情报学报 ISSN 1000-0135
第35卷第1期 77-83，2016年1月

JOURNAL OF THE CHINA SOCIETY FOR SCIENTIFIC AND TECHNICAL INFORMATION ISSN 1000-0135
Vol. 35 No. 1, 77-83 January 2016

doi:10.3772/j.issn.1000-0135.2016.001.008

一种基于网络评论的商品特征挖掘方法[1)]

郭崇慧 张倚天

（大连理工大学 系统工程研究所，大连 116024）

摘要 商品评论中含有大量的有用信息，这些信息对买方的购买行为和卖方的销售行为都有着显著的影响。商品特征作为网络评论中的关键信息，有重要的实际意义和研究价值。本文提出一种新的商品特征挖掘方法，该方法通过扩充用户词典来提升候选特征的准确性，同时引入同义词表对候选特征有效地剪枝，此外还提出情感指数的概念并以此作为从候选集中选择商品特征的依据，并从电商网站分别获取了手机和数码相机等四种商品的相关评论用于数值实验。实验结果显示该挖掘方法是可行的、有效的，不仅很好地提升现有研究结果的准确性，同时也为商品特征挖掘领域提供了新的研究思路。

关键词 评论挖掘 商品特征 用户词典 同义词 情感指数

A Product Feature Mining Method Based on Online Reviews

Guo Chonghui and Zhang Yitian

(*Institute of Systems Engineering, Dalian University of Technology, Dalian* 116024)

Abstract Product reviews contain a lot of useful information and have a significant influence on purchasing and selling behaviors. As the critical information from online reviews, product features have important theoretical and practical values. This paper presents a new product feature mining method, which improves the accuracy of candidate sets by extending user dictionary, and introduces synonyms for effectively pruning. In addition, a concept of sentiment index is proposed to select product features from candidate sets. This study respectively crawls online reviews of 4 products like cell phone and digital camera etc. For numerical experiment, and the result shows its feasibility and effectiveness. It not only well enhances the accuracy of existing research results, but also provides new research ideas for the product feature mining area.

Keywords review mining, product features, user dictionary, synonyms, sentiment index

1 引 言

随着电子商务的快速发展，越来越多的消费者习惯于进行网上购物。消费者在发生购买行为后，可以对已购买的商品进行评论，这些评论不仅是消费者对商品卖家的反馈，同时也能对其他消费者提供建议和指导[1]。商品的热销意味着商品评论的大量增加，某些火爆的商品动辄数万条的评论让卖家和买家都难以处理，这就需要双方从海量的商品评论中快速地筛选出有用的评论，从大量冗余的信息中提取出真正可以指导销售和购买的有用信息。

对网络商品评论信息提取的迫切需求使得国内外研究者都不约而同地关注起了文本挖掘的一个具体的应用领域——评论挖掘[2-4]。评论挖掘旨在对非结构化的评论进行有用信息的提取，将消费者所关注的商品特征、商品质量、其他消费者对商品的喜好等全面准确地表示出来。作为商品评论挖掘的第

收稿日期：2015年7月2日

作者简介：**郭崇慧**，男，1973年生，教授，博士生导师，主要研究方向：系统优化方法，数据挖掘与机器学习。E-mail: guochonghui@tsinghua.org.cn。**张倚天**，男，1991年生，硕士研究生，主要研究方向：文本挖掘，信息检索。

1） 国家自然科学基金资助项目（71171030，71421001）；国家软科学研究计划项目（2013GXS2D018）。

一个环节，商品特征的挖掘不仅能对评论进行有效降维，把复杂的非结构化评论用该商品的若干特征进行有效表示，还可以为后续情感分析等其他步骤提供挖掘的对象和数据基础，是整个评论挖掘过程中极为重要的一环。

