第37卷 第12期
Vol.37 No.12

计 算 机 工 程
Computer Engineering

2011年6月
June 2011

·软件技术与数据库·

文章编号：1000—3428(2011)12—0026—03 文献标识码：A 中图分类号：TP393

中文产品评论中属性词抽取方法研究

栗春亮，朱艳辉，徐叶强

(湖南工业大学计算机与通信学院，湖南 株洲 412008)

摘 要：针对现有属性词抽取方法的准确率和覆盖率偏低问题，利用百度百科和分词后相邻词语同现比例识别专业领域生词，降低分词错误对属性词识别的影响，在中文产品评论语料中通过设计词性序列模板获得候选属性词集，该词性序列模板包含名词和名词短语模板、动词和动词短语模板，采用统计技术和自然语言处理技术筛选候选属性词。实验结果表明，对于3 623篇手机评论文章，利用该方法可获得1 732个属性词，准确率为0.565、召回率为0.726、调和平均值为0.636，具有较好的抽取性能。

关键词：产品评论；生词识别；序列模板；属性词

Research of Attribute Word Extraction Method in Chinese Product Comment

LI Chun-liang, ZHU Yan-hui, XU Ye-qiang

(Institute of Computer & Communication, Hunan University of Technology, Zhuzhou 412008, China)

【Abstract】Aiming at solving problems of relatively low precision, rate of coverage when using existing attribute word extraction methods, this paper adopts Baidu Baike and co-occurrence proportion of adjacent words after word segmentation to identify new domain words, decreases impact on recognition of attribute word caused by segmentation errors. This paper designs part of speech sequence templates which contain noun and noun phrase templates, verb and verb phrase templates to obtain attribute word candidates from Chinese product comments, then utilizes statistical technique and natural language processing technique to filter attribute word candidates. Experimental results show that for the 3 623 mobile phone comments, this method obtains 1 732 attribute words, the precision, recall and f-measure reach 0.565, 0.726 and 0.636, and it has good extraction performance.

【Key words】product comment; new word recognition; sequence template; attribute word

DOI: 10.3969/j.issn.1000-3428.2011.12.009

1 概述

随着电子商务和Web2.0应用的流行，越来越多的消费者喜欢在电子商务网站、论坛、博客上写下对产品的观点态度，消费者在购买产品前总会咨询别人对产品的意见从而做出购买决定，人工浏览这些海量产品评论是费时和低效的。近年来，如何对大量的非结构化产品评论进行观点抽取，已成为一个研究的热点。产品属性词和评价词在产品评论的观点信息抽取中起到重要作用。产品属性词通常描述产品的一个特征或部分。评价词是用来表达消费者的观点态度，多数是形容词，但也有少量具有情感倾向的名词和动词。例如手机评论中的一个评论句子：“屏幕大，按键过于紧凑，总的来说，性价比非常高”。其中，“屏幕”、“按键”、“性价比”是属性词，“大”、“紧凑”、“高”是评价词，它们分别表达了对“屏幕”和“性价比”满意的态度，对“按键”持否定的态度。抽取产品属性词和评价词，构建产品属性词词典和评价词词典是产品评论中观点信息抽取的基础工作。属性词抽取与命名体识别有相似之处，都是针对某类特定的名词识别。文本中的上下文信息对这2个任务都有很强的指导意义，但不同的是，因为命名体识别的主要对象是人名、地名和机构名，所以直接利用命名体的识别方法用到属性词抽取方面并不能取得满意的效果。本文旨在设计一种新的方法在评论语料中抽取属性词，从而构建属性词词典。

