中文网络客户评论的产品特征挖掘方法研究①

李 实[1], 叶 强[1,2], 李一军[1], Rob Law[2]
(1. 哈尔滨工业大学管理学院, 哈尔滨 150001; 2. 香港理工大学, 中国香港 100085)

摘要: 随着互联网的广泛应用, 在 Blog, BBS, Wiki等网络站点中出现了大量的针对商品或服务的客户评论, 这些客户评论中所包含的丰富信息, 对企业管理具有重要的价值. 通过数据挖掘算法对客户针对某一产品的大量评论进行分析, 可以挖掘出这些产品的主要特征, 并有望进一步发现客户对这些特征的意见和态度. 在英文世界中已经有学者开始对这一研究进行探索, 然而由于语言结构等方面的差异, 英文的研究成果尚无法直接应用于中文客户评论的挖掘中. 本研究针对中文的特点, 提出了面向中文的客户评论挖掘方法. 该方法基于改进关联规则算法实现了针对中文产品评论的产品特征信息挖掘. 本研究采用通过互联网获得的针对手机、数码相机、书籍等 5种产品的评论语料, 对该方法进行了数据实验, 实验结果初步验证了该方法有效性.

关键词: 用户评论; 产品特征; 关联规则; 数据挖掘
中图分类号: TP311 文献标识码: A 文章编号: 1007-9807(2009)02-0142-11

0 引 言

过去十几年中, Internet技术与应用的快速发展不仅给企业的业务流程带来了巨大的变革, 也对消费者的行为模式产生了深刻的影响. 一方面改变了消费者表达对于产品观点和看法的方式——他们可以在销售网站, 网络论坛, 讨论小组, 以及博客(Blog)中撰写产品评论; 另一方面这些产品的"口碑"也反过来影响其他消费者做出购买决策[1,2]. DoubleClick Inc.进行了针对美国服装业、计算机硬件设备业、运动健身产品行业及旅游业网络客户的研究[3], 发现这些行业中, 近一半以上的互联网用户做出购买决定前, 会在互联网上搜索有关产品介绍及商品评论等信息.

但是, 随着网络评论数量的飞速增长, 庞杂的信息使得人工方法难以获得全部客户评论中的有用信息. 因此, 迫切需要借助一定的技术手段来使这一过程变得更为便捷. 近来, 以有效获取网络用户评论信息为目标的非结构化数据分析技术——"评论挖掘"吸引了很多学者关注[4]. 评论挖掘作为非结构化信息挖掘的一个新兴领域, 主要涉及网络评论情感倾向的分析, 评论中产品特征的提取, 以及评论中产品比较信息挖掘等等[5~8]. 消费者可以借助于评论挖掘工具了解产品的性能和其他用户对该产品的态度; 同时, 网络客户评论作为反馈机制, 可以为销售商和生产商提供哪些产品特征是客户所关注的以及客户对于产品的情感倾向分布等信息, 从而可以帮助企业改进产品、改善服务, 获得竞争优势. 面向网络用户评论的产品特征提取研究, 作为评论挖掘的研究方向之一, 旨在从客户评论中挖掘出备受关注的产品特征信息, 并且总结基于这些产品特征的观点, 依靠情感分类(sentiment classification)技术自动得出用户对各个属性的态度倾向, 从而可以为用户提供更为具体和有价值的信息[9]. 在英文世界的评论挖

① 收稿日期: 2007-11-23; 修订日期: 2008-04-25.
基金项目: 国家自然科学基金资助项目(70771032; 70501009); 香港理工大学研究基金资助项目(G-YX93).
作者简介: 李 实(1976-), 女, 黑龙江哈尔滨人, 博士生. Email: shishil@gmail.com

掘领域, 研究者已经初步取得一些成果, 而针对中文网络用户评论的研究还处于起步阶段. 随着我国网络用户群的不断壮大, 中国电子商务的发展也逐渐为世界所瞩目. 截至 2007年 12月, 中国网民人数已经达到 2. 1亿, 预计 2008年将超过美国成为世界第一[10]. 不断增长的中文评论已经成为互联网上一个重要的组成部分, 为了给企业和个人提供更为方便的工具, 自动化和智能化地挖掘中文评论中的有价值信息是非常必要的. 但是由于中英文语言存在着较大的差异, 目前针对英文评论的研究成果很多无法直接应用于中文评论. 这些差异主要根源在如下一些问题上:

(1) 文化差异导致语言表达方式不同. 正如 Rosenzweig[11]曾指出的, 文化差异将导致管理研究的不等价性; 而网络评论这一文本的风格毋庸置疑和商品评论的撰写者——客户的文化背景深刻相关[12].

(2) 语言结构的差异. 例如, 英语中的单词是自然分开的, 而中文评论的分析首先要进行分词处理;

(3) 中英文词汇语法的差别. 与英文评论相比较, 中文词性标注算法更为复杂, 词性标注工具本身的研究还在不断完善中.

本文正是在中英文语言差异存在的条件下, 探索中文网络客户评论的产品特征信息提取技术. 通过对基于关联规则的产品特征挖掘方法进行原理创新和技术拓展, 把目前主要面向英文的评论挖掘方法拓展到中文世界, 从而解决中文环境下, 如何对客户评论中所蕴含的产品特征信息进行自动挖掘的问题. 该方法的应用, 将有望帮助企业和客户在商务过程中更便捷地获取其他客户对相应产品或者服务的反馈信息.

