基于互自扩展模式的中文产品属性提取算法*

于明朕, 那日萨
(大连理工大学 管理与经济学部, 辽宁 大连 116024)

摘　要: 针对中文在线评论中产品属性词的提取,提出了一种基于互自扩展模式的半监督学习方法。利用较少的人工参与,通过 FP-Growth 算法挖掘频繁项集获得种子属性词,通过增量迭代发现新的属性词;在每一轮迭代中,通过计算提取词与提取模式的置信度确保了算法的准确性,同时避免了主题偏移。最后通过相似提取模式获得复合提取词,大大减少了因分词及词性标注错误所导致的属性词挖掘错误,以牺牲较少准确率的代价换取了较高的召回率。实验结果表明,该算法对产品属性提取的 F 值可以达到 78.97%,结果优于其他类似的提取算法。

关键词: 在线评论; 产品属性提取; 互自扩展; FP-Growth 算法; 置信度
中图分类号: TP391.1　　**文献标志码**: A　　**文章编号**: 1001-3695(2017)04-0977-04
doi:10.3969/j.issn.1001-3695.2017.04.004

Feature extraction method based on mutual self-expanding mode

Yu Mingzhen, Zhao Narisa
(*Faculty of Management & Economics*, *Dalian University of Technology*, *Dalian Liaoning* 116024, *China*)

Abstract: This paper proposed a feature extraction method based on mutual self-expanding in Chinese product comment. With little manual work, this method found seed features by FP-Growth, then found the other new features by an incremental iterative procedure. During the iteration, the confidence coefficient of the extracted-word and the extracted-mode insured a high precision, avoided deviating theme at the same time. At last, this method found combination extracted-word by similarity extracted-mode. It could reduce many feature extraction mistakes caused by word segmentation technology and part-of-speech tagging technology, and got a high precision with reducing little recall rate. The experimental results indicate that the F-score of the proposed method for product feature extraction can be 78.97%, is better than the other method of the literatures of this paper.

Key words: online comment; product features extraction; mutual sel-expanding; FP-Growth method; confidence coefficient

0　引言

近些年来,伴随互联网的高速发展,越来越多的消费者选择通过网络来购买产品。而在购物之前,消费者往往习惯通过在网上查看产品的相关评论来了解其相关信息。然而,由于消费者间的个体差异,导致个人喜好不同;关注点不同。并且产品相关的在线评论的信息量往往很大,消费者通过人工来选择浏览自己关注的方面将会消耗大量精力,且效果不佳。因此,如何从产品在线评论这种非结构化数据中提取用户关注的信息,尤其关于产品相关属性的信息显得格外重要。同时,提取产品各方面属性词是在线评论中多维度情感研究的基础性工作,评论中往往包含着用户对产品及其相关服务的复杂情感,针对不同属性表达出不同程度的情感倾向。然而,早期的情感分析研究往往是针对整个产品,这种粗糙的全局情感分析方法往往很难满足消费者多方面的需求[1]。在产品评论中往往包含数以百计的属性词,将这些属性词正确地提取[2]将对产品评论信息的深度挖掘意义重大。目前针对评论提取产品相关属性多采用机器学习的方法。这些方法中,可以根据对训练语料中是否进行人工参与标注、标注形式以及标注量等方面分为有监督、半监督和无监督三种形式。有监督方法中,Jin 等人[3,4]提出了一种基于词汇化隐马尔可夫模型产品属性提取算法。Jakob 等人[5]采用条件随机场模型来提取产品属性,以词、词性、与情感词的依存关系、与情感词间的距离等为特征,获得了比较理想的效果。Liu 等人[6]则通过对训练集人工标注,然后根据从中学习到的提取规则对测试集中的评论提取产品属性。

以上方法尽管可以取得较好的提取效果,然而需要大量人工标注的训练集,费时费力。因此,更多人开始对无监督方法进行深入的研究。Hu 等人[7]利用关联规则得到高频属性词,再通过剪枝提高准确率与召回率,最后通过邻近的形容词发现低频属性词作为补充。该方法的提取效果依赖最小支持度阈值、评论数量等多个参数,同时对低频属性词提取的准确率不高。Qiu 等人[8]将文献[7]的方法进行了改进,提出了在属性词与情感词之间依存关系的基础上进行确定,而非简单地通过邻接关系来确定。Wei 等人[9]在利用关联规则提取频繁属性后,进一步利用已知的情感词对这些频繁属性进行过滤,以提高属性提取的准确性。除了利用关联规则外,文献[10]提出了另一种产品属性提取方法,利用八条与领域无关的模板提取

