情报分析与研究

中文网络评论中产品特征提取方法研究*

王 永 张 勤 杨晓洁
(重庆邮电大学经济管理学院 重庆 400065)

【摘要】针对中文网络客户评论中产品特征提取问题,提出采用 FP 增长算法获取候选产品特征集,再根据独立支持度、频繁项名词非特征规则及 PMI 阈值过滤技术对候选产品特征进行筛选,得到最终产品特征集,从而实现对中文网络客户评论中产品特征信息的自动挖掘。采用数据堂提供的手机评论语料,对该方法进行数据实验,实验结果可以验证该方法的有效性。

【关键词】产品特征 特征提取 关联规则 评论挖掘
【分类号】TP393

Research on the Method of Extracting Features from Chinese Product Reviews on the Internet

Wang Yong Zhang Qin Yang Xiaojie
(School of Economics and Management, Chongqing University of Posts and Telecommunications, Chongqing 400065, China)

【Abstract】Aim for better solving the problem of extracting features from Chinese product reviews on the Internet, an approach using FP - growth algorithm is proposed to obtain the set of candidate product features. Then, the candidate product features are filtered according to the rules of p - support, non - features frequent nouns and PMI threshold filtering technology. Finally, the final product features set are obtained. Thus, the automatic mining of product features information from Chinese customer reviews on the Internet is achieved. The proposed method is tested with the cell phone reviews from Datatang and the results show that the presented method is valid and effective.

【Keywords】Product features Features extracting Association rules Review mining

1 引 言

随着大数据时代的到来,数据成为商业活动中的一种重要资源,基于数据的科学决策和精细化管理将成为现代商业管理发展的必然趋势。在电子商务领域,海量的商品评论数据蕴含着巨大的社会价值和商业价值。对海量商品评论中产品特征数据进行分析挖掘,可为潜在消费者提供商品属性粒度级别的购买决策依据;为企业提供产品设计的依据和其他企业的竞争情报,还能对用户的需求和产品的改进方向做出有效反应[1],提高企业竞争力。

目前产品特征提取的研究方法主要分为人工定义和自动提取两类。Zhuang 等[2]、Kobayashi 等[3]、娄德成等[4]采用人工或半自动的方式对电影、游戏和中文汽车领域进行产品特征提取研究。Shi 等[5]人工定义了基于产品属性的概念模型,并以此模型对中文领域产品特征进行研究。但是这些方法移植性较差,当产品功能发生改变

收稿日期:2013 - 08 - 12
收修改稿日期:2013 - 09 - 26
* 本文系国家社会科学基金项目"差错管理气氛对企业创新行为的影响机理及对策研究"(项目编号:12CGL049)和重庆市自然科学基金项目"基于在线社交网络的舆情演化及社会化协同过滤推荐算法研究"(项目编号:CSTC2011jjA40045)的研究成果之一。

时,需要重新构建产品特征集合,效率不高。Yi 等[8]提出产品特征词一般是具有 BNP(Base Noun Phrase)结构的名词或名词短语,并采用信息检索算法判别该特征与指定产品是否相关。余传明[7]采用基于 SOM 的产品属性挖掘方法对餐馆评论进行研究,取得了较好的效果。Hu 等[9]首先提出使用关联规则分类方法 Apriori 算法对英文评论中的产品特征进行提取。李实等[9,10]参考 Hu 等的研究方法,针对中文语言特点,提出中文文本评论中的产品特征提取方法。虽然上述方法结构简单便于实现,也具有良好的移植性,但是由于没有充分考虑短语评价对象的结构特征以及评价对象的领域相关性,会产生较多噪声信息,因此准确率有待提高,而且 Apriori 算法会产生大量的候选项集,并反复扫描数据库,其运算效率不高。

鉴于此,本文提出了一种新的产品特征提取方法。该方法针对中文产品评论,以效率远高于 Apriori 算法的 FP 增长算法[11]来提取候选产品特征; 然后从产品和属性之间的语义关系角度出发,进行产品特征的筛选,弥补了关联规则算法只从数量上考虑关注程度的不足。实验结果表明,本文提出的方法能有效降低噪声,提高中文产品评论领域特征提取的挖掘性能。

2　方法设计

在本文的方法中,首先使用 FP 增长算法提取候选产品特征,并根据独立支持度规则对候选产品特征进行初步筛选; 然后制定频繁名词非特征规则,并建立相应名词集合,从中文语义角度进一步筛选产品特征; 最后采用 PMI 算法从语义相关角度降低噪声,得出最终的产品特征集合。