本文的结构如下：第一节介绍了中英文评论特征挖掘的相关研究工作，并指出已有的进展和存在的不足；第二节对基于评论的商品特征挖掘研究进行问题描述；第三节给出本文提出的商品特征挖掘方法，详述了扩充词典、引入同义词表和利用情感指数选择商品特征等创新点；第四节介绍了该方法的数值实验和结果的讨论；第五节对整篇文章做出了总结。

2 相关工作

商品评论特征挖掘作为文本挖掘的一个具体应用领域，因其在电子商务客户评论中重要的实际意义而被国内外学者广泛关注。Hu 等[4]首先提出了一个系统完整的评论挖掘过程，包括商品特征挖掘、主观句定位、情感分析、极性判定以及结果显示等部分，并在文献[5]中详细阐述了商品特征挖掘的方法。文献[6]在电影评论的特征挖掘中借鉴了文献[4]搜索情感词的方法，首先给出一部分特征词，然后在 WordNet 词典中搜索出这些特征词的同义词作为电影评论的特征。该方法考虑到了同义词在特征挖掘中的应用，然而没能通过对同义特征词的剪枝来得到更精确的挖掘结果。由于较早的特征挖掘结果在查准率和查全率上并不尽如人意，Popescu 等[7]将改进的点互信息（Point Mutual Information，PMI）值引入特征词剪枝，查全率略有降低而查准率有了显著提升，该方法同样没有考虑同义词对同一特征表示的重复性。Scaffidi 等[8]开发了一种新的检索系统 Red Opal，它只识别单个词和二词短语作为特征，并对每个产品的每个特征都进行打分，输出结果按照产品特征的打分综合排序。该方法用打分的方式来区分特征的重要程度，但是在识别特征的过程中依然没有用更加有效的剪枝来剔除冗余特征。通过以上研究可以看出，相近词义的特征词在表达同一特征时有很大的重复性，同义词剪枝作为过滤特征的关键步骤能显著提升查准率，却往往被忽略；其他相关研究也大都采用关联规则来挖掘特征[9,10]，因此设计其他高效挖掘方法也值得进一步研究。

近年来中文评论的特征挖掘研究也逐渐深入，认识到其与英文评论特征挖掘的区别与联系是研究的前提。李实等[11]指出，相较英文评论挖掘，中文评论挖掘新增的难点主要有中文分词的准确性、词性标注的灵活性和名词短语结构等。因而，中文评论挖掘要关注分词、剪枝以及特征选取方法等多个方面。早期的中文文本挖掘研究[12,13]也指出了分词和特征词选取精度的重要性。文献[14]用中国科学院计算技术研究所开发的分词器 ICTCLAS 来对中文评论进行分词和词性标注，结果显示该分词器有较高的精度，但分词器的词典中不包含的领域词汇却很难被挖掘出来。文献[15]借鉴了文献[6]和文献[14]的成果，用分词器 ICTCLAS 对评论进行处理，并用关联规则挖掘出特征，该结果的精确度已经接近了英文商品评论挖掘，但是同样没有优化分词过程来进一步提升结果。在特征挖掘的关键方法选择上，周茜等[16]介绍了 8 种文本特征选择的方法，包括 TF-IDF 和信息增益等，给特征挖掘研究提供了新的思路。郁亚辉[17]考虑到特征与观点之间的联系，通过双向传播算法来找到商品特征和用户观点并利用领域相关度来优化特征抽取结果，但是该方法没有对分词阶段做进一步的优化处理。文献[18]运用改进的 TF-IDF 算法提取特征，并用于文本的分类，得到了很好的效果，但没有对特征提取前的分词和剪枝步骤做出优化。另外，孙春华等[19]还研究了评价商品特征的倾向性合成等内容，以此来提升评论接收者的信息处理和整合能力。

