

基于 CRF 和 HITS 算法的特征情感对提取

唐 莉, 刘 臣

(上海理工大学, 上海 200093)

摘 要:作为情感分析的子任务之一,特征级情感分析备受关注。条件随机场(CRF)是情感分析任务的常用方法之一,特别是对于产品特征的提取,但是针对特征词与情感词之间的长依存问题难以解决。针对该问题,提出一种基于 CRF 和 HITS 算法的两阶段方法来提取(产品特征-情感词)对。使用 CRF 并利用词、词性、依存句法关系三种文本特征来对产品评论中的评价特征和情感词进行提取,并利用已提取的特征和情感词分别作为权威节点和枢纽节点来构建特征情感词二分网。使用一种称为 MHITS 的扩展 HITS 算法在二分网上计算并对(产品特征-情感词)对进行排序。实验使用了京东平台上三种不同类型产品的评论数据,并与基准方法进行比较,结果表明该模型在准确率、召回率和 F_1 值上表现更平均。

关键词:条件随机场;情感分析;依存句法分析;二分网;扩展 HITS 算法

中图分类号:TP391.1

文献标识码:A

文章编号:1673-629X(2019)07-0071-05

doi:10.3969/j.issn.1673-629X.2019.07.014

Extraction of Feature and Sentiment Word Pair Based on Conditional Random Fields and HITS Algorithm

TANG Li, LIU Chen

(University of Shanghai for Science and Technology, Shanghai 200093, China)

Abstract: As one of the sub-tasks of sentiment analysis, feature-level sentiment analysis has attracted much more attention. In the past, conditional random field (CRF) was one of the commonly used methods for sentiment analysis tasks, especially for feature extraction. However, it was difficult for this method to solve the long-range dependence problem between feature words and sentiment words. Therefore, a two-stage method based on CRF and HITS algorithm is proposed for extracting the pair of product feature-sentiment word. CRF and the three kinds of text features, including word, part of speech and dependency parsing, are utilized to extract features and sentiment words. These features and sentiment words are taken as the authority node and the hub node respectively to constitute a bipartite feature-sentiment relation network. An extended HITS algorithm called MHITS is applied to calculate and sort the features and sentiment word pairs on a bipartite network. The experiment uses the reviews over three different types of products on Jingdong platform. Compared with the benchmark method, the results show that the model performs more evenly in terms of precision, recall and F_1 measure.

Key words: conditional random field; sentiment analysis; dependency parsing; bipartite network; extended HITS algorithm

0 引 言

随着电子商务的迅猛发展,网络购物用户数量不断增加。中国作为全球规模最大的网络零售市场,2017 年 12 月中国网络购物用户规模已达 5.33 亿,由此产生的网络评论数据呈几何式增长。庞大的评论数据带来的信息超载问题日益严重,海量的评论信息的分析和提炼成为需要解决的首要问题。

文本情感分析采用计算机自动分析在线评论中表达的观点或情感,具有广泛的应用前景,已成为近年来的研究热点。

文本情感分析可以被大致分成特征级、短语级^[1]、句子级、文档级^[2]。其中特征级情感分析是对产品属性及其对应情感的提取和分析。它一般被分为有监督和无监督学习方法。

无监督学习方法不需要标记数据,但是依赖于利用启发式程序和规则来找到隐藏的结构^[3]。常见的无监督学习有频繁模式挖掘^[4-5]、语义规则挖掘^[6]、话题模型^[7]等。有监督学习则需要标记数据来训练和标注。常见的有监督学习方法有隐马尔可夫模型和条件随机场。其中,条件随机场能够使用局部信息进行信

收稿日期:2018-08-10

修回日期:2018-12-18

网络出版时间:2019-03-21

基金项目:国家自然科学基金(71401107)

作者简介:唐 莉(1994-),女,硕士研究生,研究方向为自然语言处理;刘 臣,副教授,研究方向为自然语言处理。

网络出版地址: <http://kns.cnki.net/kcms/detail/61.1450.TP.20190321.0904.004.html>

息提取,但对于特征词与情感词之间的依赖关系这种信息并不能有效利用。对此,文中首先使用融合“SBV”,“ATT”,“COO”三种依存句法关系的条件随机场对句子中的特征和情感词进行抽取,然后利用已抽取的特征情感词构成一个二分网,最后使用一种基于点互信息的 HITS 算法来抽取特征和情感词对。