收稿日期: 2016-03-21; **修回日期**: 2016-05-06　　**基金项目**: 国家自然科学基金面上项目(61471083);国家教育部人文社科研究规划基金资助项目(14YJA630044)

作者简介:于明朕(1989-),男,辽宁大连人,硕士研究生,主要研究方向为机器学习、文本挖掘、中文情感分析(21311013@mail.dlut.edu.cn);那日萨(1970-),男,辽宁大连人,教授,博士,主要研究方向为情感计算、文本挖掘.

候选产品属性,计算其与产品的点互信息值(PMI)。这种方法以牺牲较少召回率的条件获得了较多准确率的提高。文献[11]是文献[10]对中文的应用。近年来,研究者尝试使用主题模型提取产品属性,也取得了一定的效果。Titov 等人[12,13]采用一种扩展的 LDA 提取显式产品属性,提出了全局特征属性和局部特征属性两个概念,以全局特征属性为主题,将局部特征属性视为主题相关项;但是此方法无法得到关于产品精确特征属性集。Sauper 等人[14]使用另一种扩展的 LDA 方法,将评论的子句与产品的属性词和情感词相联系;该方法需要预设多个参数,因而降低了该方法的普适性与扩展性。

从现在的研究来看,对中文评论中提取产品属性的研究还比较少,并且由于中文中存在分词及词性标注不准确、句法复杂等困难,其效果也远未达到英文评论中的提取效果。针对以上问题,本文在上述研究的基础上定义了提取词及提取模式置信度的概念,并给出了其求解方法,提出了一种以基于互自扩展模式的产品属性提取算法。

1 提取模式、提取词置信度的定义与求解

每句评论都由一个或多个短句构成,通过换行标志、逗号、句号、叹号、问号可以将在线评论划分为若干个子句。每个子句通过分词,然后将每个词替换为该词对应的词性,则称之为该子句的词性集。每个词性集可对应多个子句。将词性集与需要挖掘的属性词在词性集中的位置标志合称为提取模式。通过词性集可找到多个子句,通过位置标志可找到这些子句中对应的目标词。通过提取模式找到的词称之为提取词。

1.1 提取模式、提取词置信度的定义

提取词的置信度 $\mathrm{word}_{\mathrm{belief}}$ 表示该词为产品属性词的可信度。$\mathrm{word}_{\mathrm{belief}}$ 为实数,取值范围为 $(-\infty,+\infty)$,置信度越高,该提取词为属性词的可能性越大。算法中定义如果 $\mathrm{word}_{\mathrm{belief}} \leqslant -1$,该提取词为非属性提取词;如果 $-1 < \mathrm{word}_{\mathrm{belief}} < 1$,该提取词为待定属性提取词;如果 $\mathrm{word}_{\mathrm{belief}} \geqslant 1$,则该提取词为属性提取词,非属性提取词与属性提取词可统称为确定属性提取词。

提取模式的置信度 $\mathrm{mode}_{\mathrm{belief}}$ 表示该通过该模板提取的词为产品属性词的可信度。$\mathrm{mode}_{\mathrm{belief}}$ 为实数,取值范围为 $(-\infty,+\infty)$,置信度越高,该模板提取到的提取词为属性词的可能性越大。算法中定义如果 $\mathrm{mode}_{\mathrm{belief}} \leqslant -1$,该提取模式为不可信提取模式,如果 $-1 < \mathrm{mode}_{\mathrm{belief}} < 1$,该提取模式为待确定提取模式,如果 $\mathrm{mode}_{\mathrm{belief}} \geqslant 1$,该提取模式为可信提取模式。不可信提取模式与可信提取模式统称为确定提取模式。

