第35卷 第13期 计 算 机 工 程 2009年7月
Vol.35 No.13 Computer Engineering July 2009
·人工智能及识别技术· 文章编号：1000—3428(2009)13—0199—03 文献标识码：A 中图分类号：TP18

基于弱监督学习的产品特征抽取

伍 星，何中市，黄永文
(重庆大学计算机学院，重庆 400044)

摘 要：产品评论挖掘是从自然语言描述的用户评论中获取信息的过程，产品特征抽取是产品评论挖掘的第1个阶段，产品特征的好坏决定了产品评论挖掘中后续阶段的质量。采用弱监督的学习方法，只需要提供少量的产品特征作为种子，从这些种子出现的语句中抽取文本模式，利用文本模式来发现新的产品特征。实验结果表明，从英文文本中自动抽取产品特征的实验系统，取得了较好的效果。

关键词：BootStrapping算法；文本模式抽取；产品评论挖掘

Product Feature Extraction Based on Weakly Supervised Learning

WU Xing, HE Zhong-shi, HUANG Yong-wen
(College of Computer, Chongqing University, Chongqing 400044)

【Abstract】The mining of product reviews is the process of extracting information in reviews which is expressed by natural language, the extraction of product feature is the first phrase of the mining of product reviews. The quality of product feature decides the quality of subsequent phrases. This paper adopts weakly supervised methods, which just need a hand of product features as the seeds, using the occurrence sentences of seed to extract text patterns, and using the text patterns to find new product features. Experimental results show that it can extract product feature from English plain text and receive good result.

【Key words】BootStrapping algorithm; text pattern extraction; product review mining

1 概述

随着Internet的广泛应用，用户使用产品会通过Web对产品进行评论，这些评论中包含用户对产品的各个方面的性能持有肯定还是否定的意见。产品评论中蕴涵了丰富的信息，生产厂商分析产品评论可以了解产品的不足和用户实际需求以改进产品，用户浏览产品评论可以在购买产品之前更多地了解产品，从而更加合理地购买产品。要从大量使用自然语言进行描述用户评论获取信息，只有通过人工逐一阅读，这是一个需要大量时间和精力的过程，因此，需要自动化的产品评论挖掘来更快地从大量的用户评论中获取信息。

产品评论大多用自然语言进行描述，生产厂商和用户只有采用人工阅读的方式才能从中提取信息，而这是一个费时、费力且容易产生错误的过程，因此，产生了自动产品评论挖掘的需求。产品评论挖掘一般分为产品特征提取、主观句定位、用户态度提取、用户态度极性判断和挖掘结果显示等5个阶段。产品特征提取作为产品评论挖掘的第1个阶段，目的是从众多的用户评论中挖掘出用户所关心的产品特征，如相机的产品特征包括：重量，大小，图片的质量，电池的使用时间，存储容量等。

产品特征抽取分为人工定义和自动抽取2类方法。文献[1]以人工方式提出了针对汽车的产品特征，文献[2]以人工方式提出了针对电影的产品特征。由于人工定义一类产品的产品特征都必须有该领域专家参与，不同的领域需要不同的领域专家来定义，因此人工定义产品特征移植性差，并且定义完成后，产品特征并不随该类产品的变化而变化，只有再次通过领域专家的定义加入新的产品特征，因此，人工定义产品特征是静态的。自动抽取使用词性标注、句法分析等自然语言处理的技术来自动抽取产品特征。文献[3]首先对评论中的主观性语句进行句法分析，找到句子中的名词或名词短语，然后从中寻找频繁项，将得到的频繁项作为产品的特征，该方法召回率较高，但准确率低。文献[4]采用人工定义的通用文本模板，根据具体应用领域实例化通用文本模板以形成抽取规则，再利用抽取规则进行产品特征的抽取，取得了较高召回率和准确率，但该方法存在部署困难和移植性差的缺点。

