2013 年 11 月 第 11 期
Vol. 34 No. 11 2013

面向电子商务网站的产品属性提取算法

李 俊,陈 黎,王亚强,秦湘清,于中华

(四川大学 计算机学院,成都 610065)
E-mail: u-tyuuka@163.com

摘 要: 从商品评论中抽取作为评价对象的产品属性及判断评价的极性(正面评价、负面评价、中性评价),对于充分挖掘利用电子商务网站上积累的大量商品评论,为消费者的购物决策和生产者的生产决策提供支持,具有重要意义.本文针对现有算法的不足,结合中文电子商务网站中商品评论的特点,提出了综合模板、频率和 HITS 的无监督学习算法,用于从中文商品评论中识别产品属性.充分的实验结果表明,所提出的无监督算法对产品属性识别的 F 值可以达到 77.3%,优于文献中提出的其他类似算法.

关 键 词: 商品评论; 产品属性; 抽取; HITS; 抽取模板
中图分类号: TP391 **文献标识码:** A **文 章 编 号:** 1000-1220(2013)11-2477-05

Algorithm for Extracting Product Attributes from E-commerce Websites

LI Jun, CHEN Li, WANG Ya-qiang, QIN Xiang-qing, YU Zhong-hua
(College of Computer Science, Sichuan University, Chengdu 610065, China)

Abstract: Extracting product attributes which are treated as comment targets from product review and determining the opinion orientation (positive, negative and neutral) are of great significance for fully mining a large number of product reviews on e-commerce website and providing support for shopping decision, production decision of consumers and producers respectively. In view of the deficiency of existing approaches and combining with the characteristics of the product reviews in Chinese e-commerce sites, this paper proposes an unsupervised method which integrates POS patterns, term frequency and HITS to identify product attributes from Chinese product reviews. The experimental results show that the proposed method can reach an F-score of 77.3% on product attributes extraction, outperform the existing other methods.

Key words: product reviews; product attributes; extraction; HITS; extraction patterns

1 引 言

随着电子商务的迅速发展,网上购物已经成为一种时尚.借助于 Web 2.0 技术,用户在购物的同时,还可以撰写商品评论,或给商品打分.通过上述方式产生的商品评论,蕴涵着大量丰富的信息,对消费者的购物决策和生产者的生产决策都具有重要价值.然而,由于这些评论的数量庞大,且多为短文本描述,手工分析和挖掘异常困难,因此,设计实现面向商品评论的文本挖掘工具变得越来越迫切,成为自然语言处理和数据挖掘界的热点问题之一.

尽管商品评论的文本挖掘研究只有短短的十年历史,却取得了丰富的成果.早期的研究重点是文档层面的情感分类,其目的是判断一篇商品评论的整体情感极性.由于商品评论整体情感极性的应用范围有限,消费者单纯依靠对产品的整体评价无法确定该产品在自己关注的属性方面是否具有良好的口碑,一个整体上具有良好口碑的产品,不一定在每一个属性方面都具有良好的口碑.因此,产品属性识别及面向产品属性的情感分析,成为商品评论文本挖掘的亟待解决的问题,正引起广泛的关注和研究[1].

本文的研究内容是从电子商务网站的中文商品评论中识别产品属性,为后续的细粒度情感分析做准备.这里的"产品属性"包括与产品相关的任何可能的评价对象,包括产品本身、产品的构件(成分)、产品本身的属性、产品构件的属性等,例如,在评论"手机屏幕很清晰."中,"手机"和"屏幕"都是需要识别的产品属性,其中"手机"是产品本身,"屏幕"是"手机"的构件.评论"手机电池的待机时间太短"中,"手机"、"电池"、"待机时间"是需要识别的产品属性,其中"手机"是产品本身,"电池"为其构件,"待机时间"为构件"电池"的属性.

本文针对现有算法的不足,结合中文电子商务网站中商品评论的特点,提出了综合模板、频率和 HITS 的无监督学习算法,用于从中文商品评论中识别产品属性.充分的实验结果表明,所提出的无监督算法对产品属性识别的 F 值可以达到 77.3%,优于文献中提出的其他类似算法.