1.2 提取模式、提取词置信度的求解

设在训练集中,提取词出现总次数为 $\mathrm{word}_{\mathrm{all}}$,提取词出现在可信提取模式的次数为 $\mathrm{word}_{\mathrm{high}}$,提取词出现在不可信提取模式的次数为 $\mathrm{word}_{\mathrm{low}}$。可得到提取词的提取模式中,模板总数量为 $\mathrm{word}_{\mathrm{allmode}}$,可信模板的数量为 $\mathrm{word}_{\mathrm{highmode}}$,不可信模板的数量为 $\mathrm{word}_{\mathrm{lowmode}}$。则 $\mathrm{word}_{\mathrm{belief}}$ 由三部分决定,第一部分称为基础置信度,用 $\mathrm{word}_{\mathrm{base}}$ 来表示。

$$\mathrm{word}_{\mathrm{base}} = a \times \frac{(\mathrm{word}_{\mathrm{high}} - \mathrm{word}_{\mathrm{low}})}{\mathrm{word}_{\mathrm{all}}} \tag{1}$$

因为提取词均为评论中的词,所以 $\mathrm{word}_{\mathrm{all}} \geqslant 1$。如果 $\mathrm{word}_{\mathrm{high}} > \mathrm{word}_{\mathrm{low}}$,$\mathrm{word}_{\mathrm{base}}$ 为正数,表明该提取词更偏向于属性提取词。如果 $\mathrm{word}_{\mathrm{high}} < \mathrm{word}_{\mathrm{low}}$,$\mathrm{word}_{\mathrm{base}}$ 为负数,表明该提取词更偏向于非属性提取词;同时,两者之间的差距应达到总次数的某种比例才可以确定是属性提取词还是非属性提取词,如果差距很小应作为待定属性提取词。a 为调整参数,a 越大,置信度越大,越容易召回,但准确率会相应降低。

$\mathrm{word}_{\mathrm{belief}}$ 第二部分为待定词影响因素,用 $\mathrm{word}_{\mathrm{hold}}$ 来表示。

$$\mathrm{word}_{\mathrm{hold}} = b \times \frac{(\mathrm{word}_{\mathrm{high}} + \mathrm{word}_{\mathrm{low}})}{\mathrm{word}_{\mathrm{all}}} \tag{2}$$

如果提取词出现在确定提取模式中的次数占比越多,其可信度就越大。其中 b 为调整参数,b 越大,置信度越大,则越容易召回,但准确率会相应降低。

$\mathrm{word}_{\mathrm{belief}}$ 第三部分为模板数量影响因素,用 $\mathrm{word}_{\mathrm{mode}}$ 来表示。

$$\mathrm{word}_{\mathrm{mode}} = c \times \frac{\mathrm{word}_{\mathrm{highmode}}}{\mathrm{word}_{\mathrm{allmode}}} (\mathrm{word}_{\mathrm{high}} > \mathrm{word}_{\mathrm{low}}) \tag{3}$$

$$\mathrm{word}_{\mathrm{mode}} = c \times \frac{\mathrm{word}_{\mathrm{lowmode}}}{\mathrm{word}_{\mathrm{allmode}}} (\mathrm{word}_{\mathrm{high}} < \mathrm{word}_{\mathrm{low}}) \tag{4}$$

只由提取词出现的次数来判断仍然不够准确,如果该提取词可以由多个可信模板提取到,则其更加可能为属性提取词,反之亦然。所以可得到提取词的提取模式中,可信模板比例越高,其置信度应该越高,即 $\frac{\mathrm{word}_{\mathrm{highmode}}}{\mathrm{word}_{\mathrm{allmode}}}$ 越大,$\mathrm{word}_{\mathrm{belief}}$ 越大。

