DOI: 10.3969/j.issn.1001-0505.2016.03.010

二次剪枝算法在评论特征提取中的应用

吴含前[1] 周立凤[1] 谢 珏[2]

([1] 东南大学计算机科学与工程学院,南京 211189)
([2] 东南大学蒙纳士大学苏州联合研究生院,苏州 215123)

摘要: 针对序列模式挖掘(GSP)算法在中文产品评论特征提取中准确率不够高的问题,提出了一种二次剪枝算法,即利用GSP算法产生候选特征集,然后采用词对共现度作为阈值对其进行进一步筛选,从而达到提高准确率的目的.利用定制化的爬虫工具从京东网站上抓取摄像头产品的中文评论,选取其中1 000条作为试验数据,采用分词工具ICTCLAS对评论进行分词和数据预处理,并将所提算法与GSP算法、交叉语言模型(CLM)和似然比检验(LRT)进行对比试验.结果表明,利用所提算法获得的中文产品评论特征提取准确率达到76.37%,较GSP算法、CLM和LRT的准确率分别提高2.94%,5.77%和7.57%.

关键词: 特征提取;二次剪枝;词对共现度;似然比检验;交叉语言模型

中图分类号: TP315.69 **文献标志码:** A **文章编号:** 1001-0505(2016)03-0513-05

Application of secondary pruning algorithm in commentary feature extraction

Wu Hanqian[1] Zhou Lifeng[1] Xie Jue[2]

([1] School of Computer Science and Engineering, Southeast University, Nanjing 211189, China)
([2] Southeast University-Monash University Joint Graduate School, Suzhou 215123, China)

Abstract: Aiming at the low accuracy rate of the generalized sequence pattern (GSP) algorithm on product feature extraction from Chinese online reviews, a secondary pruning algorithm is proposed. In this algorithm, based on the candidate collection of the output of the GSP algorithm, the term pair co-occurrence weight (TPCW) is used as the threshold for further filtering to improve the accuracy rate. The customized tools are used to crawl the product Chinese reviews of cameras from Jingdong website. 1 000 reviews are selected as the experimental data and the segmentation tool ICTCLAS is used on the word segmentation and data preprocessing. The proposed algorithm is compared with the GSP algorithm, the cross language model (CLM), and the likelihood ratio test (LRT). The results show that the accuracy rate of the proposed algorithm on product feature extraction from Chinese online reviews is 76.37%, which is higher than those of the GSP algorithm, CLM and LRT by 2.94%, 5.77% and 7.57%, respectively.

Key words: feature extraction; secondary pruning; term pair co-occurrence weight; likelihood ratio test; cross language model

随着互联网应用的不断发展,网络购物逐渐成为一种消费时尚.在线评论作为网络购物平台的重要组成部分,为网购用户做出购买决策及制造商改善产品提供重要依据.在线评论是网购用户对产品及产品属性的主观评价,它包含了产品符合度(包括产品整体及部分的性能、外观、手感、质量优

收稿日期: 2015-08-22. 作者简介: 吴含前(1972—),男,博士,副教授, hanqian@seu.edu.cn.
基金项目: 中央高校基本科研业务费专项资金资助项目、国家高技术研究发展计划(863计划)资助项目(2015AA015904).
引用本文: 吴含前,周立凤,谢珏.二次剪枝算法在评论特征提取中的应用[J].东南大学学报(自然科学版),2016,46(3):513-517. DOI: 10.3969/j.issn.1001-0505.2016.03.010.

劣情况等)、卖家服务态度、物流速度等丰富信息. 如何从这些海量的评论中挖掘有价值的信息成为当前学术界和工业界的一个研究热点.

产品特征的提取是在线评论挖掘研究工作中的重要组成部分[1]. 学术界通常将评论中出现的频繁名词或名词短语作为产品特征[2-4]. 通过提取产品特征,可以帮助生产商、销售商和网购用户全面分析和了解产品的各个属性,进而为生产商改进产品、销售商推广产品及网购用户做出购买决策等提供重要参考信息. 此外,产品特征的提取作为后续评论挖掘工作的研究基础,其结果的准确率将直接影响到最终评论挖掘系统的可靠性.