2 相关工作

对网上大量商品评论进行细粒度的情感分析,由于其重要的应用价值,引起了研究者的广泛关注.所谓细粒度的情感分析,是指从商品评论中识别抽取出评价的对象及对该对象的评价极性,从而为进一步的分析汇总、个性化的商品推荐以

收稿日期: 2012-05-16 收修改稿日期: 2012-09-11 基金项目: 高等学校博士学科点专项科研基金项目(20100181120029)资助. 作者简介: 李 俊,男,1987 年生,硕士研究生,研究方向为数据挖掘与计算语言学; 陈 黎,女,1977 年生,博士,研究方向为数据挖掘与计算语言学; 王亚强,男,1984 年生,博士研究生,研究方向为数据挖掘与计算语言学; 秦湘清,男,1986 年生,硕士研究生,研究方向为数据挖掘与计算语言学; 于中华(通信作者),男,1967 年生,博士,副教授,研究方向为数据挖掘与计算语言学.

及分析用户的喜好等应用打下基础. 显然,产品属性的自动识别是细粒度情感分析的基础和前提,而且也是其中的难点问题之一. 因此,从商品评论文本中识别产品属性描述,成为最近几年来自然语言处理和文本挖掘界热门的研究课题之一,国际上一些著名的评测如 TREC Blog Track 和 NTCIR 等都将产品属性识别作为其任务之一,而国内第一届中文倾向性分析评测 COAE2008[1] 的任务三即为产品属性抽取(COAE2008 共设有六个评测任务).

从商品评论中自动识别抽取产品属性的研究始于 2004 年[2],研究者提出了一系列算法,这些算法可以归结为有监督和无监督两大类. 有监督的产品属性抽取需要人工标注的训练语料,在此基础上将产品属性识别和抽取问题看成是标注问题. 这方面的工作包括[7-12]. 文[7,11]提出了一种基于词汇化隐马尔可夫模型(Lexicalized HMM)的产品属性自动识别算法,为了减少有监督学习所需要的手工标注量,采用了 Bootstrapping 方法进行自动标注. 文[8-10]均采用 CRF 进行产品属性抽取,不同点在于[8]以潜在属性词前后文中的单词及其相关信息作为特征,而[9]首先使用规则从原始句中提取出核心子句,再以核心句的句法结构信息作为特征,[10]采用的特征包括前后文中的词、词性、与情感词之间的依存关系、与情感词之间的距离等,并且实验验证了 CRF 和这些特征的领域通用性,[12]从人工标注的训练集中学习抽取规则(类似于关联规则),利用学习到的规则对测试评论进行产品属性抽取.

尽管有监督的方法可以取得较好的产品属性抽取效果,然而需要人工标注大量的训练样例,这是一项既枯燥又费时费力的工作,而且不同领域(如电子类商品评论和图书类商品评论)的训练数据是无法共享的,这大大影响了有监督方法的适用性. 因此,无监督成为目前产品属性抽取方面更被普遍采用的方法,并进行了深入的研究,这方面的工作包括[2,5,13-16].

文[2]首次提出了产品属性的自动抽取问题,提出了采用关联规则发现热点属性、根据与热点属性邻接的形容词进一步识别其他属性的方法,以名词短语作为候选属性,在关联规则发现频繁模式后进一步进行紧凑型剪枝(Compactness Pruning)和冗余性剪枝(Redundancy pruning). 文[2]算法的不足是需要设置最小支持度阈值等多个参数,这些设置对算法抽取效果的影响又与商品评论的数量等密切相关,因此,普通用户很难胜任这样的设置工作. 此外,由于假定热点属性局部前后文内的形容词为情感词,被情感词修饰的名词性短语也是产品属性(低频属性),因此,对低频属性的识别显得粗糙,准确率不高. 为此,文[5,13-16]对其进行了进一步的改进.