尽管目前中文评论挖掘研究已经有了很大的进展，但在特征挖掘的研究领域仍然有提高的空间。为了进一步提升特征挖掘结果的准确性，本文对已有研究的分词、剪枝和特征提取三个步骤进行了有效改进：对于分词步骤，通过扩充用户词典来尽可能地保留更多的特征词；对于剪枝步骤，引入同义词典来剔除及合并同义词，有效减少语义重复词项；对于特征提取步骤，充分考虑到用户情感在评论中的作用，提出情感指数的概念并以此来挖掘出商品特征。结合以上改进，本文给出一套完整的基于网络评论的商品特征挖掘方法。

3 问题描述

给定电子商务网站上某店铺的某一特定商品 G，则该商品的原始评论语料为 $C_0=\{r_1, r_2, \cdots, r_k, \cdots, r_n\}$，其中 r_k 为该商品的第 k 条评论。为了得到

商品的特征，需要首先对每一篇评论进行分词和词性标注处理，对于分词和词性标注后的评论 r_k，有词语集合 $s_k = \{w_{k1}, w_{k2}, \cdots, w_{ki}, \cdots, w_{km}\}$，其中 w_{ki} 指评论 r_k 中包含的第 i 个词语，记第 i 个词语在评论 r_k 中的词频为 tf_{ki}。由于商品特征词的词性特点，本文从分词后的词语中抽出所有的名词作为特征候选初始集合 $S_0 = \{t_1, t_2, \cdots, t_i, \cdots, t_p\}$，其中 t_i 为候选集中的某个候选特征词，记该词语在评论语料中的词频为 tf_i。通过本文给出的特征挖掘方法对该特征候选集进行剪枝、排序和筛选，将得到最终的商品特征集合 $T = \{t'_1, t'_2, \cdots, t'_q\}$。

4 商品特征挖掘方法

对网络评论的商品特征挖掘方法包括评论分词、提取候选特征集、选择商品特征等步骤。在分词步骤中，将重点介绍扩充用户词典这一环节；在提取候选特征集步骤中，将同义词表作为优化结果的一个重要因素加以解释；在选取商品特征步骤中，将重点介绍情感指数的概念，并指出如何用情感指数来合理地选取商品的特征词。

4.1 网络评论分词

分词是中文评论挖掘的第一步，只有将整段的评论分割成不同词性的单词或短语，才能进行下一步的特征提取工作。然而，绝大多数研究工作在分词的步骤只关注分词的粒度以及词性的标注等问题，往往忽略了用户词典的问题。通过研究可以发现，不同的用户词典对最终特征提取的差异有显著影响。

“这款手机的性价比是很高的!”和“分辨率不错，屏幕看着非常清晰!”是两则关于手机的评价，可以明显看出“性价比”和“分辨率”是用户评论的商品特征对象。如果用户词典中有这两个词，那么这两个特征就有很大的机会被挖掘出来，但如果用户词典中没有这两个词，而是将这两个词分成了“性价”“比”和“分辨”“率”，这两个词被挖掘出来的难度就会加大。针对这种问题，如果在分词之前将手机评价的热词添加到用户词典中，甚至构建跟手机评价相关的领域词典，就可以在分词的环节大大增加结果的准确性。

为了验证扩充用户词典对结果的促进作用，本研究在某商品购买页面的“大家印象”板块挑选和总结了相关领域的 20 个关键词，人工添加到用户词典中。“大家印象”板块集中了不同用户对该商品的主观印象，从中总结出相应的关键词很有可能出现在评论中。在分词的步骤中，利用分词软件 ICTCLAS 对原始评论语料 C_0 进行分词和词性标注。值得注意的是，分词粒度和词性标注都有两个级别，本文选择细粒度分词和一级词性标注得到相应分词结果。

4.2 特征候选集提取

从已分出的词中选择所有词性标注为名词的单词和词组作为商品特征的候选集 S_0，对每个候选特征进行词频统计，并对该候选集做两次剪枝处理，即单字剪枝和同义词剪枝。单字剪枝是将候选集 S_0 中的所有单字名词从候选集去掉，得到候选集 S_1。用户对商品特征的关注集中在两个字以上的名词或名词短语，因而将单字名词从候选集中剔除可以有效缩小候选特征集合。