文本模式[5-6]广泛地应用于信息抽取领域，用于发现命名实体和实体之间的关系，而文本模式的获取存在人工定义和自动获取2种方式，人工定义文本模式需要专家总结，移植性差。自动获取分为有基于监督学习和基于弱监督学习2种方法，基于监督的学习方法需要大量的人力进行领域相关的语料库的标注，领域改变以后需要重新标注，因此，同样存在移植性差的缺点。基于弱监督的方法只需要提供领域相关少量的正确实例和领域相关的语料库，通过迭代方式自动学习文本模式，迭代过程中使用提供的少量的正确实例作为知识对新产生的文本模式和新的实例进行评估。

本文采用弱监督方法来自动获取文本模式，并用学习得到的文本模式来抽取新的产品特征，取得了较高的召回率和准确率，系统具有良好的移植性。

作者简介：伍 星(1978－)，男，博士研究生，主研方向：自然语言处理；何中市，教授、博士、博士生导师；黄永文，博士研究生
收稿日期：2009-01-20 **E-mail：**wuxing-078@163.com

2 系统介绍

本文采用基于 BootStrapping 的弱监督机器学习方法，只需提供少量的产品特征作为种子集合，自动进行文本模式的抽取，再用抽取得到的模式抽取新的产品特征。系统以人工提供的少量产品特征作为种子集合，发现产品评论语料库中的产品特征出现语句，将这些语句按照给定的文本模式结构进行模式化表示，从中生成新的文本模式，再用这些自动获取的文本模式来抽取新的产品特征，并将新的产品特征加入产品特征种子集合。对该过程不断地迭代，直到系统不能产生新种子或新的文本模式和达到人工指定迭代次数停止迭代，将产品特征种子集合中的种子输出作为结果，系统流程如图 1 所示。

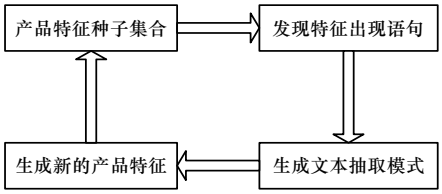

图 1 系统流程

整个系统流程分为 4 个阶段：

(1)发现产品特征种子出现语句

将语料库中的产品评论分解为语句，并对每一个句子进行词性标注，提取句子中的名词和名词短语，如果它是产品特征种子集合的一个元素，那么将出现该名词或名词短语的语句加入产品特征出现语句集合$|S_O|$。

(2)生成文本模式

从$|S_O|$中自动发现可用于抽取新的产品特征的文本模式，该过程分解为 3 个子过程：

1)模式化表示语句

通过对产品特征语料库的观察发现，一个名词所在的依存关系对该名词是否是产品特征有极大的指示作用，因此，选用基于依存关系的文本模式的表示方法。

定义 1 弱监督产品特征抽取的文本模式：

<Rel, MPos, D, L>

其中，*Rel* 是依存关系的类型；*MPos* 是依存关系中修饰中心词的修饰词词性；*D* 是中心词和修饰词之间所在语句中的位置距离的差值，中心词在修饰词之前该值为正，反之为负；*L* 是依存关系中的中心词在所在语句的句法树结构中层次数。

2)生成侯选文本模式

对$|S_O|$中的语句进行句法分析，并生成依存关系，将每个出现名词或名词短语作为中心词的依存关系按照定义 1 中的文本模式结构进行模式化表示，每一个模式化表示的依存关系记为 T_i，将所有模式化表示的依存关系加入集合$|T|$。本文选用文本模式的 4 个元组中包含有非数值型的值(*Rel, MPos*)和数值型的值(*D,L*)，因此，侯选文本模式的产生分成 2 步：

第 1 步：将文本模式的(*Rel, MPos*)看作一个整体对模式化表示的语句进行分类，统计每一类中包含 T_i 的个数，如果该类中包含的文本模式数目大于设定的阈值，那么该类作为一个可以产生文本候选模式的类。

第 2 步：根据模式的(*D, L*)的值，计算同类中的 T_i 之间的近似度，并根据近似度选择中心值的 T_i 作为该类的代表模式，并将该代表模式作为一个候选文本模式(C_i)加入候选文本模式集合$|C|$。定义 2 用于计算 T_i 和 T_j 之间的相似度，T_i 和 T_j 均来自于同一类。

定义 2 文本模式相似度：

$$Sim(T_i,T_j)=\begin{cases}\sqrt{(D_i-D_j)^2+(L_i-L_j)^2} & Rel_i=Rel_j \text{并且} MPos_i=MPos_j\\ 0 & \text{其他}\end{cases} \tag{1}$$