1 相关工作

1.1 无监督学习

基于高频名词通常是真实的特征的想法,Hu 等^[4]提出使用关联规则挖掘找到高频名词作为候选特征,然后使用剪枝挑选出的结果作为特征。基于这个研究,Popescu 等^[5]利用名词与产品类别之间的点互信息(PMI)来抽取产品特征。

除了使用频率的方法,利用语义规则提取特征也是一类重要的方法。Zhuang 等^[8]使用特征情感词之间的出现次数最多的依赖关系抽取电影评论数据集中的特征。Qiu 等^[9]考虑到句子中词间存在着直接关系和间接关系。直接关系就是两个词之间直接依赖或者都同时依赖同一个词。间接关系则是两个词之间都非直接依赖第三个词。因此,他们提出了基于直接关系的双向传播算法并设计一些规则来抽取语料库中特征和情感词。

此外,还有通过对候选特征排序来提取特征的方法。Eirinaki 等^[10]认为一个名词与越多的形容词进行搭配,那么这个名词越有可能是一个特征。基于这种想法,他们提出了 HAC 算法,通过对与名词搭配的形容词进行计数来确定名词的分数,搭配越多形容词的名词分数越高。Yan 等^[11]和 Zhang 等^[12]分别利用 PageRank 算法和 HITS 算法根据候选特征和形容词之间的搭配关系以及共现频率对候选特征进行排序,从而识别特征词。

1.2 有监督学习

条件随机场(conditional random field, CRF)是由 John Lafferty 于 2001 年提出,是一种流行的用于结构化预测的概率方法,广泛用于各领域。该模型结构类似于有条件训练的隐马尔可夫模型,并且具有有效的推理算法^[13]。

在近年来情感分析任务中,CRFs 模型表现优异。在使用条件随机场时,最常使用的结构是线性链 CRFs。为了解决长范围依赖关系问题和能够在同一序列上执行多个级联标记任务,Charles 等^[14]提出了动态条件随机场(DCRF)。它是线性链 CRFs 的泛化,其将每个时间片都包含一组状态变量和边,并且参数跨时间片绑定。Niklas 等^[15]为了解决模型的单一领域

和跨领域的特征抽取问题,将词、词性、短依存关系、词距以及观点词作为输入特征。

2 基于 CRF 的特征和情感提取

条件随机场是给定一组输入随机变量条件下另一组输出随机变量的条件概率分布模型,其特点是假设输出随机变量构成马尔可夫随机场^[16]。假设 $X = (X_1, X_2, \dots, X_n)$ 和 $Y = (Y_1, Y_2, \dots, Y_n)$ 都是随机变量, X 是需要标注的观测序列, Y 是 X 对应的标记序列。 $P(Y|X)$ 是条件随机场所要构建的条件概率模型。无向图 $G = (V, E)$ 表示由 Y 构成的马尔可夫随机场。

即:

$$P(Y_v | X, Y_w, w \neq v) = P(Y_v | X, Y_w, w \sim v) \quad (1)$$

对任意节点 v 成立,则 $P(Y|X)$ 为条件随机场。式 1 中, $w \sim v$ 表示节点 v 与节点 w 是无向图 $G = (V, E)$ 中的邻居节点。 Y_v 和 Y_w 是节点 v 与 w 对应的随机变量^[13]。

文中主要使用线性链条件随机场。线性链条件随机场中,随机变量 $X = (X_1, X_2, \dots, X_n)$ 和 $Y = (Y_1, Y_2, \dots, Y_n)$ 都是线性链表示的随机变量序列,它可以被表示为:

$$P(Y_i | X, Y_1, \dots, Y_{i-1}, Y_{i+1}, \dots, Y_n) = P(Y_i | X, Y_{i-1}, Y_{i+1}), i = 1, 2, \dots, n \quad (2)$$

在情感分析任务中,特征情感词的提取可被看作作为一个序列标注问题,因此可采用线性链 CRF 模型。下文中描述用作 CRF 模型输入的特征。

2.1 分词与词性标注

词与词性是在自然语言处理任务中经常被使用到的特征。这两个特征是将句子中当前词与该词的词性作为特征,两者对于特征和情感词的提取有相当大的影响。

对于句子的分词和词性标注都是通过哈尔滨工业大学开发的 LTP (HIT - SCIR, <http://www.ltp-cloud.com/>)。

2.2 依存句法关系

另外,文中使用了依存句法关系作为特征。依存句法关系的提取同样使用的是 LTP。保留了三种特征、情感词间经常存在的依存句法关系,分别是定中关系(ATT)、平行关系(COO)和主谓关系(SBV)^[17]。