最后,$\mathrm{word}_{\mathrm{belief}} = \mathrm{word}_{\mathrm{base}} \times \mathrm{word}_{\mathrm{hold}} \times \mathrm{word}_{\mathrm{mode}}$。整理后得到

$$\mathrm{word}_{\mathrm{belief}} = A \times \frac{(\mathrm{word}_{\mathrm{high}}^2 - \mathrm{word}_{\mathrm{low}}^2)\,\mathrm{word}_{\mathrm{highmode}}}{\mathrm{word}_{\mathrm{all}}^2\,\mathrm{word}_{\mathrm{allmode}}} \quad \mathrm{word}_{\mathrm{high}} > \mathrm{word}_{\mathrm{low}} \tag{5}$$

$$\mathrm{word}_{\mathrm{belief}} = A \times \frac{(\mathrm{word}_{\mathrm{high}}^2 - \mathrm{word}_{\mathrm{low}}^2)\,\mathrm{word}_{\mathrm{lowmode}}}{\mathrm{word}_{\mathrm{all}}^2\,\mathrm{word}_{\mathrm{allmode}}} \quad \mathrm{word}_{\mathrm{high}} < \mathrm{word}_{\mathrm{low}} \tag{6}$$

其中:A 为调整参数,A 值越大,准确率越低,召回率越高。

与提取词置信度计算方法相似,设在训练集中,由模板提取到的提取词出现总次数为 $\mathrm{mode}_{\mathrm{all}}$,提取到的提取词为属性提取词的出现次数为 $\mathrm{mode}_{\mathrm{high}}$,提取到的提取词为非属性提取词的出现次数为 $\mathrm{mode}_{\mathrm{low}}$。通过提取模式提取到的提取词中,总数量为 $\mathrm{mode}_{\mathrm{allword}}$,属性提取词的数量为 $\mathrm{mode}_{\mathrm{highword}}$,不可信模板的数量为 $\mathrm{mode}_{\mathrm{lowword}}$,则提取模式置信度的计算公式如下:

$$\mathrm{mode}_{\mathrm{belief}} = A \times \frac{(\mathrm{mode}_{\mathrm{high}}^2 - \mathrm{mode}_{\mathrm{low}}^2)\,\mathrm{mode}_{\mathrm{highword}}}{\mathrm{mode}_{\mathrm{all}}^2\,\mathrm{mode}_{\mathrm{allword}}} \quad \mathrm{mode}_{\mathrm{high}} > \mathrm{mode}_{\mathrm{low}} \tag{7}$$

$$\mathrm{mode}_{\mathrm{belief}} = A \times \frac{(\mathrm{mode}_{\mathrm{high}}^2 - \mathrm{mode}_{\mathrm{low}}^2)\,\mathrm{mode}_{\mathrm{lowword}}}{\mathrm{mode}_{\mathrm{all}}^2\,\mathrm{mode}_{\mathrm{allword}}} \quad \mathrm{mode}_{\mathrm{high}} < \mathrm{mode}_{\mathrm{low}} \tag{8}$$

2 基于互自扩展模式的产品属性提取算法

在定义及求解提取词、提取模式的基础上,首先对数据进行采集与预处理,然后基于 FP-Growth 算法筛选出初始种子提取词,最后通过互自扩展模式获得评论中的产品属性词。最后通过复合提取词的挖掘方法获得更好的属性提取效果。

2.1 数据的采集与预处理

a)语料采集。针对亚马逊(中国)网站上不同产品的在线评论,实现了可以定期采集包含指定产品在线评论网页的网络爬虫,并利用正则表达式获得网页中的在线评论。

b)评论分词。使用分词工具 ICTCLAS 对采集到的在线评论进行分词并标注词性。ICTCLAS(Institute of Computing Technology, Chinese lexical analysis system)是由中国科学院计算技术研究所通过多年研究,研制而成的汉语词法分析系统。其功能包括中文文本分词、词性的标注等,同时可支持用户自定义

词典。

c)建立情感词库。使用台湾大学中文极性情感词典 NTUSD 与大连理工大学信息检索研究室中文情感词汇本体库两个词典组成本算法中的情感词库。其中中文情感极性词典 NTUSD 是基于文本情感二元划分方法的一个中文词语数据库,包含 11 086 个词语,其中 2 810 个积极情感词语和 8 276 个消极情感词语。中文情感词汇本体库则是由大连理工大学信息检索研究室林鸿飞教授指导,所有研究室成员参与整理和标注的一个中文词语数据。该数据从不同角度来描述词汇或者短语,其中有词语的词性、情感的极性、情感的类别、情感的强度等信息,本文算法中只利用其词性与情感极性两方面维度。由于情感词库中词汇量巨大,无法导入到 ICTCLAS 的用户自定义词典中,有可能导致数据经过分词后原始情感词或短语被分为多个词语,所以对情感词库中的词汇及短语进行分词,将分词后的结果作为最终的情感词库。