特征候选集 S_1 中可能会存在表示着相同或相近意思的不同词语，例如“礼品”和“礼物”这样一对同义词。通常希望挖掘出的商品特征能够分别表示商品的不同属性，彼此间在词意上有较大差别，因而需要对同义词进行合并和剪枝。为了实现这个过程，引入哈尔滨工业大学构建的同义词词林，如图 1 所示。

同义词词林中，每一个编号代表一组同义词，于是可以将候选集 S_1 里的每个特征都在同义词词林中查找到其所在位置，若有不同的特征在同一个编号中出现，就认为这些不同的特征属于同义词。考虑到同一个词语可能出现在不同的编号中，本文规定：若特征词 t_i、t_j 出现在编号 a 中，且特征词 t_j、t_k 出现在编号 b 中（$i \neq j \neq k, a \neq b$），则 t_i 和 t_k 也属于同义词。由此，可以得到特征候选集中所有的同义词特征集合。

在得到的每一个同义词集合中，选择出本集合词频最高的特征词作为唯一保留的特征词并剔除其他特征词，然后将本组其他被剔除的特征词的词频叠加到被保留特征词的词频上。最后，在原始评论语料 C_0 中，用本集合保留的唯一特征词替代同一集合中其他被删除的特征词。剔除同义词后的特征候选集记为 S_2，同义词替代后的评论语料记为 C_1，在接下来的步骤中还将会对二者做进一步处理。下面给出一个完整的剪枝步骤示例，如图 2 所示。

4.3 商品特征的选择

上一个步骤已经初步把无用的名词做了剔除，

Aa01A01= 人 士 人物 人士 人氏 人选
Aa01A02= 人类 生人 全人类
Aa01A03= 人手 人员 人口 人丁 口 食指
Aa01A04= 劳力 劳动力 工作者
Aa01A05= 匹夫 个人
Aa01A06= 家伙 东西 货色 厮 崽子 兔崽子 狗崽子 小子 杂种 畜生 混蛋 王八蛋 竖子 鼠辈 小崽子
Aa01A07= 者 手 匠 客 主 子 家 夫 翁 汉 员 分子 鬼 货 棍 徒
Aa01A08= 每人 各人 每位
Aa01A09= 该人 此人
Aa01B01= 人民 民 国民 公民 平民 黎民 庶 庶民 老百姓 苍生 生灵 生人 布衣 白丁 赤子 氓 群氓 黔首 黎民百姓 庶人 百姓 全民 全员 萌
Aa01B02= 群众 大众 公众 民众 万众 众生 千夫