2.3 特征模板

文中采用 CRF++0.58 工具包来执行模型的训练和标注。使用 CRF++工具包需要设计特征模板。特征模板将从特征集中提取特征并添加到模型中。文中基于上述语言相关特征设计了特征模板。

表 1 展示了标注特征和情感词使用的模板。

表 1 标注特征和情感词使用的特征模板

序号	特征模板	说明
1	$w[t-2], w[t-1], w[t], w[t+1], w[t+2]$	词本身
2	$w[t-1] w[t], w[t] w[t+1]$	相邻的两个词
3	$pos[t-2], pos[t-1], pos[t], pos[t+1], pos[t+2]$	词性标注本身
4	$pos[t-2] pos[t-1], pos[t-1] pos[t], pos[t] pos[t+1], pos[t+1] pos[t+2]$	相邻的两个词性标注
5	$pos[t-2] pos[t-1] pos[t], pos[t-1] pos[t] pos[t+1], pos[t] pos[t+1] pos[t+2]$	相邻的三个词性标注
6	$dp[t-2], dp[t-1], dp[t], dp[t+1], dp[t+2]$	依存句法关系本身
7	$dp[t-2] dp[t-1], dp[t-1] dp[t], dp[t] dp[t+1], dp[t+1] dp[t+2]$	相邻的两个依存句法关系

其中, t 表示当前词的标记, w 表示词本身, pos 表示对应词的词性, dp 表示对应词的依存关系标注。

3 评价特征和情感词对抽取

使用依存关系是提取特征-情感词对的一种常用方法,但是使用依存关系提取的特征-情感词对的效果依赖于句子语法表达的正确性以及依存关系的人工选择。为了避免以上问题,文中利用 MHITS 算法^[17]来提取特征情感词对,其考虑了特征和情感词间的共现频率,以及特征与情感各自的重要性。

为了使用 MHITS 算法对特征-情感词对的值进行计算,首先需要构建特征-情感词二分网络。上阶段使用 CRF 提取的特征词和情感词分别作为网络中的权威节点 f_i 和枢纽节点 s_j 。在测试数据集中,同一句话中出现的任意两个特征词和情感词之间都存在着一条边 e_{ij} 。边 e_{ij} 的权重 w_{ij} 、权威节点的值 $p(f_i)$ 和枢纽节点的值 $p(s_j)$ 都使用 MHITS 算法进行计算。MHITS 算法具体如下:

初始步:对于在全部数据集中特征词与情感词在同句话中出现过的词对的权重为其共现频次。其余网络中边的权重 w_{ij} 的初始值为 1。

迭代过程:
权威(特征)节点:每个权威节点都更新为与其相连的枢纽节点的边权之和。

$$p(f_i(k)) = \frac{\sum_{j \in T} w_{ij}(k-1)}{\sum_{i \in F} \sum_{j \in S} w_{ij}(k-1)} \tag{3}$$

枢纽(情感)节点:每个枢纽节点更新为与其相连的权威节点的边权之和。

$$p(s_j(k)) = \frac{\sum_{j \in U} w_{ji}(k-1)}{\sum_{i \in F} \sum_{j \in S} w_{ij}(k-1)} \tag{4}$$

边权:边连接的权威节点和枢纽节点了点互信息。

$$w_{ji}(k) = \log \frac{w_{ji}(k-1)^2}{p(f_i(k))p(s_j(k))} \tag{5}$$

其中, k 表示当前的迭代次数; T 表示权威节点 f_i 连接的枢纽节点的集合; U 表示枢纽节点 s_j 连接的权

威节点的集合; F 表示所有权威节点的集合; S 表示所有枢纽节点的集合。

第一轮迭代中网络的边权是根据特征词与情感词之间共现频率计算两者之间的点互信息,然后更新特征和情感节点的值,之后的每一轮都是根据上一轮的值计算边权和节点的值。

算法在不断迭代中收敛,最终结果会趋于平稳。根据 MHITS 算法计算的边权的值对(特征-情感)对进行排序,边权值大于某个阈值的节点对会被保存在最终列表中。

4 实验设置与结果分析

为了验证算法的有效性,使用京东商城上的评论数据进行实验。

4.1 实验数据

实验数据是 Liu 等^[17]使用的评论数据集集中的三种商品,分别是华为手机、洗面奶以及羽毛球拍,见表 2。

表 2 数据集

产品名称	评论数量	评论字数	标记词对数
洗面奶	1 006	23 845	94
羽毛球拍	1 486	37 481	117
华为手机	1 359	35 149	276