1　相关研究背景

近年来在客户关系管理的研究中有学者强调客户感知利失在影响顾客满意、品牌忠诚和 CRM效果中的作用必将逐渐增大[13]. 而目前网络经济时代中, 从网络评论中挖掘客户的感知利失信息是新兴起的研究领域. 对于网络评论的挖掘问题, 学者 Popescu[4]明确将其分为以下几个子任务: 1) 挖掘重要产品特征; 2) 挖掘用户对于产品特征的主观观点; 3) 判断评论观点的情感导向; 4) 根据观点的重要性进行排名. 相关研究包括评论的情感分析[9], 评论的主客观分析[14]以及评论中商品特征挖掘[4,5,9]等.

1.1　网络客户评论情感分析

情感分析以客户在互联网上发布的产品评论为研究对象, 挖掘客户的情感倾向, 从而自动判断该评论的极性 (the polarity of the review), 即正面评论或负面评论[6]. 通过对大量客户评论的情感分类, 可以综合得出这些客户对该种产品或服务的普遍看法.

目前情感分析技术, 主要包括机器学习方法及语义方法 (semantic orientation)两类[15]. 一些学者已经开始应用这两种方法对英文客户评论的情感分类进行了一定的研究. 最初 Pang[16]在研究中提出来用机器学习的方法进行情感倾向的挖掘工作, 准确率达到 87. 5%, 之后的一些学者在此研究基础上扩展和延伸, 取得了很好的研究结果. Sanjiv[17]等针对 yahoo网站股票留言板中的评论进行了研究, 提取了投资者对其所关注股票的态度. Beineke[18]等用机器学习和人的注释评论相结合, 提高了英文文本情感分析的准确度. Fei[19]等利用机器学习方法, 针对 yahoo网站的英文体育评论研究情感分析.

基于机器学习的情感分类方法在针对每一种产品使用前, 都需要用大量的训练样本对分类模型进行训练, 而训练样本集的建立则需要采用人工方法对大量的评论文章逐一阅读甄别, 并进行手工标识, 这与利用自动情感分类降低人的阅读负担这一初衷还有着一定的差距. 因此, 近来许多研究者将情感分析研究的重点集中在对训练样本的需求量较低的语义方法上. Turney[6]最早提出了基于 PMI-IR算法的语义情感分类思想, 该方法将点互信息 (PMI)与信息汲取方法 (IR)相结合, 借助搜索引擎的后台数据库获得语义倾向信息, 从而做出情感判断, 得到汽车评论的准确率是 84%, 电影评论的准确率是 66%. 其可靠性已经在英文客户情感分类的研究中得到了初步的验证. 2003年, Dave[20]利用该方法对亚马逊 (Amazon)和 C-Net等网上商店的客户评论进行了情感分析, 再次验证了该方法的性能. Zhou[15]

利用电影评论数据对基于语义倾向的情感分类方法和基于机器学习的情感分类方法进行了对比分析,发现语义方法的结果与机器学习方法具有相似性.上述研究均证实了该语义倾向的客户情感分析方法的有效性.除此之外,还有一些学者采用由普林斯顿大学开发的英文词网(wordnet)[21]进行英文语义方法的情感分析,也取得了较好的分析结果[22]. Liu[5]等在对于产品特征挖掘后,针对某一特征的情感导向分析正是利用了英文词网中对于词的语义关系定义.

而由于语言结构的差别,现有的面向英文客户评论情感分类的语义方法,无法直接用于中文客户评论的情感分类.叶强,李一军等探索了中文环境下的情感分析理论与方法,在 PMI-IR方法基础上,初步建立了中文语义倾向情感分析方法,并分别将中文搜索引擎 www.Google.com和 www.Baidu.com提供的 API集成于实验平台中,对手机、图书、电影的中文客户评论进行了情感分析,获得了接近英文同类研究的分析结果,显示出了该方法在中文情感分析上的应用前景[23~26].另外, Yao[27]等在研究中提出了使用电子汉英翻译词典结合英文词网的方法,也是对中文评论情感分析的一个有益尝试.

1.2　网络客户评论的主客观分析

用户的情感倾向主要是通过主观句来表达的,所以在现有的评论挖掘技术中,主观性模式的自动识别与判断是非常重要的基础性技术. Wiebe等[14]针对英文主观情感识别进行了研究,选择某些词类(代词、形容词、序数词、情态动词和副词)、标点和句子位置作为特征,实现对主观句识别的平均准确率 72.17%. Riloff等[28]利用 bootstrapping算法学习得到了 1052个主观性名词,单独使用主观性名词为特征,采用朴素贝叶斯分类器对主观句识别的查准率为 77%,查全率为 64%;如果加上先前确定的主观线索(来自词典和已有的研究结论)和句子的背景信息,那么分类器对主观句判断的查准率和查全率分别能达到 81%和 77%. Riloff和 Wiebe[29]进一步提出了从未经过人工标注的文本中自动提取主观句的方法.他们依靠先前研究中确定的主观特征,分别建立了主观分类器和客观分类器,自动从未标注的文本中获得大量主观句(查准率为 91.5%,查全率为 31.9%)和客观句,再从这些句子中得到更多主观性词语搭配,再用准确性很高词语搭配更新原始的主观特征.通过重复上述过程进一步提高主观分类器和客观分类器的准确率,最终主观分类器的查准率和查全率分别达到 90.2%和 40.1%. Yu和 Hatzivassiloglou[30]利用相似性方法、朴素贝叶斯分类和多重朴素贝叶斯分类 3种统计方法进行主客观句的识别研究.其中,朴素贝叶斯分类器在原有研究的基础上采用词、2-gram、3-gram和词类、具有情感倾向的词序列、主语和其直接修饰成分等作为特征项,对主观句识别的查准率和查全率达到 80%~90%,而客观句的查准率和查全率大约在 50%左右.叶强等探索了中文主观性的自动识别方法,提出了基于连续双词词类组合模式(2-POS)的主观程度自动判别算法,为中文客户评论挖掘提供了一种可能的方法选择[31].