文[5]直接将[2]的方法应用于中文商品评论的产品属性抽取上,达到了 71.05% 的 F 值,接近[2]在英文商品评论上达到的 73.89%,[13]将[2]的方法进一步发展,提出了根据属性和情感词之间的依赖关系,利用种子情感词和双向传播策略,增量地识别新情感词和产品属性,属性和情感词之间的依赖关系在依存分析的基础上确定,而[2]根据简单的邻接关系来确定. 文[14]利用[2]的关联规则挖掘方法识别频繁属性后,进一步利用已知的情感词对这些频繁属性进行过滤(根据与情感词的共现情况),以提高热点产品属性识别的精度. 该方法的不足是要求情感词表作为系统输入,而面向特定商品领域的情感词表的构造和维护是件繁琐复杂的工作,特定领域情感词的识别本身也是细粒度情感分析的一个难点和重点[13]. 文[15]利用不可能出现在产品属性中的单词词表来对关联规则的挖掘结果进行过滤,而这种词表的构造甚至难于情感词表的构造,用户必须对被处理商品的相关情况有了全面了解后才可能胜任构造这种词表的工作. 文[16]对[2]算法用于中文时的一些剪枝策略进行了调整和实验.

文[3]提出了另外一种产品属性的抽取方法,该方法源自 Konwitall 信息抽取系统[4]. 具体思想是:利用[4]中系统的 8 条领域无关模板抽取产品属性的候选,例如要抽取扫描仪的属性,可以利用模板"NP1 such as NPList2",NP1 实例化成"Scanner",与 NPList2 匹配的名词短语即作为扫描仪产品属性的候选. 对每一个产品属性候选,计算其与品名(对于前面的例子,品名为 Scanner)的点互信息值(PMI,根据产品属性候选和品名在搜索引擎中的命中情况估计概率,利用概率计算点互信息),互信息越大,表示候选确实为产品属性的可能性越大. 这种方法的精确率比[2]高 22%,召回率低 3%,但不足之处是需要用户给定产品名. 由于商品评论的口语化风格,同物异名的现象更加普遍存在,因此,很难保证用户给定的产品名在评论中出现,而且多数产品属性都伴随有相应的品名. 文[17]是[3]针对中文的应用,具体方法是:基于百度百科和词的共现信息识别专业领域生词,然后利用词性序列模板抽取产品属性候选,对每个候选计算其与领域典型产品属性(如对于手机领域,手机、屏幕、电池等为典型产品属性,这些典型产品属性由人工给定)的点互信息(利用 Google 搜索引擎的输出计算得到)进行过滤. 此外,还利用人工搜集的不可能出现在属性名中的动词词表进行过滤. 该方法对中文手机评论的精确率为 56.5%,召回率为 72.6%,F 值为 63.6%. 文[17]方法的不足是需要人工确定产品的典型属性,还需要不可能出现在属性名中的动词词表的支持,因此,已经不属于完全的无监督方法. 文[6]在词性标注和句法分析的基础上,利用频率过滤、PMI 值过滤和名词剪枝三步实现产品属性的识别,PMI 值的计算同样利用了搜索引擎. 该方法在 COAE2008 评测语料[1]上达到 F 值 51.69%.

最近,以 PLSA 和 LDA 为代表的主题模型被引入到无监督的产品属性抽取中,成为目前产品属性抽取方面的热点研究方向之一,这方面的工作包括[18-22]. 这些方法一般假定每个产品属性对应一个局部主题(Local Topic),即一个定义在词表上的多项式分布,利用生成模型实现对评论中单词的聚类,即确定每个单词属于每个局部主题的概率,这样,不但可以识别作为产品属性的单词,而且可以解决同物异名的问题. 然而,已经提出的模型多数不能显式地区分开产品属性和情感词,有些虽然可以区分开,但是需要利用一些额外的信息资源来评价模型的部分参数,如有些需要初始的种子情感词表[22],有些需要用有监督的最大熵模型来评价主题模型的部分参数[20]. 此外,目前还没有明确的证据可以证明主题模型的效果优于其他无监督算法的效果.