4.2 评价指标

文中使用精确率(P)、召回率(R)、 F 指标(F_1)去评估模型的质量。精确率计算的是模型提取的所有特征或情感中正确的特征-情感词对的比例。召回率计算的是模型提取的正确的特征-情感对占评论中所有特征-情感词对的比例。 F_1 指标是精确率和召回率的调和平均值。

其计算公式分别为:

$$P = \frac{TP}{TP + FP} \tag{6}$$

$$R = \frac{TP}{TP + FN} \tag{7}$$

$$F_1 = \frac{2 \times P \times R}{P + R} \tag{8}$$

其中,TP 表示真正例,即模型提取出的正确的特征-情感词对;FP 表示假正例,即模型提取出来的错误的特征-情感词对;FN 表示假负例,是模型未能提取出来的评论中的特征-情感词对。

4.3 实验结果

文中首先使用融合词、词性和依存句法关系的线性链 CRF 提取特征和情感词。特征模板使用表 1 中序号 1-7 的模板。然后在 CRF 标注的情感词和特感词构建特征-情感网络并运行 MHITS 算法。对比的方法分别是:

依存关系:对评论进行句法分析后提取符合“SBV”,“ATT”,“COO”关系的特征情感对。具体的提取规则有 5 种,可分为“SBV”,“ATT”两个大类。具体提取的规则见表 3。

表 3 提取特征情感词对的规则

关系	具体规则	解释
SBV	Adj+Noun	符合 SBV 且形容词修饰名词
	COO+Noun	符合 SBV 且多个形容词共同修饰名词
	Noun+Adj	符合 ATT 且形容词修饰名词
ATT	COO+Adj	符合 ATT 且单个形容词修饰多个名词
	COO+Noun	符合 ATT 且多个形容词修饰一个名词

依存+CRF:在使用 CRF 标注特征情感词后利用“SBV”,“ATT”,“COO”关系抽取特征-情感词对。

HITS+CRF:使用 CRF 标注的特征和情感词构建二分网,然后使用 Zhang 等提出的基于二分网的 HITS 算法对特征和情感词组成的词对进行排序^[12]。

MHITS+CRF:使用 CRF 标注的特征和情感词构建二分网,然后使用 MHITS 算法对特征和情感词组成的词对进行排序。

实验使用排序的最佳阈值为在评论上的最高 F 值。最终在三个数据集上的结果如表 4 所示。

表 4 在华为手机、羽毛球和洗面奶数据集上的比较结果

算法	华为手机			羽毛球			洗面奶		
	P	R	F ₁	P	R	F ₁	P	R	F ₁
依存关系	0.24	0.40	0.30	0.16	0.67	0.25	0.23	0.72	0.36
依存+CRF	0.64	0.30	0.41	0.76	0.32	0.45	0.78	0.31	0.44
HITS+CRF	0.70	0.36	0.48	0.62	0.28	0.39	0.66	0.27	0.38
MHITS+CRF	0.57	0.41	0.48	0.5	0.39	0.45	0.42	0.34	0.38

从表 4 中可以看出,使用依存+CRF 和 HITS+CRF 提取特征-情感词对具有较高的准确性,但召回率比较低。而仅使用依存关系则具有较高的召回率,准确率非常低。MHITS+CRF 算法在准确率和召回率上都表现中等。

从表 5 可以看出,MHITS+CRF 的平均 F₁ 值相对较高。

表 5 在全部数据集上的平均值

算法	P	R	F ₁
依存关系	0.21	0.6	0.3
依存+CRF	0.73	0.31	0.43
HITS+CRF	0.66	0.3	0.42
MHITS+CRF	0.5	0.38	0.44

通过比较依存关系和依存+CRF 两种方法可以看出,使用 CRF 提取特征和情感后再使用依存关系提取特征情感对能够显著提升提取特征和情感对的效果。MHITS+CRF 与依存+CRF、HITS+CRF 方法比较之下,准确率较低,但是召回率较高。另外,由于 MHITS 算法考虑词对的共现频率,最终 F₁ 值稍高于使用 HITS 算法排序的结果。

5 结束语

文中提出了一种两阶段的特征-情感词对提取方法。首先使用 CRF 模型,融合词、词性、依存句法关系三种文本特征对评论文本中的特征和情感词进行识别,然后使用 MHITS 算法对特征-情感词对进行提取。实验结果表明,相较于基准方法,该方法在特征-情感词对的提取上取得了一定的效果。但该方法也存在一定的不足:使用的是有监督的 CRF 模型,需要预先标注大量的数据;会受到数据量大小的影响,特别是对于数据量小和特征与情感词少的评论效果较差。因此,下一步的改进方向是减少数据量大小对方法的影响。

参考文献:

[1] 乌达巴拉,汪增福. 一种扩展式 CRFs 的短语情感倾向性分析方法研究[J]. 中文信息学报,2015,29(1):155-162.