1.3　网络客户评论中的产品特征挖掘

网络客户评论中的产品特征挖掘是指通过机器从大量的网络客户产品评论中自动地获取所关注的产品特征[5],这项技术是分析用户对于产品具体特征所持情感倾向的前提,其准确性和全面性是非常重要的.对于英文评论中的产品特征挖掘研究已经取得了一些成果. Hu[5]和 Liu等[32]首先提出应用关联规则分类方法提取英文评论中的产品特征,利用该方法对于包括手机,数码相机等产品评论进行挖掘,平均查全率达到 80%,平均查准率达到 72%,而且他们在此基础上进行了后续的研究,判断用户对这些特征的观点以及情感导向.也有一些研究人员采用了其他方法实现这一功能,比如 Kobayashi[33]采用了半自动化的循环方法提取产品特征和用户观点,但是需要大量的人工参与; Popescu等[4]利用了 Etzioni研发的 kownital系统,计算点互信息值(PMI),然后进行贝叶斯分类,从而提取产品特征,虽然提高 Hu的准确率(平均提高了 22%),但是查准率却有所下降(平均下降了 3%).另外, Liu[8]重点研究存在多种产品互相比较内容的评论,这与 Liu等挖掘同一产品的重要特征在研究内容上有些差别;而且与 Popescu所提出的技术类似,其中对于产品特征的提取采用有导师学习方法(supervised training),需要建立一个产品特征集合,以及产品

相关领域的训练样本集, 而训练样本集的建立则需要人为对这些大量的评论文章进行逐一阅读, 这与自动评论挖掘的目的有些矛盾.

目前, 对于英文评论的产品特征挖掘中, Hu[5]和 Liu[9]等提出的基于关联规则的方法主要步骤为:

第 1步, 标注词性;

第 2步, 将名词和名词短语组成事务文件 (transaction file);

第 3步, 基于关联规则分类方法提取频繁规则项产生候选特征项集合;

第 4步, 对于特征项进行邻近规则剪枝. 邻近的定义为: 假设 是频繁规则项, 而且 包含 n个单词, 假设一个句子 包含 , 而且在 中的词出现在 中的顺序为: w_1, w_2, …, w_n. 假设 中任何两个相连的单词 (w_i和 w_{i+1}) 的距离不超过 3个单词, 则可以说 在 中是邻近的. 如果 出现在评论数据库中的 m个句子中, 而且至少在 2个句子中是邻近的, 就可以称 是一个邻近的特征短语. 非临近的特征短语将不是需要的产品特征;

第 5步, 对于特征项进行独立支持度剪枝, 形成频繁特征项所构成的产品特征集合. 独立支持度的定义为: 特征 的独立支持度 (p-support) 是包含 而且句子中不包 的父集作为特征的句子的数量. 在 Hu的研究中采用最小的独立出现支持度为 3, 也就是说如果一个特征的独立支持度小于 3, 那么就从候选特征集合里面去掉;

第 6步, 补充评论中非频繁特征项的产品特征.

目前尚缺乏对于中文网络客户评论的产品特征挖掘研究. 虽然英文中相关研究已经得到了有效的验证, 但是无法直接应用于中文, 其根本原因是引言中所提到的中文和英文语言特点及文化背景不同, 具体有下面几个技术困难:

(1) 中文在进行语言处理中首先需要进行中文分词;

(2) 中文词性标注也和英文有差别, 特别是中文语言比较复杂, 有些单词的词性随着语言环境的变化可能会发生转化, 而形式上却没有变化;

(3) 在英文方法中标注词性的过程中就可以标出名词短语, 而对于中文名词短语的定义则非常复杂, 词性标注工具只能标注一些专有名词短语;

(4) 中英文语言表达中, 名词短语的构成不同. 例如英文中过去分词 +名词可以表示名词性短语, 中文中没有这种形式, 但是具有其他形式;

(5) 中文中有字的概念, 而英文没有. 中文名词可以由一个或者一个以上的字构成, 这样对于表达产品特征的名词可能具有其特殊的规律.

本文将参考 Hu等学者基于关联规则分类的产品特征挖掘算法, 针对中文评论的语言特点和风格特征, 解决上面的技术困难, 探索面向中文网络客户评论中的产品特征挖掘方法和理论, 并且通过实验验证这一方法的有效性.

2　中文网络客户评论产品特征挖掘方法

2.1　方法具体内容

本文所提出的中文网络客户评论的产品特征挖掘技术, 由以下 8个步骤构成.

步骤 1　对评论语料进行分词.

本文采用中国科学院计算机所软件室编写的中文分词工具 ICTCLAS (institute of computing technology, Chinese lexical analysis system) (http://mtgroup.ict.ac.cn/), 对评论文本语料进行分词.

步骤 2　对分词后的评论语料进行词性标注.

同样采用 ICTCLAS工具. 词性标注方法可以根据需要进行一级或者二级标注, 其差别在于: 一级只标注名词, 动词等; 二级可以标注出更为具体的情况, 包括具有名词功能的形容词或者动词, 专有名词, 词素等等. 为了提高挖掘查准率, 采用二级标注.