图1 同义词词林部分示例

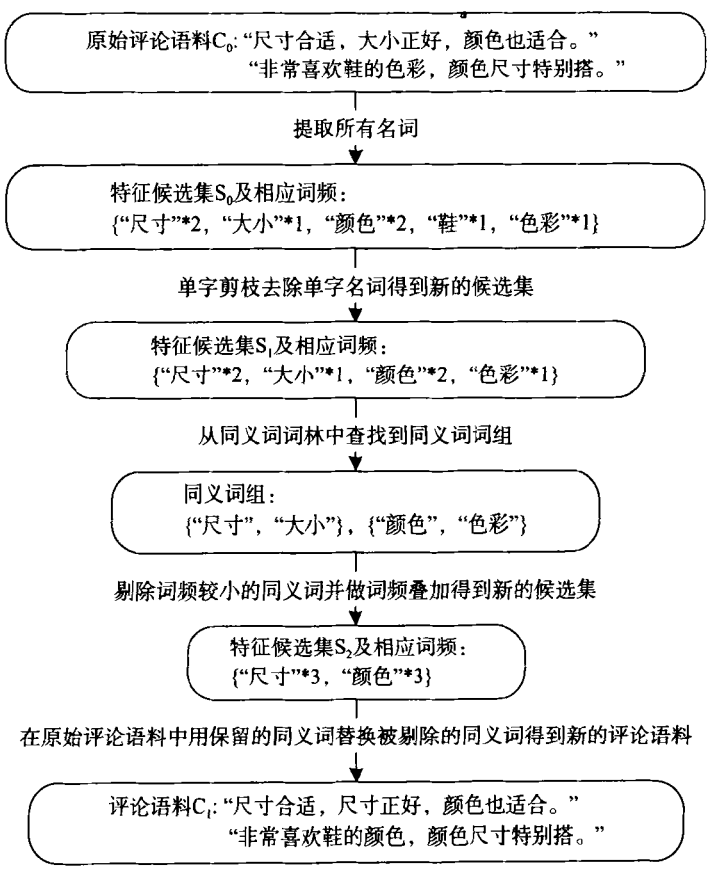

图2 剪枝步骤示例

下一步将对特征候选集 S_2 中的名词进行进一步的操作和选择，最终得到商品的特征词。本步骤包括两部分，第一部分考虑不同特征词有不同的重要程度，于是计算特征词的权重并排序；第二部分利用第一部分的排序结果，并考虑情感词和特征词之间的联系，提出情感指数的概念并得到最后的特征挖掘结果。

商品有各种各样的特征，而这些特征的重要程度是不一样的。对候选特征的权重计算可以从很大程度上分析出哪些候选特征是真正的商品特征，又有哪些候选特征不属于商品特征。本文采用 TF-IDF 权重公式来计算每个候选特征的权重，并对所有的候选特征按照权重的大小进行降幂排序。

TF-IDF 公式早在 1971 年就由 Salton 在向量空间模型中提出[20]，并在文献[21]中做了各种改进模型的比较。此后被广泛应用于特征权重的计算。该公式综合考虑了词语出现的频繁程度和密集程

度,用作特征权重的计算有很好的效果。每个候选特征 t_i 的权重计算公式为:

$$w_i = tf_i \times \log(n/n_i + 0.01) \quad (1)$$

其中,tf_i 表示特征 t_i 的词频,n 表示商品评论总数,n_i 表示特征 t_i 在多少篇文档中出现过。在对每个候选特征计算完权重后,按照权重由大到小的顺序排列候选特征。

在从原始评论语料中获得候选特征集并对其进行处理后,最后一个步骤就是从这些候选特征中选择出商品的特征。考虑到用户对商品的评论归根结底是对商品特征的评论,而评论中用户的主观性评价词语(即情感词)往往出现在商品特征词周围,因而本文提出"情感指数"的概念,利用评论语料中的形容词和副词对商品特征进行选择。假设用户倾向于用情感词来评价商品特征,名词周围出现越多的情感词,那么就越有可能是商品特征。情感指数表示用户评论商品特征时使用情感词(指形容词和副词)的多寡,具体计算情感指数的方法步骤如下:

步骤一:对评论语料 C_1 以逗号","、句号"。"和叹号"!"为分隔符号进行断句,将断句后的所有新句作为新的评论语料 C_2;

步骤二:查找某候选特征 t 在语料 C_2 中的位置 P,若该候选特征所在位置 P_j 上下文中距离(字数)2 以内(包括 2)有形容词,则特征在该位置的情感 p_j 标记为 1,若没有,则转下一步;

步骤三:继续查找该候选词所在位置 P_j 上下文中距离 2 以内(包括 2)有无副词,若有副词且该副词与另一形容词相连,则特征在该位置的情感 p_j 标记为 1,若没有则标记记为 0;