国内外学者已对产品特征的自动化提取展开了深入研究. Hu 等[1]于2004年采用关联规则算法进行了特征提取; Popescu 等[5]利用互信息去除不是属性的高频名词; Li 等[2]基于每个语句与结构化的语义信息来构建解析树,从而自动化识别特征; Chen 等[6]利用条件熵和通用算法来自动化提取特征; 李实等[7]利用关联规则挖掘算法进行中文产品特征识别; Javed 等[8]利用最大熵模型进行特征提取. 本文针对 GSP 算法[9]自动提取产品特征准确率不高的问题,提出了一种二次剪枝算法(GSP-TPCW 算法). 该算法汲取了 GSP 获取频繁名词和词组的优点,利用概率统计方法来计算词语与主题的相关性,通过设定相关度阈值来过滤噪声数据,从而提高中文产品评论特征提取的准确率. 将二次剪枝算法与 GSP 算法、交叉语言模型(CLM)算法[10]、似然比检验(LRT)算法[11]进行比较,以验证所提算法的有效性.

1 二次剪枝算法

二次剪枝算法是在 GSP 算法的基础上,利用概率统计的方法计算词对共现度,以获取词语与主题的相关性,并通过设定词语与主题相关度的阈值来实现噪声数据的过滤,从而提高中文产品评论特征的提取准确率.

1.1 GSP 算法原理

GSP 算法采用冗余候选模式中的剪枝策略进行剪枝,最初是为处理数据库中形如〈客户 ID 交易时间交易项〉的客户交易记录而提出的. GSP 算法考虑了项与项之间的顺序. 例如,组合词“性价比”通过 ICTCLAS 分词工具处理之后显示为“性价/n 比/p”,“性价”与“比”之间有着严格的先后顺序,而关联规则挖掘算法并不适合提取此类特征,类似的组合词还有“摄像头”、“硅胶套”等. GSP 算法中的相关概念如下[12]:

1) 序列模式

序列模式是由频繁出现的数据项构建的模式,即一个序列模式 s 表示一个排过序的项集列表 $\langle\alpha_1,\alpha_2,\cdots,\alpha_n\rangle$. 其中,$\alpha_i(i=1,2,\cdots,n)$ 为序列模式 s 中的一个序列元素. 同时,α_i 又可表示为一组项目的集合$\{\chi_1,\chi_2,\cdots,\chi_m\}$,其中,$\chi_j$ 为一个项目元素. GSP 算法的目标是从序列模式集合中找出所有满足用户指定的最小支持度序列模式(频繁序列),可将序列模式集合中该序列模式的个数作为其支持度.

2) 词对共现度

词对共现度是用来度量词语与主题之间关联程度的指标,可以通过统计概率的方法计算得到[12],其计算公式为

$$w(t_{ij},t_{ik})=\frac{2f(t_{ij}\cap t_{ik})}{f(t_{ij})+f(t_{ik})}$$

式中,$f(t_{ij})$ 表示词 t_j 在第 i 条评论中出现的次数; $f(t_{ij}\cap t_{ik})$ 表示词 t_j,t_k 在第 i 条评论中连续出现的次数. w 值越大,说明特征名词与主题关系越紧密.

1.2 GSP-TPCW 算法步骤

令 F_k 表示满足最小支持度且长度为 k 的频繁序列集合,C_k 表示所有长度为 k 的频繁候选序列集合,则 GSP-TPCW 算法步骤如下:

① 将原始数据集合存入模式数据库 S 中,形成候选集 C_1.

② 遍历数据库 S,找出满足最小支持度的序列模式并将其存入 F_1.

③ 将元素长度为 k 的 F_k 作为种子集,通过连接和剪枝操作,形成长度为 $k+1$ 的候选集 C_{k+1} 并将其作为种子集. 重复该步骤,直到没有新的序列模式或候选序列模式产生为止.

④ 通过计算词对共现度,对获得的候选序列模式集合进行二次剪枝. 检查序列模式集合中任意2个序列模式在分别去掉其第1个项目和最后1个项目后得到的序列是否相同,如果相同,则对这2个序列模式进行连接处理; 否则,不进行连接处理. 然后,计算每个候选序列模式的支持度,如果结果小于最小支持度,则将其从候选序列模式集合中删除,否则保留该候选序列模式,直到最终获得满足要求的序列模式集合.

2 交叉语言模型

Zhai 等[10]提出了一种基于概率方法的交叉语言模型. 该模型被广泛用于语音识别和信息检索

领域.