中文客户评论中所讨论的商品特征可能由名词短语构成, 但是值得注意的是, 中文评论的词性标注过程中并没有直接标注出名词短语 (除了专有名词短语以外, 例如地名、单位名称), 所以需要对于基本名词短语进行人为界定. 在中文语料学研究中, 基本名词短语的定义有一些不同, 本文根据周雅倩等[34]的定义: 基本名词短语为非嵌套的名词短语, 它包括单个名词、没有任何修饰成分

的名词短语、难以确定修饰关系的一串名词、并列名词性成分、专有名词、时间、地点等,这种基本名词短语占语料中所有基本短语的 60.8% (用 Chinese treebank做统计).很显然,专有名词和时间、地点名词一般情况下不是普通产品特征 (对于一些特殊商品的特点挖掘可能需要,比如旅游目的地),所以在本文中,名词短语将按照以下两种情况界定:

(1)由两个或三个相邻的名词所连接成的短语 (不包含专有名词和时间、地点名词,但包含具有名词功能的形容词或者动词);

(2)两个名词之间仅用结构助词 "的" 连接成的短语.

本文根据这两种情况提出了中文网络评论中基本名词短语的提取模式,如表 1所示,其中名词不包含专有名词和时间、地点名词,但是包含二级分词标注出来具有名词功能的形容词或者动词.在应用标注工具进行词性标注后,再按照这几种模式提取出基本名词短语.

表 1　中文基本名词短语提取模式

Table 1　Extracting patterns of Chinese basic noun phrase

序号	第 1个词	第 2个词	第 3个词
1	名词	名词	不是名词
2	名词	名词	名词
3	名词	助词 "的"	名词

步骤 3　利用词性标注后的评论语料创建关联规则事务文件 (transaction file).

本文所提出的方法基于关联规则分类算法,需要对于文本评论进行形式化预处理.所以首先需要建立事务数据库,这里事务数据库以文本文件的形式存储.在这一步骤中以句子为事务单位,提取评论中的所有名词或者基本名词短语作为项 (item)构成一个事务文件,为下面提取频繁项集 (frequent itemset)做好数据准备.

步骤 4　基于关联规则 Apriori算法找到频繁项集作为候选产品特征集合 I_0.

一般来讲关联规则的挖掘分为两步:一是找出所有的频繁项集,这些项集出现的频繁性至少和预定义的最小支持计数 (min support count)一样;二是由频繁项集产生强关联规则.对于评论中产品特征的挖掘研究只用到第一步,挖掘出满足最小支持度的频繁规则项,作为商品的候选特点[35].利用 Apriori算法从上一步所生成事务文件中找到频繁项集作为候选的商品特征集合 I_0,采用的最小支持度为 1% (参考英文评论处理方法);3项以上的频繁项可以很明显的看出不是产品特征,这一特点在英文评论的商品特点挖掘中也是一样的,采用同类研究的解决办法,不考虑 3项以上的频繁项[5].

步骤 5　将候选产品特征集 I_0按照邻近规则剪枝,成为候选特征集 I_1.

参考英文邻近规则定义,可以定义中文评论中的邻近规则.

定义 1　在中文评论中,假设 f是频繁规则项,而且 f包含 n个名词 (或名词短语),假设一个句子 s包含 f,而且在 f中的词 (或名词短语)出现在 s中的顺序为: w_1, w_2, …, w_n.假设 s中任何两个相连的名词 (或名词短语) w_i和 w_{i+1}的距离不超过 3个词 (根据中文分词结果),则可以说 f在 s中是邻近的.

例如下面三句话:

"这款手机功能非常强大."

"摄像功能已经成为重要的手机功能之一."

"作为一款女士手机,外观是非常重要的,而一些商务功能则不是必须的."

对于 "手机功能" 这一候选特征, "手机" 和 "功能" 这两个词在前两句话中满足邻近规则,最后一句话中不满足.但已经在两句话中邻近,可以说 "手机功能" 是一个邻近的特征名词短语.

在这一步骤中,遍历每一个名词短语、2—项和 3—项频繁项,如果 f出现在评论数据库中的 m个句子中,而且至少在 2个句子中是邻近的,就可以称 f是一个邻近的特征名词短语,加入到候选项集合 I_1中.

步骤 6　将候选产品特征集 I_1按照独立支持度规则进行修正,形成候选特征集 I_2.

参考英文独立支持度的定义,可以定义中文评论中的独立支持度:

定义 2　在中文评论中名词或者基本名词短语 f_t的独立支持度 (p_support)是包含 f_t的而且句子中不包 f_t的父集作为频繁特征项的句子数量.

例如 "屏幕" 作为频繁项,出现的句子为 10

个, “屏幕分辨率”, “屏幕效果”也是频繁项, 它们出现的次数分别为 3和 4, 则“屏幕”的独立支持度为 3.

本文采用最小的独立支持度为 3, 即一个特征项的 p_support< 3, 那么这个特征项就从候选特征集合里面去掉. 过滤掉所有不满足独立支持度要求的候选特征项, 形成新的候选特征集合 I2.

步骤 7　建立常见中文频繁项名词却非产品特征的集合, 将 I2过滤形成特征集合 I3.

常见的中文名词或者名词短语而确定非产品特征在本研究中主要划定为以下几种的情况:

1)在候选特征项中去掉关于表示商品型号的名词, 第 1位为字母后面全部为数字的名词例如“N70”.

2)常见商品的品牌. 例如对于某型号手机产品特征的挖掘可以排除掉“诺基亚”, “摩托罗拉”等名词.

3)一些常见的口语化名词. 例如: “机子”, “东西”.

4)一些常见的人称名词. 例如“朋友”, “先生”.