步骤四:将候选特征 t 在所有位置的情感进行加和,再除以该特征的词频 tf,得到的数值即该候选特征的情感指数,即

$$M = \sum p_i / tf, (i = 1, 2, 3, \cdots, tf) \quad (2)$$

用该计算方法得到每个候选特征的情感指数,并以此作为选择商品特征的参考。具体的选择步骤如下:

步骤一:将排列后的候选特征集分成两部分,第一部分是权重较高的前 50% 候选特征,第二部分是权重较低的后 50% 候选特征;

步骤二:对于权值在前 50% 的候选特征,保留情感指数 $M \geq 0.4$ 的候选特征,并剔除情感指数 $M < 0.4$ 的候选特征;

步骤三:对于权值在后 50% 的候选特征,保留情感指数 $M \geq 0.6$ 的候选特征,并剔除情感指数 $M < 0.6$ 的候选特征;

步骤四:将之前两个步骤所保留的候选特征作为最终的商品特征。

5 数值实验和讨论

为了验证本文提出的商品特征挖掘方法的有效性,从天猫商城分别下载了 100 篇 iPhone5 手机、索尼数码相机、飞利浦 MP3 以及纸质图书《冰与火之歌》的买家评论进行数值实验。针对这四部分评论,分别采用人工标注的方法从评论中识别并提取出所有的相应商品特征作为与实验结果的对照。在实验过程中(以手机评论为例),首先将包括"性价比"、"通话音质"、"分辨率"在内的 20 个关键词添加到用户词典中,再对所有评论进行分词和词性标注,接着提取高频名词并剪去单字名词和同义词名词,然后用 TF-IDF 计算每个词的权重,最后引入情感指数筛选出符合要求的词项作为产品特征。表 1 是实验前人工标注的 iPhone5 手机特征同实验结果提取的特征对比。

为了评价本研究商品特征挖掘方法的性能,采用查准率(precision)、查全率(recall)和 F-measure

表 1 人工标注与实验提取的 iPhone5 手机特征对比

	商品名称:iPhone5	特征数量
人工标注特征集合	客服态度,礼品,物流,性价比,性能,手感,质感,电量,价位,包装,运行,卖家描述,正品,配件,通话音质,上网,像素,功能,反应,质量,屏幕,音质,卖家态度,触摸,速度,分辨率,色彩,电池,信号,发货速度,拍照,尺寸,服务态度,外观,散热,画面,游戏,电影,售后,做工,款式,摄像头,系统,操作,服务	45
实验提取特征集合	客服态度,礼品,物流,性价比,性能,手感,质感,电量,价位,包装,运行,卖家描述,配件,通话音质,上网,像素,功能,反应,质量,屏幕,速度,分辨率,卖家态度,色彩,电池,信号,尺寸,服务态度,外观,画面,游戏,电影,款式,摄像头,服务,东西,老板,实体店,上网速度,感觉,苹果,专卖店,通话,态度,整体,容量,网购,人员,情况,货物	50

作为评价指标来综合考量。为了理解这三个指标的意义,表2给出了这三个指标所包含的实验信息和计算方法。

从表2可以看出,较高的查准率要求实验提取的非人工标注的特征数尽可能少,而较高的查全率又要求人工标注出的特征中没有在实验中提取出来的数量尽可能少,这两者是有一定的冲突的,一方的提高很有可能因其另一方的降低,因此还需要引入*F*-measure来平衡查准率和查全率的关系,使二者都保持较高的水平。以手机和数码相机为例,从这三个指标分别评价扩充用户词典步骤和剪枝步骤,结果如表3和表4所示。

表2 评价指标包含的实验信息和计算方法

	人工标注出的特征数	非人工标注的特征数
实验提取出的特征数	a	b
非实验提取的特征数	c	
查准率(precision)	$P=a/(a+b)$	
查全率(recall)	$R=a/(a+c)$	
F-measure	$F=2*P*R/(P+R)$	

表3 扩充词典及剪枝步骤评价(手机)