2 相关研究

目前，已有不少学者针对英文产品评论中属性词抽取方法进行研究。文献[1-2]利用关联规则抽取高频的属性词，然后利用修剪规则提高准确率和覆盖率，进而利用邻近的形容词发现低频的属性词补充属性词列表。文献[3]提供一个可视化评价信息的原型系统，它们利用 tf-idf 获得属性词，并且把属性词分成一般属性词和特殊属性词。文献[4]利用半监督的学习技术抽取属性词-评价词关系对。文献[5]通过预处理，聚类、相似度计算、名词短语聚集、修剪等非监督的方法抽取产品属性词。国内也有不少学者对中文产品评论属性词抽取方法进行研究。文献[6]抓取大量介绍产品的网页，从中手工整理构建属性词表。文献[7]构建商品名称字典，然后利用商品名称字典抽取属性词，手机属性词仅抽取到180个，属性词的覆盖率比较低。

因为目前产品评论属性词抽取的准确率、覆盖率、调和平均值值都不太高，所以有必要进行深入研究，本文主要工作如下：(1)利用百度百科以及分词后的词语同现比例识别专业领域生词，降低分词错误对属性词识别的影响。(2)设计词性序列模板产生候选属性词集，词性序列模板不仅包含了名词和名词短语模板，还包括了动词和动词短语模板，这样可

基金项目：教育部人文社会科学研究青年基金资助项目(09YJCZH019)；湖南省自然科学基金资助项目(10JJ3002)；中国包装总公司科研基金资助项目(2008-XK13)

作者简介：栗春亮(1984－)，男，硕士研究生，主研方向：文本分类；朱艳辉，教授；徐叶强，硕士研究生

收稿日期：2011-01-14 **E-mail：**liliang546@qq.com

以提高覆盖率。(3)利用统计技术和自然语言处理技术筛选候选属性词，构建属性词词典。

3　候选属性词集的产生

3.1　语料集的建立和预处理

本文以手机产品作为实验对象。在百货网、欧酷网、京东网上抓取了 3 500 篇手机评论文章，去掉 Html 等标记，整理成纯文本文件，然后加上文献[8]的 123 篇手机评论文本组成实验语料，共有 3.84 MB，平均每篇文章约有 1.1 KB。采用中文 ICTCLAS 分词系统对评论文本进行分词处理，并标注词性。

3.2　专业领域生词的识别

ICTCLAS 分词系统的词库是通用性词库，会对一些专业领域的词汇造成分词错误，例如:“短信”分成了“短/a 信/n”、“性价比”分成了“性/ng 价/n 比/p”等，为了降低分词错误对产品属性词识别的影响，提出一种利用百度百科和分词后相邻词语同现比例的方法识别专业领域生词。

N-Gram 是最常用的统计语言模型，以马尔科夫模型为理论基础。在语言模型的构造中，以字、词、词性或词义等作为 N-Gram 的统计单元。本文选择 ICTCLAS 分词后的词语作为统计单元。对分词后的词语进行 2-Gram 和 3-Gram 处理。例如一个分词后的句子:“这/rzv 款/q 手机/n 对/p 蓝/a 牙/n 的/ude1 实现/v 不好/a 。/wj”，可以得到“这款”、“款手机”、“手机对”、“对蓝”、“蓝牙”、“牙的”、“的实现”、“实现不好”8 个 2-Gram 项，“这款手机”、“款手机对”、“手机对蓝”、“对蓝牙”，“蓝牙的”、“牙的实现”、“的实现不好”7 个 3-Gram 项。把语料中所有的 2-Gram 项和 3-Gram 项置于百度百科中进行词条搜索，如果出现则作为候选生词。然后利用式(1)中的 p 值进一步过滤从而确定最后的生词:

$$p=\frac{hits(w_1)}{hits(w_2)} \tag{1}$$

其中，w_1 是 2-Gram 项或 3-Gram 项；w_2 是组成 2-Gram 项或 3-Gram 项的词语，并且词语和词语之间用空格分开，例如 2-Gram 项：“蓝牙”，w_1 是“蓝牙”，w_2 是“蓝 牙”；$hits(w_1)$ 是 w_1 在 Google 中利用双引号技巧进行精确匹配搜索返回的页面数；$hits(w_2)$是 w_2 在 Google 中不使用精确匹配搜索返回的页面数。p 值越大，词语间的相关性越大，2-Gram 或 3-Gram 项是一个生词的可能性越大。去除小于某一个阈值的词条，汉语言文学专家将最终确定的生词进行人工词性标注，加入到自定义词库中，重新对语料进行分词，降低分词错误率。