步骤 8　从 I3中去掉单字名词的候选项, 包括含单字名词的 m项频繁项(m≤ 3), 形成最后的产品特征集合 I.

在中文中, 一个单字可以标注为名词, 这是中文所特有的情况. 从中文评论中产品特征的人工标注结果就可以看到, 基本上不用单字名词作为特点的名称. 而且在后面的数据实验中, 采用 5样产品, 共挖掘出来属于产品特征的有 139项, 而其中是单字名词或者包含单字名词的特征一个也没有. 而去掉候选特征集中只有一个单字的名词, 例如“手”, “信”等会大大提高挖掘的准确率. 这一步骤所带来的效果将在下面的数据实验中得到验证.

2.2　非频繁特征项的处理

正如前面提出的方法步骤中所介绍的, 本文和 Hu[9] 的英文评论挖掘研究都基于关联规则的频繁特征项挖掘技术. 对于非频繁项的产品特征处理, 在英文评论的方法中最后进行了补充. 所应用的方法为: 找到所有修饰频繁特征项的形容词作为句子的主观观点, 形成用户观点数据库, 然后再重新回到所有评论中遍历, 如果一个句子中的形容词是主观观点则离它最近的名词或者名词短语补充为非频繁特征项产品特征. 但通过这一步并不能使查全率和查准率都有所提高. 这是因为补充的非频繁特征项产品特征可能与用户讨论的商品对象没有关系, 导致了准确率的降低. 但是他们考虑到这种非频繁项的数量比较小, 对于用户购买决策影响不大, 所以为了提高结果的查全率以及挖掘方法的综合性能而增加了这一步.

但是对于中文网络评论中的产品特征挖掘, 是否补充非频繁项为产品特征需要针对中文评论的特点进行分析. 本文提出的中文评论挖掘方法性能结果通过实验验证(详见后面的数据实验结果)是查全率比较高, 而查准率比较低. 如果增加非频繁项作为产品特征使得查全率和查准率的差异更为增大, 整体性能会降低; 而且错误非频繁项特征的产生即使对于用户决策影响比较小, 也还是有影响. 所以在中文网络客户评论挖掘方法中补充非频繁项的产品特征带来的负面作用比较大, 并不适合. 本方法中将不考虑补充非频繁项作为产品特征.

3　数据实验

3.1　语料数据

本文选取了 5种商品的网络评论作为实验语料进行数据实验, 这 5种商品分别是一款手机(Nokia N70), 两款数码相机(Canon A710, Canon 850), 一款 MP3播放器(魅族 E3)和一本图书(《达芬奇的密码》). 其中手机, 数码相机及 MP3播放器的评论从 it168网站下载(http://www.it168.com), 图书评论从卓越网下载(http://www.joyo.com.cn). 每样商品各选取 100篇评论, 针对每一种商品的全部评论, 用人工标注的方法对这些评论中所提到的该商品属性进行识别和标注. 根据最小最大覆盖原则建立最小的属性集合, 使这个集合可以覆盖所有这 100个评论中提到的该商品的属性. 以手机为例, 手机(Nokia N70)的商品属性集合如表 2所示.

表 2 手机(Nokia N70)属性的人工标注结果

Table 2 The manual features of mobile phone (Nokia N70)

商品名称	人工标注属性集合	人工标注属性数量
手机(Nokia N70)	屏幕, 软件, 电池, 体积, 游戏, 外形, 输入, 字库, 收音机, 内存, 语音, 摄像, 按键, mp3, 多媒体, 耳机, 待机时间, 键盘, 铃声, 拍照, 速度, 系统, 功能, 摄像头, 手感, 售后服务, 声音, 机身, 价格, 接口, 电话簿, 菜单, 语音拨号, 版本, 快捷键, 兼容性, 闪光灯, 充电器, 质量, 智能, 屏幕效果, 桌面, 运行速度, 音质	45

3.2 性能评估方法

按照前面提出的方法采用 JAVA语言构造实验系统. 为了评估挖掘方法的性能, 本文采取了在文本处理问题研究中普遍使用的性能评估指标: 查全率(recall), 查准率(precision). 本文中的研究问题为判断所挖掘的产品属性是否为人工标注的真实属性, 这可以归结为二值分类, 评估一般使用 2维列联表(contingency table).

实验所采用的列联表如表 3所示. 这里真实产品属性数即人工标注结果的属性数量, 其中作为挖掘性能度量的查全率和查准率计算方法如下:

$$\text{查准率(precision)} = \frac{A}{A+B}$$

$$\text{查全率(recall)} = \frac{A}{A+C}$$

表 3 评估方法性能的列联表

Table 3 The contingency table for performance of experiment

	真正产品属性数	非真正产品属性数
本文方法挖掘出来的产品属性数	A	B
本文方法没有挖掘出来的产品属性数	C	D

3.3 实验结果

综合 5种商品的实验结果(如表 4所示), 平均查全率 77.8%, 平均查准率 63.6%, 说明本研究所提出方法具有一定有效性. 从表 4可以看到通过去掉单字名词候选项, 查准率获得了大幅度的提高. 为了深入验证方法的实际性能, 需要进行中英文客户评论产品特征挖掘结果的差异显著性检验.