	A	B	C	D
查准率	63.3%	63.6%	64.8%	70.0%
查全率	68.9%	77.8%	77.8%	77.8%
F-measure	66.0%	70.0%	70.7%	73.7%

表4 扩充词典及剪枝步骤评价(数码相机)

	A	B	C	D
查准率	66.7%	65.2%	64.8%	69.8%
查全率	74.3%	76.9%	76.9%	76.9%
F-measure	70.3%	70.6%	70.4%	73.2%

表3和表4中ABCD分别表示不扩充词典但剪枝(A)、扩充词典但不做单字剪枝(B)、扩充词典但不做同义词剪枝(C)、扩充词典并剪枝(D)。从结果中可以看到,扩充词典后三个指标都有显著的增加,这说明分词时如果能保留更多类似商品特征的候选特征,那么将很大程度上提升结果的有效性;两个剪枝步骤对查全率基本无影响,但是通过去除冗余候选特征来提升查准率,进而提高了F-measure以及实验结果的有效性,这说明从产品的候选特征词适当地剔除无用词汇是很有必要的,而其中的单字特征和同义特征是考虑的重点。

本文将结果与文献[15]中基于关联规则的评论特征选取结果进行比较。本文与文献[15]都用电子商务网站下载的手机和数码相机等四种商品的评论(具体评论内容不同)各100篇用于实验,尽管具体数据不同,但由于形式的一致性,仍有一定的可比性。依然用查准率,查全率和F-measure作为评价指标,比较结果如表5所示。

通过比较可以看出,本文实验在查准率、查全率和*F*-measure评价指标上大都处于更优的位置,只有图书相关特征的查全率处于较低水平,可以认为本文提出的商品特征挖掘方法是有效的。通过实验结果,对基于网络评论的商品特征挖掘方法还有以下几点分析:

(1)用户词典和分词工具对挖掘结果的影响是很大的,如果能在评论特征挖掘前首先建立当前研究的领域词典,将会大大改善挖掘的结果。而对评论的分词以及词性标注的准确性也关系着最终结果的好坏,同一个词语在不同的上下文中有不同的词性,准确的区分并标注是此部分的关键点之一。

(2)初步提取出来的候选特征会含有大量的冗余,如何在保留真正商品特征的前提下将冗余的部分去除是个难点。本文指出的单字名词和同义词属于较为简单的结构,其实还有很多复杂的情况,例如

表5 本文与文[15]结果比较

	手机		数码相机		MP3		图书	
	文献[15]	本文	文献[15]	本文	文献[15]	本文	文献[15]	本文
查准率	63.3%	70.0%	63.6%	69.8%	66.7%	76.9%	62.9%	66.7%
查全率	68.9%	77.8%	73.2%	76.9%	82.4%	81.1%	91.7%	76.9%
F-measure	66.0%	73.7%	68.1%	73.2%	73.7%	78.9%	74.6%	71.4%

"感觉"、"情况"这种宽泛无意义的冗余特征,以及"货物"、"专卖店"这种看似相关却并非评论对象的冗余特征等。

(3)本研究关注的是如何自动地从评论中获取商品特征,但却无法忽视人的因素。评论中个性化的语法句法会影响特征提取的准确性,此类研究更要求评论规范的结构和直接的语义表达;而人工标注特征的时候也带有很强的主观理解,不同的人对评论特征的人工提取也会有一定的差别。

6 结束语

针对网络评论的挖掘研究能够指导实践,反过来促进电子商务的发展。商品特征挖掘作为评论挖掘的重要组成部分,有重要的研究意义。本文提出了一套新的商品特征挖掘方法,其中通过扩充用户词典、引入同义词表来挖掘商品候选特征,并利用TF-IDF计算权重,且在最后考虑到用户情感而提出情感指数作为选择商品特征的依据。实验结果表明本文所提方法在前人的基础上有新的提升,对电子商务领域的具体应用也有现实的指导作用。