d)划分子句。将采集到的在线评论,通过指定标点符号与换行标志划分为若干子句。指定标点符号包括逗号、句号、叹号、问号。最后去掉只有标点符号或分词数量小于等于 2 的子句;之后利用分词工具对子句进行分词,生成其对应的词性集。

e)初始提取模式。结合以往研究以及对标注后的数据集进行人工分析,得到三个可信提取模式作为初始提取模式,其表示方法如下:〈n a sw,1〉〈n d a sw,1〉〈n v n sw,1〉。逗号前边部分为词性集,逗号后边部分为需要挖掘的属性词在词性集中的位置标志。三个可信提取模式的置信度赋值为 1。初始提取模式的置信度并非一直不变,随着算法的不断迭代进一步地优化。

f)初始提取词。选取评论中所有名词,若该名词从未与情感词出现在同一子句中,认为其为非属性提取词,置信度赋值为 -1。由种子提取模式提取到的词作为属性提取词,置信度赋值为 1。初始提取模式的置信度同样是可以变化的,会随着算法的不断迭代进一步地优化。

2.2　基于 FP-Growth 算法选择初始种子提取词

FP-Growth 算法是常见的挖掘频繁项集的两种算法之一,另一种是 Apriori 算法。由于 Apriori 算法需要构建候选集,然后对候选集进行进一步筛选,最后挖掘出频繁项集。所以 Apriori 算法需要反复读取原始数据,当数据较大时,效率比较低下。FP-Growth 算法因其 FP-tree 数据结构压缩了原始数据,所以只需读取两遍原始数据,具有较高的效率。为了满足在大数据背景下的产品属性挖掘,这里采用 FP-Growth 算法选择初始种子提取词。

由于产品属性词多为名词或名词短语[10],并且消费者在评价产品属性时,同时表达了对该属性的情感,以此为依据可以从数据中挖掘出一定数量的名词作为算法中的种子提取词。具体方法如下:对数据进行进一步的处理,只保留评论中的名词、名词短语以及情感词,同时将所有情感词以统一的符号 emo 来表示。通过 FP-Growth 算法挖掘包含情感词符号 emo 与名词的频繁项集,然后将所挖掘的名词作为种子提取词。

2.3　互自扩展算法

互自扩展算法是一种半监督的机器学习方法,通过所得到的种子进行增量迭代学习,直至收敛。每轮迭代都会产生新的数据,新的数据训练出新的提取模式,新的提取模式继续产生新的数据,直至收敛或达到预定的迭代次数后结束算法。

在本文中,通过 FP-Growth 算法获得第一轮种子提取词集合,通过训练集找到包含种子提取词的评论子句,进而获得该子句的词性集及该种子提取词的所在位置,组合成为候选提取模式,所有候选提取模式作为候选提取模式集合;然后通过训练集找到符合提取模式的评论子句,进而获得模板指定位置的词作为下一轮种子提取词集合。按此方法不断迭代直至收敛或达到迭代次数。

但是这种互自扩展算法的缺点是容易发生主题漂移,比如〈n d a sw,1〉提取模式,大部分情况下,通过该模板寻找到评论类似“系统很稳定,”等,从而挖掘出“系统”这个属性词;但该模板也会找到评论“女朋友很喜欢。”,而误认为“女朋友”同样为该产品的属性词,进而继续挖掘出错误的提取模式。因此引入了提取词与提取模式置信度的概念,通过筛选高置信度的提取词与提取模式保证了主题不会发生漂移。当挖掘到“女朋友”提取词时,由于其他评论“给女朋友买的。”“女朋友说不错。”提取到〈p n v u sw,2〉〈n v a sw,1〉等不可信提取模式和待定提取模式,降低了该提取词的置信度,从而减少了此种错误。

2.4　复合提取词的挖掘方法

由于对中文文本的分词与词性标注,目前还无法做到 100% 正确,这大大影响了中文在线评论中属性提取的效果。本文中,通过挖掘到的可信提取模式寻找相似提取模式,继而获得高置信度的复合提取词。