假定一个文档由一个目标短语集合和一个资料库短语集合构成,即

$$\theta = \alpha\theta_{\text{corpus}} + \beta\theta_{\text{query}} \tag{1}$$

式中,θ 表示从评论集合中获取的名词集合; θ_{corpus} 表示资料库名词集合; θ_{query} 表示与主题相关的名词集合; α,β 分别为 θ_{corpus} 和 θ_{query} 的噪声干扰因子,且 $\alpha+\beta=1$.

Zhang 等[4] 提出了一种时间复杂度为 $O(n\log(n))$ 的算法(n 为数据规模),以获取交叉语言模型中的 θ. 交叉语言模型可表述为

$$\boldsymbol{r} = \alpha\boldsymbol{p} + \beta\boldsymbol{q} \tag{2}$$

式中,$\boldsymbol{r}$,$\boldsymbol{p}$ 和 $\boldsymbol{q}$ 为多维向量.

假设 f_i 和 p_i 分别为 $\boldsymbol{r}$ 和 $\boldsymbol{p}$ 中第 i 个词出现的频率,构造式(2)的对数似然函数为

$$l = \sum_{i=1}^{k} f_i\log(r_i) = \sum_{i=1}^{k} f_i\log(\alpha p_i + \beta q_i)$$

$$\text{s.t.} \quad \sum_i q_i = 1,\ q_i \geqslant 0$$

对于所有满足 $q_i>0$ 的 q_i,采用拉格朗日乘子方法,得到如下公式:

$$L = l - \lambda\left(\sum_i q_i - 1\right)$$

通过求解目标函数的极大值可得

$$q_i = \begin{cases} \dfrac{f_i}{\lambda} - \dfrac{\alpha}{\beta}p_i & 1\leqslant i\leqslant t \\ 0 & \text{其他} \end{cases}$$

式中,$\lambda = \sum_{i=1}^{t} f_i \Big/ \left(1 + \alpha/\beta\sum_{i=1}^{t} p_i\right)$.

根据计算所得的 q_i 值进行排序,并设定一个阈值,选取大于该阈值的词构建最终的特征集合.

3　似然比检验

似然比检验是一种基于统计学的检验方法,用于比较 2 个模型之间的拟合情况[11]. 令 D_+ 为与主题 T 相关的文档集合; D_- 为与主题 T 不相关的文档集合; w 为从 D_+ 提取的候选特征,一般为名词或者名词短语; $\overline{w}$ 为 D_+ 中除去候选特征 w 以外的其他特征; d 为单个文档. 则似然比 $-2\log\lambda$ 定义如下:

$$-2\log\lambda = -2\log\frac{\max\limits_{p_1\leqslant p_2} L(p_1,p_2)}{\max L(p_1,p_2)} \tag{3}$$

式中,$p_1 = p(d\in D_+ \mid w\in d)$; $p_2 = p(d\in D_+ \mid \overline{w}\in d)$; $L(p_1,p_2)$ 为词 w 既在 D_+ 又在 D_- 中的似然值. 词频统计结果见表 1.

表 1　词频统计

词	D_+	D_-
w	C_{11}	C_{12}
$\overline{w}$	C_{21}	C_{22}

假设每个词都服从贝努利分布 $b(p,k,n) = p^k\cdot(1-p)^{n-k}$,表 1 中的计数服从二项式分布,则式(3)中的似然比近似服从 χ^2 分布,即

$$-2\log\lambda = -2\log\frac{x}{y} \tag{4}$$

式中,

$$x = \max_{p_1\leqslant p_2} b(p_1,C_{11},C_{11}+C_{12})\,b(p_2,C_{21},C_{21}+C_{22})$$

$$y = \max\ b(p_1,C_{11},C_{11}+C_{12})\,b(p_2,C_{21},C_{21}+C_{22})$$

式(4)可以转换为

$$-2\log\lambda = \begin{cases} -2h & r_2<r_1 \\ 0 & r_2\geqslant r_1 \end{cases} \tag{5}$$

式中,

$$r_1 = \frac{C_{11}}{C_{11}+C_{12}},\quad r_2 = \frac{C_{21}}{C_{21}+C_{22}}$$

$$h = (C_{11}+C_{21})\log(r) + (C_{12}+C_{22})\log(1-r) - C_{11}\log(r_1) - C_{12}\log(1-r_1) - C_{21}\log(r_2) - C_{22}\log(1-r_2)$$

$$r = \frac{C_{11}+C_{21}}{C_{11}+C_{12}+C_{21}+C_{22}}$$

根据式(5)可以计算得到词 w 的似然比. 似然比越大,表示词 w 与主题 T 之间的相关度也越大. 在实际应用中,可以通过设定阈值来选取最终的特征集合.