本文提出的特征挖掘方法还有一些问题值得深入探讨,例如没有对标注的词性进行不同类别的分析,TF-IDF权重计算公式过于简单,没有从深层语义语法的角度分析用户评论等。针对这些问题,今后将词性标注分析、权重公式改进、评论情感分析等多个方面进行更加深入的研究,从而进一步提升商品特征挖掘的准确性。此外,评论检索作为信息检索的一个具体研究方向在目前的文献中还少有提及,评论检索的模型和方法设计也将是下一步研究的重点。

参 考 文 献

[1] Vermeulen I E, Seegers D. Tried and tested: The impact of online hotel reviews on consumer consideration [J]. Tourism Management, 2009, 30(1): 123-127.

[2] Chen Y, Xie J. Online consumer review: Word-of-mouth as a new element of marketing communication mix [J]. Management Science, 2008, 54(3): 477-491.

[3] 史伟,王洪伟,何绍义. 基于微博的产品评论挖掘:情感分析的方法[J]. 情报学报,2014,33(12):1311-1321.

[4] Hu M, Liu B. Mining opinion features in customer reviews [C]. Proceedings of 19th National Conference on Artificial Intelligence, AAAI, 2004: 755-760.

[5] Hu M, Liu B. Mining and summarizing customer reviews [C]. Proceedings of the 10th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, ACM, 2004: 168-177.

[6] Zhuang L, Jing F, Zhu X Y. Movie review mining and summarization [C]//Proceedings of the 15th ACM International Conference on Information and Knowledge Management, ACM, 2006: 43-50.

[7] Popescu A M, Etzioni O. Extracting product features and opinions from reviews [M]. Natural Language Processing and Text Mining, Springer London, 2007: 9-28.

[8] Scaffidi C, Bierhoff K, Chang E, et al. Red Opal: product-feature scoring from reviews [J]. Proceedings of the 8th ACM Conference on Electronic Commerce, ACM, 2007: 182-191.

[9] Pang B, Lee L. Opinion mining and sentiment analysis [J]. Foundations and Trends in Information Retrieval, 2008, 2(1-2): 1-135.

[10] Pak A, Paroubek P. Twitter as a corpus for sentiment analysis and opinion mining [C]//Proceedings of the Seventh Conference on International Language Resources and Evaluation, 2010, 10: 1320-1326.

[11] 李实,叶强,李一军. 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.

[12] 王继成, 潘金贵, 张福炎. Web文本挖掘技术研究[J]. 计算机研究与发展, 2000, 37(5): 513-520.

[13] 苏金树, 张博锋, 徐昕. 基于机器学习的文本分类技术研究进展[J]. 软件学报,2006,17(9): 1848-1859.

[14] Shi B, Chang K. Mining Chinese Reviews[C]. Sixth IEEE International Conference on ICDM Workshops, IEEE Computer Society, 2006: 585-589.

[15] 李实, 叶强, 李一军, Rob Law, 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, 12(2): 142-152.

[16] 周茜, 赵明生, 扈旻. 中文文本分类中的特征选择研究[J]. 中文信息学报, 2004, 18(3): 17-23.

[17] 郗亚辉. 产品评论特征及观点抽取研究[J]. 情报学报, 2014, 33(3): 326-336.

[18] 张玉芳, 彭时名, 吕佳. 基于文本分类TFIDF方法的改进与应用[J]. 计算机工程,2006,32(19):76-78.

[19] 孙春华, 刘业政. 基于产品特征词关系识别的评论倾向性合成方法[J]. 情报学报, 2013, 32(8): 844-852.

[20] Salton G. The SMART retrieval system—experiments in automatic document processing [M]. Prentice-hall, Inc Upper Saddle River, 1971.

[21] Salton G, Buckley C. Term-weighting approaches in automatic text retrieval [J]. Information Processing & Management, 1988, 24(88): 513-523.

(责任编辑 魏瑞斌)