表 4 实验结果

Table 4 The experimental results

商品名称	人工标注属性数	未去除中文单字属性查准率	去除中文单字属性查准率	查全率
手机(Nokia N70)	45	56.4%	63.3%	68.9%
数码相机(Cannon A710)	41	50.8%	61.1%	80.5%
数码相机(Cannon 850)	38	44.6%	64.1%	65.8%
MP3播放器(魅族 E3)	34	52.8%	66.7%	82.4%
书籍(《达芬奇密码》)	24	51.2%	62.9%	91.7%
平均值	36	51.2%	63.6%	77.8%

3.4 差异显著性检验

对于英文评论的产品特征挖掘, Hu等的研究结果被验证为有效, 并且得到相关研究领域的承认, 为了进一步确认本文所提出方法的有效性, 将本文研究结果和 Hu等人的研究结果作以比较, 并对两者差异做显著性检验, 如果两个结果接近(即本文结果显著好于或与 Hu的研究结果的差距不明显)则可以进一步验证本方法的有效性.

检验过程为利用本文所提出的方法, 数据采用与 Hu实验数据中的相同种类和数量商品评论进行特征挖掘, 最后将实验结果与 Hu的实验结果即查准率和查全率分别进行差异 T检验, 同时考虑了与分类随机比率 50%的差异检验. 本研究所使用的具体商品以及评论内容数据和前面的验证实验相同, 例如手机类采用的是 Nokia N70的评论数据; Hu等对应每一类商品的实验结果从文

献[9]中得到;所利用的评论都为 100篇. 结果列于表 5、表 6和表 7.

对于查准率的比较结果显示, 在去掉单字属性之前, 本文挖掘方法的查准率比较差 (如表 5所示), 与英文实验结果差异在 0.01水平上显著, 与随机比率差异不显著, 很难令人满意. 但是经过去除中文单字属性的改进后 (如表 6所示), 查准率取得了大幅度提高, 从 51.2% 提高到 63.8%, 与 Hu等的实验结果基本一致 (查准率差异在 0.01水平上不显著).

从总体数据试验结果 (表 6和表 7)可以看到, 本文和 Hu等的平均实验结果查准率差异在 0.05水平上显著, 0.01水平上不显著, 查全率差异不显著, 说明本文所提出的对于中文网络用户评论挖掘方法和英文评论挖掘的方法基本上性能差异不大, 进一步验证了本文方法的有效性.

表 5　查准率差异检验结果 (未去除中文单字属性)

Table 5　T-test results of precision (with features of single Chinese character)

产品类别	Hu和 Liu实验的查准率	本文的查准率	自由度	P值	与随机比率 0.5比较的 P值
手机	0.718	0.564	53	0.014 1*	0.346 9
数码相机 1	0.71	0.508	63	0.000 6**	0.897 8
数码相机 2	0.71	0.446	54	0.000 0**	0.422 7
MP3播放器	0.692	0.528	51	0.012 6*	0.685 0
平均值	0.708	0.512	227	0.000 0**	0.717 7

注: **, * 分别表示结果在 0.01和 0.05水平上显著, 没有 * 表示不显著.

表 6　查准率差异检验结果 (去除中文单字属性)

Table 6　T-test results of precision (without features of single Chinese character)

产品名称	Hu和 Liu实验的查准率	本研究的查准率	自由度	P值	与随机比率 0.5比较的 P值
手机	0.718	0.633	47	0.192 6	0.068 9
数码相机 1	0.71	0.611	52	0.115 0	0.108 7
数码相机 2	0.71	0.641	37	0.348 3	0.086 5
MP3播放器	0.692	0.667	40	0.727 4	0.036 4*
平均值	0.708	0.638	182	0.038 2*	0.000 2**

注: **, * 分别表示结果在 0.01和 0.05水平上显著, 没有 * 表示不显著.

表 7　查全率差异检验结果

Table 7　T-test results of recall

产品名称	M. Hu和 B. Liu实验的查全率	本研究的查全率	自由度	P值	与随机比率 0.5比较的 P值
手机	0.761	0.689	43	0.263 5	0.015 0*
数码相机 1	0.792	0.805	39	0.838 6	0.000 4**
数码相机 2	0.792	0.658	36	0.049 3*	0.059 2
MP3播放器	0.818	0.824	32	0.371 2	0.000 7**
平均值	0.791	0.744	156	0.148 2	0.000 0**

注: **, * 分别表示结果在 0.01和 0.05水平上显著, 没有 * 表示不显著.

4　结果讨论

通过上面的数据实验, 证明了本文所提出的方法是有效的, 在面向中文网络评论的产品特征挖掘领域进行了初步的理论探索和实践检验. 为了进一步提高方法的性能需要了解偏差产生的原因, 可以从下面几点进行分析:

(1) 分词工具对于结果的影响 分词工具是中文自然语言处理的基础, 对于本文准确率起到了很大的作用, 但是目前中文分词工具本身还有一定的误差.

(2) 词性标注对于结果的影响 中文文本的词性标注同样是后面算法的基础, 而词性标注工具本身也是有误差的. 例如 "售后服务"是一个很重要的商品特点, 但是在词性标注的时候标注为 "售后 /名词, 服务 /动词", 这样对于后面的特点挖掘来讲, 无法识别出来这样非名词词组. 另外对于中文名词短语的挖掘非常复杂, 也使得创建事务文件的时候产生偏差.

(3) 人工标注和产品特征概念内涵对于结果的影响 对于评论中的产品特征内涵概念目前还没有比较准确的定义. 在很多文献中按照特点出现的方式分为两类: 一类是显性的, 也就是在评论中比较明确的可以提出来的比如手机的外形, 酒店的房间; 另外一类为隐性的, 是评论中所隐含的特点, 比如 "小孩子也能用", 就是指手机的操作简单. 在本文中主要针对评论中的显性特点. 另外本数据实验结果对照的是人工标注的特征, 尽管通过一些原则处理了人工标注的结果, 但是对于产品特征内涵的主观理解仍然可能会影响标注的客观结果从而使得实验结果统计产生偏差.

