采集，从新浪网进行评论采集，有30000 条微博电影评论数据。并进行人工标注，把文本分类成褒义、贬义和中性 3 种。

(2)评价方法

对采集的文本进行情感倾向分析，将自动分析的结果和人工标注的对比。测试结果越接近人工标注，则说明实验越正确。

评价指标采用最被接受的，评测时使用准确率(precision)和召回率(recall)，并用综合评分指标 F 来衡量正确率。

准确率(precision)=分析正确的文本数/总的文本数

召回率(recall)=分析正确的文本数/总的正确的文本数

$$F=2PR/(P+R)$$

其中 P 表示正确率，R 表示召回率。

(3)实验设计及分析

基于依存句法的算法，通过系统分析这 30000 条文本，在不同阈值下的 F 值的曲线变化如下图：

表 1

阈值	正面 F 值	负面 F 值
0	0.9248	0.9213
0.05	0.9321	0.9335
0.1	0.9264	0.9478
0.15	0.9281	0.9505
0.2	0.9012	0.9219
0.25	0.8974	0.9081
0.3	0.8861	0.8945
0.35	0.8392	0.8676

由上表可知，当阈值达到 0.15 时，情感分析结果达到最优。取阈值 0.15 进行实验，与协同训练算法[11]进行比较。

表 2

	协同训练算法		本文算法	
	正面	负面	正面	负面
正确率	0.9471	0.9035	0.9423	0.9398
召回率	0.9183	0.9247	0.9507	0.9192
F 值	0.9285	0.9164	0.9496	0.9323

由上表可以看出本文的算法的各方面都是高于协同训练算法的，结果证明本文的实验结果是达到预期效果的。

4 结语

本文针对微博电影评论进行了情感分类研究，提出了基于依存句法规则对微博电影评论分类的方法。

并进行了实验和比对,取得了一些效果,但仍存在许多的不足。比如,情感词典的多维构造,以及有效地解决特征稀疏问题等。所有还需要我们更加努力地进行下一步的研究和改进。

参考文献:

[1]Chaovalit P, Zhou L. Movie Review Mining: A Comparison Between Supervised and Unsupervised Classification Approaches[C]. System Sciences,2005:112-148.

[2]Zhuang L, Jing F, Zhu X Y. Movie Review Mining and Summarization[C]. 2006: 99-132.

[3]庞观松,蒋盛益. 文本自动分类技术研究综述[J]. 情报理论与实践,2012,35 (1):96-123.

[4]Hung C, Lin H K. Using Objective Words in SentiWordNet to Improve Word-of-Mouth Sentiment Classification[J]. IEEE Intelligent Systems,2013,28(2):147-154.

[5]Philip J. Stone,Dexter C Dunphy,Marshall S. Smith, Daniel M. Ogilvie. The General Inquirer: A Computer Approach to Content Analysis. MIT Press,1966.

[6]Pennebaker, J.W., Booth, R.J., & Francis, M.E. Linguistic Inquiry and Word Count: LIWC 2007. Austin, TX. 2007.

[7]Stefano Baccianella, Andrea Esuli,Fabrizio Sebastiani. An Enhanced Lexical Resource for Sentiment Analysis and Opinion Mining. LREC. 2010.

[8]HowNet[R/OL]. HowNet's Home Page. http //www.keenage.com. 2011,12,10.

[9]刘海涛. 依存语法的理论与实践[M]. 北京:科学出版社,2009.

[10]施寒潇. 细粒度情感分析研究[D]. 苏州大学,2013.

[11]Blum A,Mitchell T. Combining Labeled and Unlabeled Data with Co-Training[C]. Proceedings of the Eleventh Annual Conference on Computational Learning Theory. ACM,1998: 92-100.

作者简介:

钱慎一(1975-),男,江苏扬州人,硕士,副教授,硕士生导师,研究方向为数据库与信息集成、计算机应用技术

杨铁松,男,河南商丘人,硕士,研究方向为数据挖掘为大数据分析

收稿日期:2016-11-29 修稿日期:2017-02-12

Research on Emotional Analysis Based on Micro-Blog Film Criticism

QIAN Shen-yi,YANG Tie-song

(College of Computer and Communication Engineering, Zhengzhou University of Light Industry, Zhengzhou 450001)

Abstract:

In recent years, data mining analysis has become a hot research topic, in which the text research and analysis has become a hot, micro-blog film commentary has become a new film design pattern. Mainly studies the data acquisition, feature extraction, emotion dictionary construction and emotion computation, proposes a syntax analysis algorithm and makes necessary research. And further improves the micro-blog movie comments emotional analysis of the correct rate.

Keywords:

Data Mining; Emotion Analysis; Feature Extraction