3)评估候选文本模式

评估候选文本模式的目的是对候选模式集合$|C|$中的候选文本模式进行评估以得到优秀的文本模式，并将该优秀模式加入文本模式集合以寻找新的产品特征。目前采用弱监督进行自动学习的方法中均需要对生成的候选文本模式的可靠性进行度量。

文献[5]方法中文本模式用于抽取实体和实体之间的关系，种子集合中的种子是一个二元组，因此，可以直接使用文本模式发现的实例是否和种子集合中已有的种子之间存在矛盾来判断抽取得到实例的正确性，从而可以对文本模式的可靠性进行度量。由于本文的种子是一元组，因此不能直接采用文献[5]的方法来度量候选文本模式。

文献[6]同样采用了 BootStrapping 方法来自动建立电子词典，该方法给定的种子集合中的种子直接就是人工定义的文本模式，使用这些文本模式发现新的词条。为了对新发现的词条进行度量，文献[6]采用了一个假设：如果产生该名词短语的文本模式越多，则该名词短语越可靠。因此，本文使用种子集合中产品特征来评估候选模式。

本文假设：使用候选模式抽取产品特征，能够得到越多种子集合中的特征的模式越可靠。所有的候选模式采用定义 3 中的公式进行置信度评价，将置信度最高的文本模式作为优秀文本模式加入文本模式集合$|P|$。

定义 3 候选文本模式的置信度

$$Conf(C_i)=\frac{C_{\text{positive}}}{(C_{\text{positive}}+C_{\text{negative}})} \tag{2}$$

其中，C_{positive} 表示候选文本模式；C_i 抽取的产品特征在产品特征种子集合中出现的个数；(C_{positive}+C_{negative})表示候选文本模式 C_i 抽取到的产品特征的总数。

(3)生成新的产品特征

目的是利用抽取得到的文本模式寻找新的候选产品特征，并度量候选产品特征的置信度。包括以下 2 个子任务：

1)生成侯选产品特征

利用文本模式集合中的文本模式，从产品评论的语料库抽取产品特征，如果该特征尚未在产品特征种子集合中出现，那么将该特征作为候选特征 F_i 加入候选产品特征集合$|F|$，并记录获得该候选产品特征 F_i 的文本模式$|P|$。

2)评估侯选产品特征

候选产品特征集合$|F|$中每一个候选产品特征 F_i 都可能被一个或多个文本模式得到，本文假设：能够被多个文本模式抽取得到的产品特征元组具有更高的可靠性。假设文本模式产生正确元组的概率相互独立，从而采用定义 4 中的公式来度量候选产品特征的置信度。

定义 4 候选产品特征的置信度

$$Conf(F_i)=1-\prod_{i=0}^{|P|}(1-Prob(P_i)) \tag{3}$$

其中，$Conf(F_i)$表示候选产品特征可靠性；$|P|$为抽取该候选产品特征的文本模式集合；$Prob(P_i)$表示文本模式 P_i 抽取正确产品特征的概率；$Conf(P_i)$就是模式 P_i 抽取正确产品特征

的概率[3]，因此，定义 4 变化为定义 5 中的形式。

定义 5 候选产品特征的置信度

$$Conf(F_i)=1-\prod_{i=0}^{|P|}(1-Conf(P_i)) \quad (4)$$

(4)更新产品特征种子集合

将候选产品特征集合中置信度最高的 n 个种子加入产品特征种子集合，形成新的产品特征种子集合，n 为人工制定的阈值。

3 实验结果与分析

从 Amaze 上获取的 100 个相机的用户评论作为产品评论语料库，并且以人工的方式对产品评论语料库中的产品特征进行标注，以建立一个产品特征集合，人工标注的产品特征集合包含 83 个产品特征。

实验中采用 Stanford Parser 来生成句子的句法结构和依存关系。通过简单地对产品评论语料库的观察，选用了 lens, mode, body, screen, size, picture, auto-focus 等词汇作为初始阶段的产品特征种子。