3.3　词性序列模板的设计

为获得候选属性词集，设计一个词性序列模板。首先对第 1 届中文倾向性分析评测会议[8]中任务 3 的 8 177 条标准答案进行分词，获得属性词的词性序列并统计出现次数，把出现次数大于 99 的词性序列作为词性序列模板，词性序列模板不仅包含名词和名词短语，还包含动词和动词短语，见表 1。

表 1　词性序列模板

模板	实例
n	外观/n
n+n	来电/n、铃声/n
v	操作/v
vn+n	通话/vn、质量/n
v+n	显示/v、效果/n
vi+n	拍照/vi、功能/n
n+v	系统/n、反应/v
n+vn	键盘/n、设计/vn

产品评论是消费者针对产品的属性进行评论的文章，因为属性词会频繁出现，所以候选属性词在语料中出现次数越多，它是属性词的可能性越大。首先利用词性序列模板对语料进行匹配获得符合模板的词条，并统计在语料中的出现次数，去除那些出现特殊符号(#、|、&、@、>、<、\、"、^、_、´、*、◆、δ、~等)的词条，把出现次数大于 4 的词条作为候选属性词集，共 6 751 个词条。

4　非属性词的过滤

不是所有的名词、名词短语、动词以及动词短语都是属性词，仅利用设计的词性序列模板将所有满足模板的词条作为属性词必然引入很多噪声。下面利用统计技术和自然语言处理技术，提出 3 条规则对候选属性词集进行过滤。

规则 1 利用互信息过滤。

实验发现有些词汇虽然在评论语料中出现频繁，例如，“时候”、“商品”、“问题”等，但这些词汇跟手机是不相关的，不是真正的属性词，通过量化领域相关性可对候选属性词进行过滤，而互信息 PMI(Point-wise Mutual Information)值提供了这样一种量化方法。从语料中出现次数最多的 100 个候选属性词中，手工挑选手机的 6 个典型属性，分别是屏幕、电池、短信、铃声、按键、待机时间，并加入产品类别名称“手机”，组成领域性种子词集 *Seeds*。*PMI-IR* 的计算如下:

$$PMI\text{-}IR(w_1)=\sum_{w\in Seeds}\mathrm{lb}\frac{hits(w_1\ \&\ w)}{hits(w_1)hits(w)} \tag{2}$$

其中，*Seeds*={手机, 屏幕, 电池, 短信, 铃声, 按键, 待机时间}；$hits(w_1)$、$hits(w)$是词条 w_1、w 在 Google 中利用双引号技巧精确匹配后返回的页面数；$hits(w_1\ \&\ w)$是词条 w_1 和 w 在 Google 中精确匹配并同时出现的页面数。候选属性词的 *PMI-IR* 值越高，是真正的属性词的概率就越大，例如，“通话音质”的 *PMI-IR* 值为-180.79，“地方”的 *PMI-IR* 为－209.44，尽管前者在语料中仅出现了 18 次，而后者在语料中出现了 684 次，但“通话音质”是属性词而“地方”不是。设定一个阈值，如式(3)所示，*PMI-IR* 值大于阈值 α 的为属性词。

$$IsAttri\_ByPMI(w_1)=\begin{cases}\text{Yes} & PMI\text{-}IR(w_1)\geqslant\alpha\\ \text{No} & PMI\text{-}IR(w_1)<\alpha\end{cases} \tag{3}$$