[2] LIU B. Sentiment analysis and opinion mining[M]. USA: Morgan & Claypool Publishers,2012.

[3] AZIZ N A A,MAAROF M A,ZAINAL A,et al. A review of the opinion target extraction using sequence labeling algorithms based on features combinations[J]. Journal of Internet Computing & Services,2016,17(5):111-119.

[4] HU Mingqing, LIU Bing. Mining and summarizing customer reviews[C]//Proceedings of the tenth ACM SIGKDD international conference on knowledge discovery and data mining. Seattle, WA, USA:ACM,2004:168-177.

[5] POPESCU A M, ETZIONI O. Extracting product features and opinions from reviews[C]//Proceedings of human language technology conference and conference on empirical methods in natural language processing. Stroudsburg, PA, USA: Association for Computational Linguistics, 2007: 9 -

- 28.
- [6] WU Yuanbin, ZHANG Qi, HUANG Xuanjing, et al. Phrase dependency parsing for opinion mining[C]//Proceedings of the 2009 conference on empirical methods in natural language processing. Singapore: Association for Computational Linguistics, 2009: 1533–1541.
- [7] ZHAI Zhongwu, LIU Bing, XU Hua, et al. Constrained LDA for grouping product features in opinion mining[C]//Proceedings of the 15th Pacific-Asia conference on advances in knowledge discovery and data mining. Shenzhen, China: Springer, 2011: 448–459.
- [8] ZHUANG Li, JING Feng, ZHU Xiaoyan. Movie review mining and summarization[C]//Proceedings of the 15th ACM international conference on information and knowledge management. Arlington, Virginia, USA: ACM, 2006: 43–50.
- [9] QIU Guang, LIU Bing, BU Jiajun, et al. Expanding domain sentiment lexicon through double propagation[C]//Proceedings of 21st international joint conference on artificial intelligence. Pasadena, California, USA: Kaufmann Publishers Inc, 2009: 1199–1204.
- [10] EIRINAKI M, PISAL S, SINGH J. Feature-based opinion mining and ranking[J]. Journal of Computer & System Sciences, 2012, 78(4): 1175–1184.
- [11] YAN Zhijun, XING Meiming, ZHANG Dongsong, et al. EX-PRS: an extended pagerank method for product feature extraction from online consumer reviews[J]. Information & Management, 2015, 52(7): 850–858.
- [12] ZHANG Lei, LIU Bing, LIM S H, et al. Extracting and ranking product features in opinion documents[C]//International conference on computational linguistics. Beijing, China: Association for Computational Linguistics, 2010: 1462–1470.
- [13] LAFFERTY J D, MCCALLUM A, PEREIRA F C N. Conditional random fields: probabilistic models for segmenting and labeling sequence data[C]//Eighteenth international conference on machine learning. [s. l.]: Morgan Kaufmann Publishers Inc, 2001: 282–289.
- [14] SUTTON C, ROHANIMANESH K, MCCALLUM A. Dynamic conditional random fields: factorized probabilistic models for labeling and segmenting sequence data[J]. Journal of Machine Learning Research, 2004, 8(2): 693–723.
- [15] JAKOB N, GUREVYCH I. Extracting opinion targets in a single- and cross-domain setting with conditional random fields[C]//Proceedings of the 2010 conference on empirical methods in natural language processing. Cambridge, Massachusetts: Association for Computational Linguistics, 2011: 1035–1045.
- [16] 李 航. 统计学习方法[M]. 北京: 清华大学出版社, 2012.
- [17] LIU Chen, TANG Li, SHAN Wei. An extended HITS algorithm on bipartite network for features extraction of online customer reviews[J]. Sustainability, 2018, 10(5): 1425–1440.



CNCC

2019

中国计算机大会

China National Computer Congress 2019

10.17~19 苏州金鸡湖国际会议中心

智能+

——引领社会发展

AI+ leading the development of society

010-62600336

cncc@ccf.org.cn

cncc.ccf.org.cn