(4) 网络用户评论的这一文体本身的风格特征对于分类结果的影响 根据文体理论 (genre theory), 在线产品评论是新的文体风格[12], 它包含的 3种内容会对挖掘方法的性能有所阻碍, 分别是超级链接, 求助性的疑问句, 以及全文引用的其他评论内容. 另外语言特点方面, 在线用户评论会用很多符号表示感情色彩, 例如 "太棒了!!!!", 这样在挖掘过程中会影响到结果的准确.

5 结 论

互联网上大量的客户评论内容中存在着很多有价值信息, 特别是用户集中关注的产品特征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. 评论中产品特征的提取是网络评论挖掘的基础性关键技术, 尽管这一问题在英文中已经开展了一些研究, 然而面向中文客户评论的产品特征挖掘研究目前仍很不足.

本文从中文语言特点和中文评论风格出发, 拓展了基于关联规则的英文评论产品特征挖掘方法, 通过构建中文短语提取模式, 定义中文评论中的临近规则和独立支持度概念, 以及针对中文单字名词等语言结构特点采取改进措施等一系列技术创新, 提出了包含八个步骤的面向中文网络客户评论的产品特征挖掘方法, 从理论上对中文客户评论产品特征挖掘问题进行了初步的探索. 在数据试验中, 该方法的平均查全率为 77.8%, 平均查准率为 63.6%, 这一结果与其他研究者针对英文评论的研究结果基本一致, 表明了该方法的有效性. 该方法的应用, 将有望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.

本文还深入分析了目前算法查准率不够理想的原因, 指出今后的研究将进一步解决这些影响挖掘效果的问题, 从而提高挖掘准确率. 此外, 针对所得到的商品特点, 进行情感倾向分析, 进一步判断中文用户评论中对于具体商品特点的情感倾向分布, 也将是今后的研究重点.

参 考 文 献:

[1] Senecal S, Nantel J. The Influence of Online Product Recommendations on Consumers' Online Choices[C]. Journal of Retailing. Elsevier, 2004: 159—169.

[2] Chevalier J, Mayzlin D. The Effect of Word of Mouth on Sales: Online Book Reviews[C]. NBER Working Paper Series 10148. National Bureau of Economic Research, USA, 2003.

[3] Godes D, Mayzlin D. Using online conversations to study word-of-mouth communication[J]. Marketing Science, 2004, 23(4): 545—560.

[4] Popescu A-M, Etzioni O. Extracting Product Features and Opinions From Reviews[C]. In Proceedings of HLT-EMNLP 2005. ACL, 2005: 339—346.

[5] Hu M, Liu B. Mining Opinion Features in Customer Reviews[C]. In AAAI, 2004: 755—760.

[6] Turney P D. Thumbs up or Thumbs Down: Semantic Orientation Applied to Unsupervised Classification of Reviews[C]. Proceeding of Association for Computational Linguistics 40th Anniversary Meeting, 2002. 417—424.

[7] Liu B. Opinion Observer: Analyzing and Comparing Opinions on The Web[C]. Proceedings of The 14th International World Wide Web Conference (WWW—2005), 2005. 10—14.

[8] Liu J, Wu G, Yao J. Opinion Searching in Multi-product Reviews[C]. Proceedings of The Sixth IEEE International Conference on Computer and Information Technology (CIT'06), 2006. 25—25.

[9] Hu M, Liu B. Mining and Summarizing Customer Reviews[C]. Proceedings of The Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2004. 168—177.

[10] 中国互联网络发展状况统计报告[R]. 中国互联网信息中心(CNNIC), 2008. 1.
Statistical Reports on the Internet Development in China[R]. China Internet Network Information Center, Jan. 2008. (in Chinese)

[11] Rosenzweig P M. National Culture and Management[M]. Harvard Business School: Harvard Business School Pub. Division, 1994.

[12] Pollach I. Electronic Word of Mouth: A Genre Analysis of Product Reviews on Consumer Opinion Web Sites[C]. Proceedings of the 39th Hawaii International Conference on System Sciences, 2006.

[13] 王永贵, 韩顺平, 邢金刚, 等. 基于顾客权益的价值导向型顾客关系管理——理论框架与实证分析[J]. 管理科学学报. 2005, 8(6): 27—36.
Wang Yong-gui, Han Shun-ping, Xing Jin-gang, et al. Value oriented customer relationship: Conceptual framework and empirical management based on customer equity: Conceptual framework and empirical analysis[J]. Journal of Management Sciences in China, 2005, 8(6): 27—36. (in Chinese)

[14] Wiebe J M. Learning Subjective Adjectives from Corpora[C]. Proceeding of 17th National Conference on Artificial Intelligence. Menlo Park, California: AAAI Press, 2000. 735—740.

[15] Chaovalit P, Zhou L. Movie Review Mining: A Comparison between Supervised and Unsupervised Classification Approaches[C]. Proceedings of The 38th Annual Hawaii International Conference on System Sciences, 2005. 112.3.

[16] Pang B, Lee L, Shivakumar Vaithyanathan. Thumbs up: Sentiment Classification Using Machine Learning Techniques[C]. 2002 Conference on Empirical Methods in Natural Language Processing (EMNLP 2002), 2002. 79—86.

[17] Sanjiv R D, Chen M Y. Yahoo! For Amazon: Sentiment Parsing from Small Talk on The Web[C]. Proceedings of The 8th Asia Pacific Finance Association Annual Conference, 2001.

[18] Beineke P, Trevor H, Shivakumar Vaithyanathan. The Sentimental Factor: Improving Review Classification via Human-Provided Information[C]. Proceedings of ACL, 2004. 263—270.