4　实验与结果

4.1　资料库的建立

实现似然比检验和交叉语言模型需要构建主题相关资料库和主题无关资料库. 为此,本文利用自定义爬虫工具抓取了京东商城中的 10 个电子产品(包括电脑、耳机、音响、MP3、平板、电源等),共 1×10^5 条评论,8 位研究生采用人工方式参与了对上述评论中名词和名词词组的提取工作,剔除与电子产品无关的名词和一些常用名词(如东西、产品、东东等),构建了主题相关资料库. 此外,采用同样方法,通过抓取豆瓣网站的 60 部电影评论,构建了主题无关资料库.

4.2　数据集的选取和处理

实验数据集选用了摄像头的产品评论,经过预处理后选择 1 000 条评论作为测试数据. 采用中国科学院的分词工具 ICTCLAS 对上述测试数据进行分词,并提取其中的名词. 例如,对于评论“摄像

头不错,性价比很高!”,经过分词后得到的名词序列为“摄像、头、性、价”,将该名词序列作为一行记录存入文本文件中,其中“摄像”、“头”、“性”、“价”均为序列中的项.

为了评估各种特征提取算法的有效性,将准确率 p、召回率 r 以及两者的加权平均值 f 作为测试算法性能的评价指标,其中 f 的计算公式如下:

$$f=\frac{2pr}{p+r} \tag{6}$$

4.3 结果比较及分析

根据实践经验并结合本文设计的实验环境,二次剪枝过程中词对共现度的阈值设为0.68. GSP-TPCW算法与GSP算法的实验结果见表2.

表2 GSP-TPCW算法与GSP算法的评价指标 %

算法	准确率	召回率	加权平均值
GSP-TPCW	76.37	54.18	63.34
GSP	73.43	50.54	59.87

由表2可知,就准确率、召回率以及两者加权平均值这3个评价指标而言,本文提出的GSP-TPCW算法明显高于GSP算法,这是因为所提算法通过设定的词对共现度阈值有效过滤掉一些评论中不存在的名词组合,从而提高了算法的性能. 另外,GSP-TPCW算法的召回率相对偏低,这与实验中设置的候选序列模式最小支持度有关,若最小支持度设置过低,则会产生较多噪声数据,导致准确率不高;反之,若最小支持度设置过高,准确率会提高,但召回率降低. 本实验中,将最小支持度设置为3.

为了进一步验证GSP-TPCW算法的有效性,将其与交叉语言模型和似然比检验进行对比. 实验过程中考虑了以下2种情况:① 保留单个汉字的名词;② 删除单个汉字的名词. 实验结果见表3.

表3 3种算法的评价指标 %

算法	情况1			情况2		
	准确率	召回率	加权平均值	准确率	召回率	加权平均值
GSP-TPCW	76.37	54.18	63.34	78.10	52.20	62.46
CLM	58.09	51.97	54.86	72.33	49.01	58.43
LTR	65.70	52.30	58.24	70.53	48.03	57.15

由表3可知,GSP-TPCW算法的综合性能较似然比检验和交叉语言模型好,这与分词工具的准确性和中文评论资料库的完善程度有关. 例如,ICTCLAS分词工具将“性价比很高”分词为“性/n 价/n 比/p 很/d 高/a”,提取名词为“性”与“价”;将“摄像头不错”分词为“摄像/n 头/n 不错/a”,提取名词为“摄像”与“头”;显然,“性价比”和“摄像头”这些常用名词不能通过似然比检验和交叉语言模型准确获取,而GSP-TPCW算法则可通过获取频繁词序列得到. 此外,删除单个汉字名词后的准确率和加权平均值均比保留单个汉字名词的准确率与调和平均值高,这进一步证明了GSP-TPCW算法的有效性. 实验过程中,采用似然比检验和交叉语言模型都需要构建资料库,且要求人工参与,而GSP-TPCW算法则不需要构建资料库,从而能够实现针对产品特征的高效提取.