假设有两个提取模式 M_1、M_2。M_1 为可信提取模式,提取模式 M_1 的词性集中,属性词位置标志处的词性前部词性集称为属性词前部词性集,属性词位置标志处的词性后部词性集称为属性词后部词性集。如果提取模式 M_2 的词性集中,前部分与 M_1 的属性词前部词性集相同,同时后部分与 M_1 的属性词后部词性集相同,不同的地方词性数量小于等于 3,则称 M_2 为 M_1 的相似提取模式,由这个几个词性提取到的词组合称为复合提取词。由于复合提取词有较大可能性为错误词汇,所以对复合提取词需要进行更为严格的筛选,不仅需要计算该复合提取词的置信度,并且需要结合该复合提取词在 Web 中出现的次数。本文设置,如果该复合提取词的置信度大于等于 1,并且在 Web 搜索前 10 页中出现次数大于等于 5,则认为该复合提取词为属性提取词。

通过获得复合提取词,可以在牺牲较少准确性的情况下明显地提高算法的召回率。但不会通过获得的复合提取词进一步挖掘提取模式,而是直接加入到属性提取词集合的结果中。

2.5　算法实现流程

本文基于互自扩展模式提取属性词的算法流程如图 1 所示。

3　实验结果及分析

3.1　实验数据与评价指标

本文的实验数据从亚马逊(中国)网站上获取,包括手机评论 716 条,MP3 评论 481 条、电脑评论 472 条。在每种产品的评论中随机抽取 100 条评论,人工标注评论中所包含的属性,作为测试集。对测试集进行属性提取,以准确率、召回率、F 值评价算法性能,它们的定义如下:

$$\text{precision} = \frac{\text{正确抽取的产品属性数量}}{\text{抽取的产品属性数量}} \tag{9}$$

$$\text{recall} = \frac{\text{正确抽取的产品属性数量}}{\text{全部产品属性数量}} \tag{10}$$

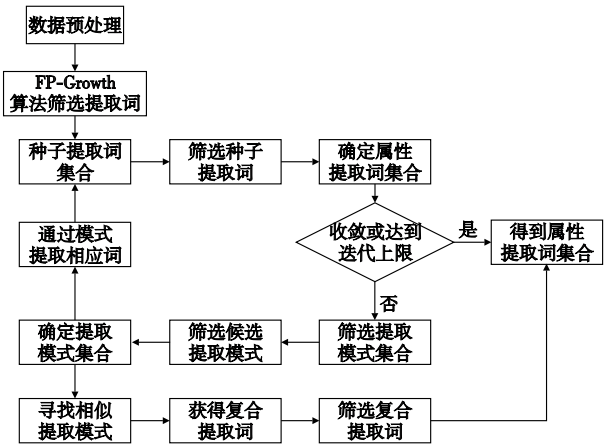

图1 属性词提取算法流程

$$F\text{-score} = \frac{2 \times \text{precision} \times \text{recall}}{\text{precision} + \text{recall}} \tag{11}$$

3.2 实验结果

为了评价不同的调整参数 A 与属性词的提取效果之间的关系,实验中利用不同的调整参数 A 测得 F 值结果如图 2 所示。图 2 中横轴表示调整参数 A 的取值,纵轴表示提取效果的 F-score。

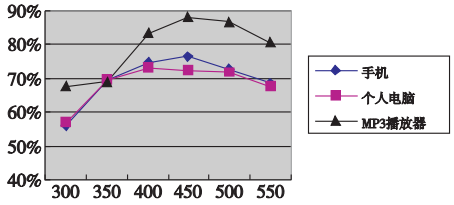

图2 不同调整参数值下的实验结果

由图 2 中的实验结果可以看出,当调整参数 A 取 450 时 F-score 的平均值最大,达到 78.97%。表 1 给出了当调整参数 A 取 450 时,每种产品提取效果的具体评价指标。

表 1 产品属性提取结果

产品名称	准确率	召回率	F 值
手机	0.779 8	0.748 6	0.763 8
个人电脑	0.751 3	0.698 1	0.723 7
MP3 播放器	0.823 9	0.947 7	0.881 5