[19] Fei Z C, Liu J, Wu G F. Sentiment Classification Using Phrase Patterns[C]. In: Proceedings of The Fourth International Conference on Computer and Information Technology (CIT'04). Wu Han, China: IEEE, 2004. 1—6.

[20] Dave K, Lawrence S, Pennock D M. Mining The Peanut Gallery: Opinion Extraction and Semantic Classification of Product Reviews[C]. Proceeding of 12th International Conference on World Wide Web, Budapest, Hungary: ACM Press, 2003. 519—528.

[21] Miller G A. WordNet: A lexical database for English[J]. Communications of The ACM, 1995, 38(11): 39—41.

[22] Andreevskaia A, Sabine B. Mining WordNet for Fuzzy Sentiment: Sentiment Tag Extraction from WordNet Glosses[C]. Proceedings of The 11th Conference of The European Chanpter of The ACL (EACL'06), April 2006. 209—216.

[23] Lin B, Lu T, Ye Q. Opinion Classification for Chinese Movie Reviews[C]. Proceeding of 12th International Conference on Management Science and Engineering, 2005.

[24] Ye Q, Li Y J, Zhang Y W. Semantic-oriented sentiment classification for Chinese product reviews: An experimental study on the reviews for books and cell phones[J]. Tsinghua Science and Technology, 2005, 10(z1): 797—802.

[25] Ye Q, Lin B, Li Y J. Sentiment Classification for Chinese Reviews: A Comparison between SVM and Semantic Approaches[C]. The 4th International Conference on Machine Learning and Cybernetics (ICMLC2005) (IEEE), 2005, 4(8): 2341—2346.

[26] Ye Q, Shi W, Li Y J. Sentiment Classification for Movie Reviews in Chinese by Proved Semantic-Oriented Approach[C]. Proceedings of the 39th Annual Hawaii International Conference on System Sciences, 2006.

[27] Yao J X, Wu G F, Liu J, et al. Using Bilingual Lexicon to Judge Sentiment Orientation of Chinese Words[C]. Proceedings of The Sixth IEEE International Conference on Computer and Information Technology (CIT'06), 2006.

[28] Riloff E, Wiebe J, Wilson T. Learning Subjective Nouns using Extraction Pattern Bootstrapping[C]. Proceedings of The Seventh Conference on Computational Natural Language Learning (CoNLL—03), 2003: 25—32.

[29] Riloff E, Wiebe J. Learning Extraction Patterns for Subjective Expressions[C]. Proceedings of The 2003 Conference on Empirical Methods in Natural Language Processing, EMNLP—03, 2003: 105—112.

[30] Yu H, Vasileios Hatzivassiloglou. Towards Answering Opinion Questions: Separating Facts from Opinions and Identifying the Polarity of Opinion Sentences[C]. Proceedings of The 2003 Conference on Empirical Methods in Natural Language Processing, 2003: 129—136.

[31] 叶 强, 张紫琼, 罗振雄. 面向互联网评论情感分析的中文主观性自动判别方法研究[J]. 信息系统学报, 2007, 1(1): 79—81.
Ye Qiang, Zhang Zi-qiong, Rob Law. Automatically measuring subjectivity of Chinese sentences for sentiment analysis to reviews on the internet[J]. China Journal of Information Systems, 2007, 1(1): 79—81. (in Chinese)

[32] Liu B, Hu M Q, Cheng J S. Opinion Observer: Analyzing and Comparing Opinions on The Web[C]. Proceedings of The 14th International World Wide Web Conference (WWW—2005), 2005: 342—351.

[33] Kobayashi N, Inui K, Matsumoto Y, et al. Collecting Evaluative Expressions for Opinion Extraction[C]. In Proceedings of The 1st International Joint Conference on Natural Language Processing, Sanya City, Hainan Island, China, 2004: 584—589.

[34] 周雅倩, 郭以昆, 黄萱菁, 等. 基于最大熵方法的中英文基本名词短语识别[J]. 计算机研究与发展, 2003, 40(3): 440—446.
Zhou Ya-qian, Guo Yi-kun, Huang Xuan-qing, et al. Chinese and English BaseNP recognition based on a maximum entropy model[J]. Journal of Computer Research and Development, 2003, 40(3): 440—446. (in Chinese)

[35] Liu B, Hsu W, Ma Y. Integrating Classification and Association Rule Mining[C]. KDD-98, 1998: 80—86.

Mining features of products from Chinese customer online reviews

LI Shi[1], YE Qiang[2], LI Yi-jun[1], LAW Rob[2]

1. School of Management, Harbin Institute of Technology, Harbin 150001, China;
2. Hong Kong Polytechnic University, Hung Kom, Kowloon, Hong Kong, China

Abstract: Nowadays, more and more customers read online reviews on products before making the decision of purchase. It is also a common practice for merchants and manufacturers to get useful feedback from reviews written by their customers on products and associated services. Therefore, mining features of products from online reviews has emerged to be an important research topic. However, most present studies focused mainly on English reviews. As China becomes a potential e-commerce market in the world, Chinese have already been to become one of the most important group of customers. The numbers of Chinese online reviews increased greatly in recent years. It makes the study of features retrieved from Chinese reviews imperatively important. The techniques used for English reviews can hardly be applied directly to Chinese reviews due to the differences in characteristics between these two languages. Association rule based methods for products' feature retrieving to English reviews were modified and improved to fit Chinese reviews in this study. Experiments demonstrate the validity of this new method.

Key words: customer reviews; product features; association rule; data mining