2.1　方法步骤

本文提出的面向中文网络客户评论的产品特征提取过程如图 1 所示:

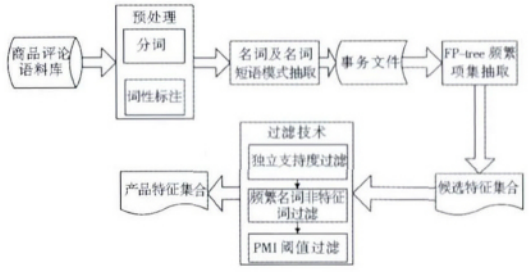

图 1　产品特征提取过程

该方法的具体步骤如下:

(1) 应用中国科学院计算技术研究所的中文分词工具 ICTCLAS 对原始评论语料进行分词和词性标注。

(2) 利用词性标注后的评论语料提取名词或名词短语并创建关联规则事务文件。ICTCLAS 分词工具所使用的词性标注标记集中与名词相关的标记子集是{ /an, /ng, /n, /nr, /ns, /nt, /nz, /vn} ,根据这些标记所代表的含义以及产品属性词的语法特点,本文选择{ /n, /vn } 子集作为抽取规则。使用计算机语言对每一条评论进行名词及动名词抽取,并生成一条记录插入到事务文件中。

(3) 采用 FP 增长算法对事务文件进行扫描,将得到的频集生成一棵频繁模式树(FP - tree) ; 随后再将 FP - tree 分化成若干与长度为 1 的频集相关的条件库; 再分别对每一个条件库进行频繁特征识别,得到频繁项集,并将它作为候选产品特征集合 I0。

(4) 采用独立支持度规则对候选产品特征集 I0 中的名词及名词短语进行过滤、修正,形成候选特征集 I1。本文将评论中包含频繁特征名词或名词短语 ftr 且不包含 ftr 父集的句子数量称为 ftr 的独立支持度(P - support) 。本研究采用最小支持度为 1% 进行实验。

(5) 制定中文频繁项名词非产品特征规则,并建立相应的名词集合,从中文语义及语法知识角度过滤 I1,形成特征集合 I2。本研究将中文频繁项名词却非产品特征主要划定为以下几种情况:

①常见的抽象性名词,如“情况”、“事情”、“原因”等。

②所评价产品名称,如“酒店”、“宾馆”、“手机”等。

③用户口语化的评论名词,如“本子”、“机子”等。

④与产品无关的称呼类名词,如“网友”、“老板”、“同事”等。

⑤常见的集合类名词,如“人员”、“大家”等。

(6) 使用基于网络搜索引擎的 PMI 算法[12]计算产品和特征集合 I2 中各个特征的共现度——PMI 值,并按照 PMI 值从大到小进行排列,PMI 值越高,二者之间的关联程度越大,通过多次实验选择最佳阈值,过滤共现度低的特征,形成最后的产品特征集合 I3。PMI 计算公式定义如下:

$$\mathrm{PMI}(\text{产品},\text{特征}) = \log_2 \frac{\mathrm{hit}(\text{"产品"}\ \mathrm{and}\ \text{"特征"})}{\sqrt{\mathrm{hit}(\text{"产品"})\,\mathrm{hit}(\text{"特征"})}} \quad (1)$$

其中,hit(x) 是以词语 x 为关键词查询时搜索引

擎所返回的页面数; hit(x and y) 是同时以 x 和 y 作为关键词查询所返回的页面数。本文选取百度搜索引擎返回的页面数作为 PMI 计算的依据。

2.2 性能评估

采用信息检索领域常用的性能评估指标: 查准率 P、查全率 R 和综合值 F - score。其中,查全率和查准率分别度量性能的某个方面,忽略任何一个都有失偏颇,综合值 F - score 是对性能的整体评估,具体计算方法如下:

$$P = \frac{A}{A + B} \quad (2)$$

$$R = \frac{A}{A + C} \quad (3)$$

$$F - score = \frac{2RP}{R + P} \quad (4)$$

其中,A 表示算法识别出来的产品特征数量,B 表示算法识别出来但不是产品特征的数量,C 表示算法未识别出但是是产品特征的数量。

3 实验结果及性能评价

3.1 实验数据

采用数据堂提供的手机评论语料(http: //www.datatang. com/data/43824) ,选择其中的 600 篇作为实验数据。通过手工标注的方法获取产品特征提取实验的参照特征。结合“中关村在线”、“京东商城”等网站对手机的评定标准,并依据最小 - 最大覆盖原则[9] ,对语料进行手工标注,获得覆盖 600 篇评论中提到的该商品特征的集合,共得到手机产品特征 86 个,如表 1 所示:

表 1 手工标注手机产品特征集合

名称	参数	手工标注特征集合	数量
手机	外观设计	外观,外形,造型,机型,外壳,按键,键盘,后盖,机身,材质,手感,颜色,重量,体积	14
	屏幕	屏幕,外屏,内屏,显示屏,色彩,分辨率,清晰度,画面	8
	基本功能	功能,短信,信息,彩信,通话,输入,录音,语音,闹钟,通讯录,电话簿,电话本,收音机,快捷键,软件,程序	16
	摄像功能	摄像头,相机,摄像,照相,照片,图片,象素,闪光灯	8
	娱乐功能	娱乐,多媒体,视频,音乐,播放器,游戏	6
	数据功能	蓝牙,红外线	2
	手机附件	配件,耳机,电池,数据线	4
	美化	界面,主题,背景,菜单	4
	性能	性能,待机,反应,速度,操作,开机,上网,信号,网络,智能	10
	声音	声音,铃声,音量,音质,扬声器,听筒	6
	硬件配置	配置,内存	2
	性价比	价格,价位,性价比	3
	售后反馈	售后服务,服务,质量	3

3.2 实验结果

(1) 产品特征提取结果

从词频和语义相关两个角度,得出手机评论语料中客户关注程度居前 10 名的手机特征,如表 2 所示:

表 2 手机产品特征提取结果(按 PMI 值排序)

排名	属性	PMI 值	词频
1	智能	0.0	14
2	软件	-0.100 05	42
3	号码	-0.444 18	30
4	屏幕	-0.652 9	194
5	价格	-0.798 37	34
6	功能	-0.846 84	297
7	图片	-0.865 12	46
8	游戏	-0.870 39	32
9	电池	-0.894 32	96
10	机型	-0.974 26	12

对 PMI 值设置不同的阈值,性能变化如图 2 所示,相应的性能指标值如表 3 所示:

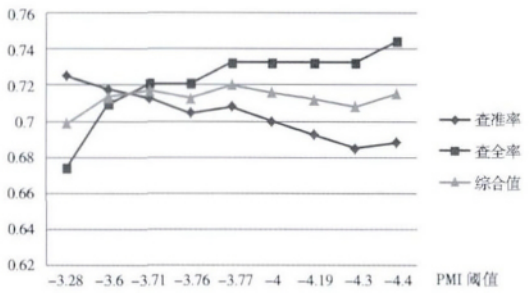

图 2 PMI 值选择不同阈值时的性能变化情况

表 3 手机评论挖掘性能

PMI 阈值	查准率	查全率	综合值
-3.28	72.5%	67.4%	69.9%
-3.6	71.8%	70.9%	71.3%
-3.71	71.3%	72.1%	71.7%
-3.76	70.5%	72.1%	71.3%
-3.77	70.8%	73.3%	72%
-4	70%	73.3%	71.6%
-4.19	69.2%	73.3%	71.2%
-4.3	68.5%	73.3%	70.8%
-4.4	68.8%	74.4%	71.5%

可以看出,PMI 阈值为 -3.77 时,挖掘结果综合性能最优,即查准率达到 70.8%,查全率达到 73.3%,综合值达到 72%。

(2) 对比分析

对于手机产品的评论特征挖掘,本文所提出的方法与其他方法的结果比较如表 4 所示。

表 4　针对手机评论的产品特征挖掘结果比较

性能指标	本文方法	文献[10]的方法	文献[5]的方法	文献[8]的方法
查准率	70.8%	62.8%	70.72%	71.8%
查全率	73.3%	81.8%	68.35%	76.1%
综合值	72%	71.05%	69.51%	73.89%

可以看出，对于中文手机评论，本文方法的查准率优于文献[5]和文献[10]；查全率优于文献[5]，但逊于文献[10]；综合评价指标优于文献[5]和文献[10]。文献[8]的研究对象集合是英文评论，由于对象不同，两者之间不具有绝对的可比性，但是挖掘性能已基本接近。