由表 1 可知,本文算法对产品属性的提取效果,平均准确率为 78.50%,平均召回率为 79.81%,平均 F-score 为 78.97%,其中 MP3 播放器的提取效果最为理想。

为了进一步验证算法的有效性,将文献[15,16]用于处理中文产品在线评论的结果与本文算法的结果进行了对比,结果表明,本文算法对准确率的提高效果较为明显,召回率基本持平,F-score 的结果良好。

表 2 不同方法下结果的对比

指标	文献[15]	文献[16]	本文
准确率	0.679	0.453	0.785 0
召回率	0.787	0.794	0.798 1
F 值	0.728	0.577	0.789 7

4 结束语

本文提出了一种半监督的产品属性提取算法,该算法基于互自扩展模式,结合 FP-Growth 算法,利用提取词与提取模式的置信度减少互自扩展算法中的主题偏移问题。相似提取模式、复合提取词方法的提出,大大提高了属性词的提取效果。实验结果表明,本文算法对产品属性提取的 F-score 可以达到 78.97%,结果优于文献中其他类似提取算法。下一步将本文算法与判断产品有用性方法相结合,探讨评论中所提及的产品属性对该评论有用性的影响,实现高有用性在线评论的自动发现。

参考文献:

[1] Thet T T, Na J C, Khoo C S. Aspect-based sentiment analysis of movie reviews on discussion boards [J]. Journal of Information Science, 2010, 36(6): 823-848.

[2] Arun A, Srinivasan P. Automated query generation Of Rdbms for informationandknowledge extraction [C]//Proc of International Conference on Information Communication and Embedded Systems. [S. l.]: IEEE Press, 2013: 468-473.

[3] Wei Jin, Ho H H, Srihari R K. OpinionMiner: a novel machine learning system for Web opinion mining and extraction [C]//Proc of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. NewYork: ACM Press, 2009: 1195-1204.

[4] Wei Jin, Hung H H. A novel lexicalized HMM-based learning framework for Web opinion mining [C]//Proc of the 26th International Conference on Machine Learning. NewYork: ACM Press, 2009: 465-472.

[5] Jakob N, Gurevych I. Extracting opinion targets in a single-and cross-domain setting with conditional random fields [C]//Proc of Conference on Empirical Methods in Natural Language Processing. Stroudsburg: Association for Computation Linguistics, 2010: 1035-1045.

[6] Liu Bing, Hu Minqing, Cheng Junsheng. Opinion observer: analyzing and comparing opinions on the Web [C]//Proc of the 14th International Conference on World Wide Web. NewYork: ACM Press, 2005: 342-351.

[7] Hu Minqing, Liu Bing. Mining and summarizing customer reviews [C]//Proc of the 10th ACM SIGKDD International Conference on Knowledge Discovery And Data Mining. NewYork: ACM Press, 2004: 168-177.

[8] Qiu Guang, Liu Bing, Bu Jiajun, *et al*. Opinion word expansion and target extraction through double propagation [J]. Computational Linguistics, 2011, 37(1): 9-27.

[9] Wei C P, Chen Y M, Yang C S, *et al*. Understanding what concerns consumers: a semantic approach to product feature extraction from consumer reviews [J]. Information Systems and E-business Management, 2010, 8(2): 149-167.

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

[11] 粟春亮,朱艳辉,徐叶强. 中文产品评论中属性词抽取分法研究[J]. 计算机工程, 2011, 37(12): 26-28.

[12] Titov I, McDonald R. Modeling online reviews with multi-grain topic models [C]//Proc of the 17th International Conference on World Wide Web. New York: ACM Press, 2008: 111-120.

[13] Titov I, McDonald R. A joint model of text and aspect ratings for sentiment summarization [C]//Proc of the 46th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies. 2008: 308-316.

[14] Sauper C, Barzilay R. Automatic aggregation by joint modeling of aspects and values [J]. Journal of Artificial Intelligence Research, 2013, 46(1): 89-127.

[15] 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.

[16] 刘鸿宇,赵妍妍,秦兵,等. 评价对象抽取及其倾向性分析[J]. 中文信息学报, 2010, 24(1): 84-88.