实验共进行了 16 次迭代，第 1 次迭代因为文本模式集合为空，所以一次加入了可靠性最高的 5 个文本模式，以后的迭代中每次将可靠性最高的一个文本模式加入文本模式集合。

在每次迭代过程中，将置信度最高的 5 个候选产品特征加入产品特征的种子集合，第 16 次迭代后共 87 个种子(80+7)。采用召回率和准确率表示实验结果如图 2、图 3 所示。

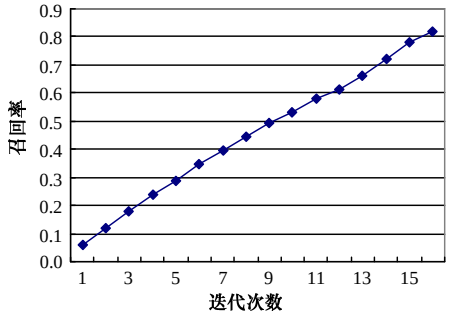

图 2 采用召回率的实验结果

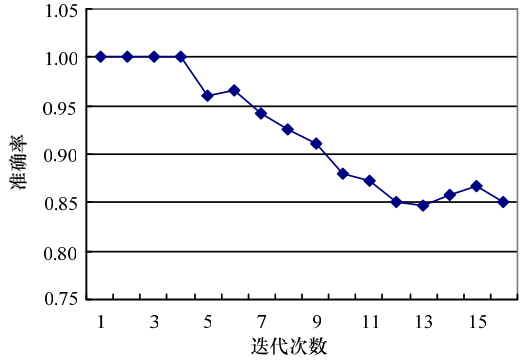

图 3 采用准确率的实验结果

文献[3-4]中的评论挖掘系统的产品特征自动抽取和本文的研究同属一个研究领域，表 1 将本文的实验结果和他们进行了比较(通过实现他们的系统在本文的语料上进行实验)。

表 1 实验结果比较

实验方法	召回率	准确率
文献[3]方法	0.80	0.71
文献[4]方法	0.77	0.89
本文方法	0.82	0.85

实验结果表明，与文献[3]的方法相比，本文的方法利用名词和名词短语所在的依存关系，句法特征等进一步度量名词是否作为产品特征，因此，准确率有了较大的提高。

与文献[4]的方法相比，召回率和精确率相差不大，但本文采用的是弱监督的学习方法，应用于一类新的产品评论，只需要提供少量的该类产品的特征，因此，具有更好的移植性。

4 结束语

本文将信息抽取领域的文本模式应用于产品特征的自动抽取，采用了基于弱监督的 BootStrapping 方法实现文本模式和产品特征的交替发现。

实验结果表明，该方法获得了较高的召回率和精确率，并且系统具有良好的移植性。下一步将致力于采用结构更加灵活的文本模式以提高准确率。

参考文献

[1] Kobayashi N, Inui K, Matsumoto Y. Collecting Evaluative Expressions for Opinion Extraction[C]//Proc. of the 1st International Joint Conference on Natural Language. Hainan, China: [s. n.], 2004: 596-605.

[2] Li Zhuang, Feng Jing, Zhu Xiaoyan. Movie Review Mining and Summarization[C]//Proceedings of the 15th ACM International Conference on Formation and Knowledge Management. Arlington, Virginia, USA: ACM Press, 2006: 43-50.

[3] Hu Minqing, Liu Bing. Mining Opinion Features in Customer Reviews[C]//Proceedings of American Association for Artificial Intelligence Conference. San Jose, USA: ACM Press, 2004: 755-760.

[4] Popescu A M, Etzioni O. Extracting Product Features and Opinions from Reviews[C]//Proceedings of Empirical Methods in Natural Language Conference on Association for Computational Linguistics. London, UK: [s. n.]. 2005: 339-346.

[5] 何婷婷, 徐 超, 李 晶, 等. 基于种子自扩展的命名实体关系抽取方法[J]. 计算机工程, 2006, 32(21): 183-184.

[6] Riloff E, Jones R. Learning Dictionaries for Information Extraction by Multi-level BootStrapping[C]//Proceedings of American Association for Artificial Intelligence Conference. Florida, USA: [s. n.], 1999: 474-479.

编辑 索书志