本文方法是基于 FP 增长算法设计的，由于 FP 算法的运行效率远高于 Apriori 算法[11]，因此，本文方法的效率远高于文献[8]和文献[10]中方法的效率。

4　结　语

产品特征作为互联网海量商品评论信息的一个重要方面，是其他用户做出购买决策的参数，更是生产商和销售商改进商品和服务的关键指标。对产品特征进行提取是文本评论挖掘的重要任务之一，直接影响着评论挖掘的性能。在英文文本评论领域，研究者已初步取得一些成果，而针对中文网络产品评论的研究还处于探索阶段，存在诸多不足。本文从理论上对中文客户评论产品特征挖掘问题进行了探索，将关联规则分类方法 FP 增长算法应用于产品特征提取领域，并采用独立支持度规则、频繁名词非特征规则以及 PMI 算法，从多个角度筛选产品特征，拓展了基于关联规则的产品特征挖掘方法。数据实验结果表明，本文方法具有良好的性能，有望在一定程度上解决网络评论数据过载以及信息非结构化等问题。

参考文献:

[1] 翟东升，徐颖，黄鲁成，等. 基于产品评论挖掘的竞争产品优势分析[J]. *情报杂志*，2013，32(2)：45－51.（Zhai Dongsheng，Xu Ying，Huang Lucheng，et al. The Advantage Analysis of Competitive Product Based on Product Reviews Mining[J]. *Journal of Intelligence*，2013，32(2)：45－51.）

[2] Zhuang L，Jing F，Zhu X Y. Movie Review Mining and Summarization[C]. In: *Proceedings of the* 15*th ACM International Conference on Information and Knowledge Management*（*CIKM'*06），Arlington，Virginia，USA. New York: ACM，2006: 43－50.

[3] Kobayashi N，Inui K，Matsumoto Y，et al. Collecting Evaluative Expressions for Opinion Extraction[C]. In: *Proceedings of the* 1*st International Joint Conference on Natural Language Processing*（*IJCNLP'*04）. Berlin，Heidelberg: Springer－Verlag，2004: 596－605.

[4] 娄德成，姚天昉. 汉语句子语义极性分析和观点抽取方法的研究[J]. *计算机应用*，2006，26(11)：2622－2625.（Lou Decheng，Yao Tianfang. Semantic Polarity Analysis and Opinion Mining on Chinese Review Sentences[J]. *Journal of Computer Applications*，2006，26(11)：2622－2625.）

[5] Shi B，Chang K. Mining Chinese Reviews[C]. In: *Proceedings of the* 6*th IEEE International Conference on Data Mining*. Washington D C: IEEE Computer Society，2006: 585－589.

[6] Yi J，Nasukawa T，Bunescu R，et al. Sentiment Analyzer: Extracting Sentiments About a Given Topic Using Natural Language Processing Techniques[C]. In: *Proceedings of the* 3*rd IEEE International Conference on Data Mining*（*ICDM'*03）. Washington D C: IEEE Computer Society，2003: 427.

[7] 余传明. 从用户评论中挖掘产品属性——基于 SOM 的实现[J]. *现代图书情报技术*，2009(5)：61－66.（Yu Chuanming. Mining Product Aspects from User Reviews—— An SOM－based Approach[J]. *New Technology of Library and Information Service*，2009(5)：61－66.）

[8] Hu M，Liu B. Mining and Summarizing Customer Reviews[C]. In: *Proceedings of the* 10*th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*（*KDD'*04）. New York: ACM，2004: 168－177.

[9] 李实，叶强，李一军，等. 中文网络客户评论的产品特征挖掘方法研究[J]. *管理科学学报*，2009(2)：142－152.（Li Shi，Ye Qiang，Li Yijun，et al. Mining Features of Products from Chinese Customer Online Reviews[J]. *Journal of Management Sciences in China*，2009(2)：142－152.）

[10] 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.（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.）

[11] Han J，Pei J，Yin Y，et al. Mining Frequent Patterns without Candidate Generation: A Frequent－Pattern Tree Approach[C]. In: *Proceedings of the* 2000 *ACM SIGMOD*，Dallas，USA. 2000: 1－12.

[12] Church K W，Hanks P. Word Association Norms，Mutual Information and Lexicography[C]. In: *Proceedings of the* 27*th Annual Meeting on Association for Computational Linguistics*，New Brunswick，NJ，Canada. Stroudsburg: Association for Computational Linguistics，1989: 76－83.

（作者 E－mail: wangyong_cqupt@163.com）