综上可知,利用GSP-TPCW算法获得的中文产品评论特征提取准确率达到76.37%,较GSP算法、CLM和LRT的准确率分别提高2.94%,5.77%和7.57%.

基于本文实验采用的资料库,采用GSP-TPCW算法、交叉语言模型和似然比检验获得的部分产品特征见表4.

表4 摄像头特征自动获取结果

算法	摄像头特征
GSP-TPCW	产品质量,摄像头,清晰度,效果,外观,摄像,价格,镜头,造型,配置,颜色,三脚架,产品效果
LRT	价格,屏幕,摄像,像素,分辨率,清晰度,外观,手感,指示灯,灵敏度,尺寸,对焦
CLM	质量,清晰度,效果,外观,摄像,价格,造型,配置,颜色,镜头,对焦,录像

5 结语

本文针对网购平台中文在线评论的产品特征提取,提出了一种二次剪枝算法. 该算法在GSP算法的基础上,采用词对共现度作为阈值对候选特征集进行进一步筛选,从而提高准确率. 采用1 000条摄像头的中文评论作为测试数据,并与GSP算法、交叉语言模型和似然比检验进行了比较. 结果表明,利用所提算法进行中文产品评论特征提取可获得较高的准确率.

参考文献(References)

[1] Hu M, Liu B. Mining opinion features in customer reviews [C]//*Proceedings of the 19th National Conference on Artifical Intelligence*. Chicago, Illinois, USA, 2004: 755-760.

[2] Li F, Pan S J, Jin O, et al. Cross-domain co-extraction of sentiment and topic lexicons [C]//*Meeting of the As-*

sociation for Computational Linguistics: *Long Papers*. Beijing,China, 2012: 410-419.

[3] Fei G, Liu B, Hsu M, et al. A dictionary-based approach to identifying aspects implied by adjectives for opinion mining [C] //24*th International Conference on Computational Linguistics*. Chicago, Illinois, USA, 2012: 309-318.

[4] Zhang Y, Xu W. Fast exact maximum likelihood estimation for mixture of language model [J]. *Information Processing & Management*, 2008, **44**(3): 1076-1085. DOI: 10.1016/j.ipm.2007.12.003.

[5] Popescu A-M, Etzioni O. Extracting product features and opinions from reviews [C] //*Proceedings of the Conference on Human Language Technology and Empirical Methods in Natural Language Processing*. Washington, DC, USA, 2005: 32-33.

[6] Chen Y, Wang X. Text feature extraction based on joint conditional entropy [C] //*Proceedings of* 2012 2*nd International Conference on Computer Science and Network Technology*. Changchun,China, 2012: 2055-2058.

[7] **李实, 叶强, 李一军, 等. 挖掘中文网络客户评论的产品特征及情感倾向** [J]. **计算机应用研究**, 2010, 27(8): 3016-3019. DOI: 10.3969/j.issn.1001-3695.2010.08.054.
Li Shi, Ye Qiang, Li Yijun, et al. Mining product features and sentiment orientation from Chinese customer reviews [J]. *Application Research of Computers*, 2010, **27**(8): 3016-3019. DOI: 10.3969/j.issn.1001-3695.2010.08.054. (in Chinese)

[8] Javed K, Babri H A, Saeed M. Feature selection based on class-dependent densities for high-dimensional binary data [J]. *IEEE Trans Knowl Data Eng*, 2012, **24**(3): 465-477. DOI: 10.1109/tkde.2010.263.

[9] Agrawal R, Srikant R. Mining sequential patterns [C] //*Proceedings of the Eleventh International Conference on Data Engineering*. Taipei,China, 1995: 3-14.

[10] Zhai C, Lafferty J. Model-based feedback in the language modeling approach to information retrieval [C] //*Proceedings of the Tenth International Conference on Information and Knowledge Management*. Pittsburgh, Pennsylvania, USA, 2001: 403-410.

[11] Ferreira L, Jakob N, Gurevych I. A comparative study of feature extraction algorithms in customer reviews [C] // 2008 *IEEE International Conference on Semantic Computing*. Santa Clara, California, USA, 2008: 144-151. DOI: 10.1109/icsc.2008.40.

[12] Zheng Y, Ye L, Wu G F, et al. Extracting product features from Chinese customer reviews [C] //2008 3*rd International Conference on Intelligent System and Knowledge Engineering*. Xiamen, China, 2008: 285-290. DOI: 10.1109/iske.2008.4730942.