规则 2 候选属性词中情感词、单字动词及特殊动词的过滤。

如果一个词语是主观词或者情感词，则它肯定不是一个属性词。基于这个观点，利用 HowNet 情感分析词语集对候选属性词中的情感词进行过滤。同时根据语言特性，单个的动词作为属性词的可能性也比较小，为提高准确率，对单字的动词进行过滤。在句子中，一些符合模板的词语序列，不可能构成属性词，最容易出现歧义的模板是“v+n”、“n+v”、“vi+n”，例如，“这/rzv 款/q 手机/n 支持/v 红外/n 。/wj”，在该句子中，“手机支持”、“支持红外”不是属性词。对于这种情况，整理了一些不能构成属性词的动词，对包含这些动词的短语进行过滤，如表 2 所示。

表 2　不能构成属性词的动词

动词	实例
形式动词	给与、进行、有、无、可以
趋向动词	到、上来、下来、进来、出来、回来、过来、起来、去、上去、下去、进去
心理动词	打算、喜欢、希望、害怕、担心、讨厌、愿意
助动词	能、要、会、带、放、当、用、让、看、允许、应当、该、能够、敢、可能、必须、支持
系动词	成为、当做、是、为
变化动词	下降、降低、增加、升高

规则 3 利用候选属性词所在句子中形容词和动词性惯用语的同现比例进行过滤。

如果一个候选属性词是真的属性词，通常句子中会同时出现评价词表达观点。所以，在句子中出现候选属性词的情况下，如果同时出现形容词和动词惯用语的比例越高，那么候选属性词是真的属性词的可能性就越大。设定：

times_with_adjorvl=出现候选属性词并同时出现形容词或动词惯用语的句子个数

times=出现候选属性词的句子数

Ratio=*times_with_adjorvl*/*times*

过滤方法如式(4)所示：

$$IsAttri\_ByRatio(w)=\begin{cases}\text{Yes} & Ratio\geqslant b\\ \text{No} & Ratio<b\end{cases} \tag{4}$$

5 实验与结果分析

实验基于 VS 2005 平台。为了评估方法的性能，首先由 4 组实验人员，每组 2 人协同工作，手工从语料中挑选属性词，然后去除那些不一致的结果，确定 1 349 个产品属性词作为标准答案。利用准确率(Precision)，覆盖率(Recall)，调和平均值(F-measure)3 个指标评估方法的性能，具体如下：

$$\text{准确率}=\frac{\text{提交正确的个数}}{\text{提交的总个数}}$$

$$\text{召回率}=\frac{\text{提交正确的个数}}{\text{标准答案的个数}}$$

$$\text{调和平均值}=\frac{2\times\text{准确率}\times\text{召回率}}{\text{准确率}+\text{召回率}} \tag{5}$$

实验 1 阈值 α 的选取

对未过滤的属性词使用规则 1 进行过滤，其准确率、召回率、调和平均值随阈值 α 的变化趋势如图 1 所示，当阈值 α 在－200 附近时，调和平均值达到最高。为了进一步找到合适的阈值，在阈值－204~－196 之间重新做实验，结果表明当阈值 α 取－201 时，调和平均值达到最高。

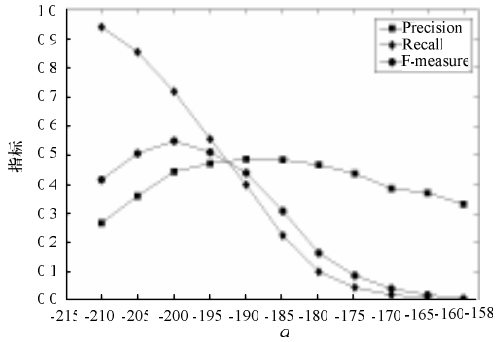

图 1 3 个指标随阈值 α 的变化趋势

实验 2 阈值 β 的选取

图 2 是在使用规则 1+规则 2 过滤候选属性集后，再使用规则 3 过滤，得到不同比例阈值 β 时准确率、召回率、调和平均值的变化趋势，结果表明当比例阈值 β 取 0.1 时，调和平均值达到最高。

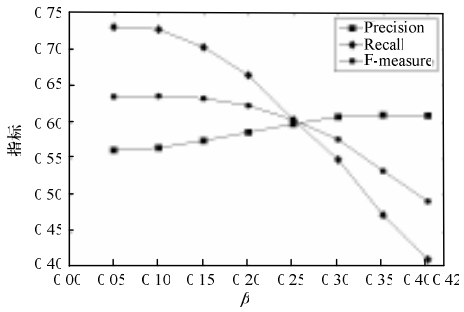

图 2 3 个指标随阈值 β 的变化趋势

实验 3 非属性词的过滤

非属性词过滤实验结果如图 3~图 5 所示，其中，*n* 表示语料库中属性词出现次数按降序排序的前 *n* 条。

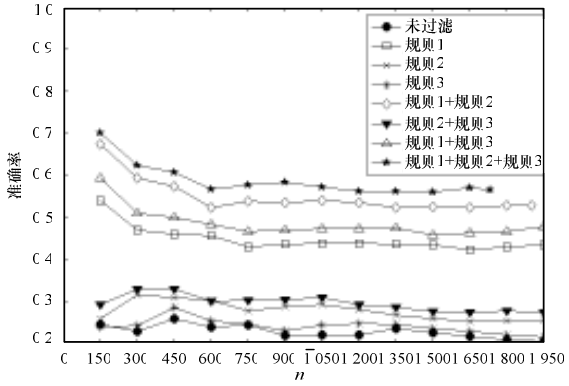

图 3 准确率随 *n* 的变化趋势

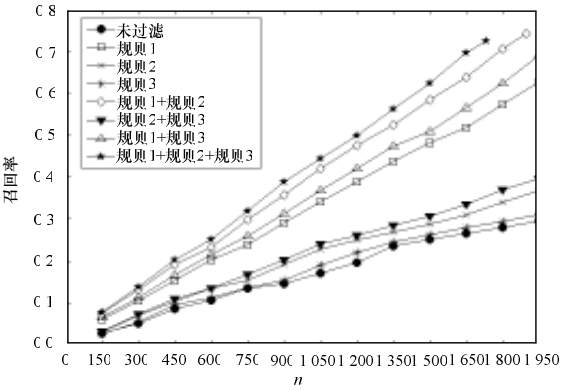

图 4 召回率随 *n* 的变化趋势

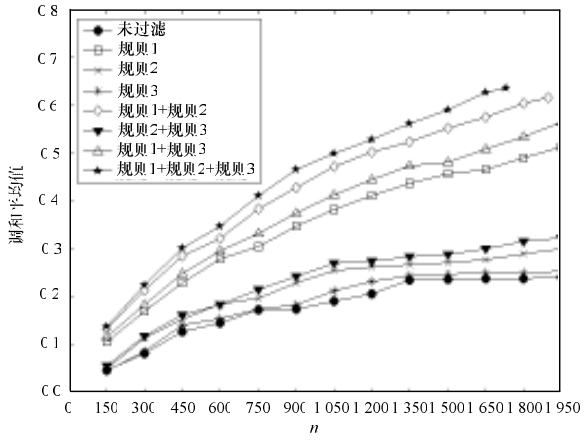

图 5 调和平均值随 *n* 的变化趋势

实验结果表明，随着提交结果数的增多，准确率呈缓慢下降趋势，提交结果数越少，准确率越高。当同时使用 3 条规则提交 150 条结果时，准确率达到最高，为 0.7。当提交 1 732 条结果时，准确率为 0.565，覆盖率为 0.726，调和平均值达到最高，为 0.636。

规则 1 的性能比规则 2 和规则 3 好很多，甚至比规则 2+规则 3 同时使用的性能还要好，可见规则 1 在过滤非属性词上最有效。

单独利用 3 条规则以及利用 3 条规则的不同组合过滤非属性词，比未过滤属性词的准确率、召回率、调和平均值都高，规则 1+规则 2+规则 3 的性能最好，说明提出的 3 条规则在过滤非属性词上是有效的。

(下转第 32 页)