中文产品属性的自动抽取,由于在分词、短语识别、未登录词识别、句法分析等方面存在特殊的困难,还远未达到英文那样的抽取效果,如[1]对 COAE2008 所综述的那样,"中文

词语倾向性评测的结果基本上能够达到较满意的效果,评价对象抽取的评测方面,最好的结果也难令人满意".

本文针对现有中文产品属性抽取算法的不足,提出了一种无监督的自动抽取算法,该算法使用领域无关的词性模版抽取产品属性的候选,然后结合频率和 HITS 分析进行筛选,确定最终的产品属性. 实验结果表明,该方法对高频和低频的产品属性都可以进行有效的识别.

3　方　法

3.1　基本思想

一般来说,商品评论中频繁出现的产品属性代表广受关注的热点属性,这些属性相对于其他属性对细粒度的情感分析具有更重要的应用价值,以 [2] 为代表的无监督算法对这类属性的识别可以达到较高的精度,但对低频属性的识别效果较差. 为了进一步提高产品属性识别的有效性,解决现有方法无法有效识别低频属性的问题,本文提出以下基本观点: 一个商品评论包含的不同产品属性的个数和它们的重要程度决定了该商品评论的重要程度,而包含一个产品属性的不同商品评论的个数和它们的重要程度决定了该产品属性的重要程度,产品属性和商品评论的重要性相互促进,越重要的候选属性成为真正产品属性的可能性越大. 基于上述观点,本文采用 HITS 算法的思想对候选产品属性按重要性进行排序和筛选,从而实现对产品属性的识别. 图 1 给出了算法的处理流程,本章后续部分将对其中的关键步骤进行详细的介绍和讨论.

```
Input:
P←set of reviews
Output:
F←set of product attributes
1.    F←φ
2.    foreach p in P do
3.       p_pos←POS( p)    //词性标注
4.       W←extract( p_pos)    //候选产品属性抽取
5.       F←F∪W
6.    F←pruner( F,P)    //候选产品属性剪枝
7.    A←HITS( F,P)    //HITS 模型计算候选属性的权威值
8.    sort F by A desc //按权威值降序排序
9.    F←choose top-k elements in F
10. return F
```

图 1　算法处理流程

Fig. 1　Algorithmic process

3.2　候选产品属性抽取

通过对大量中文商品评论的分析发现,产品属性多数是名词和名词短语. 因此,本文将识别候选产品属性的词性模板定义为

候选产品属性→NN | NP

NP→动词 + NP

NP→(动词 | NN) + NN

NN→名词 | 名动词

算法首先利用分词工具 ICTCLAS (http: //ictclas. org/) 对商品评论进行分词和命名实体识别, 然后利用上述模板识别候选产品属性,并且限定 NP 最多由 3 个单词组成,同时排除掉时间、人名、地名等命名实体.

3.3　候选产品属性剪枝

为了提高算法的有效性,在利用 HITS 进行重要性评价之前,对 3.2 节得到的候选产品属性进行如下的剪枝处理:

(1) 去除单字名词形式的候选属性. 根据观察发现,以单字名词表示产品属性的情况非常少见.

(2) 去除明显不可能成为产品属性的名词,这些名词由人工收集,组成产品属性禁忌词表. 本文实验所采用的产品属性禁忌词表包括"东西"、"机子"、"手机"、"优点"和"缺点".

在对京东商品评论进行属性识别实验时(实验数据和实验结果详见第 4 章) ,使用这 5 个词构成的产品属性禁忌词表可以过滤掉 5.2% 的非单字候选产品属性.

(3) 独立支持度(p-support) 剪枝[2]

候选属性 w 的独立支持度定义为

$$psp(w) = count(w) - \sum_{w_i \in p_w} count(w_i) \quad (1)$$

其中 p_w 表示 w 所有父串的集合,$count(w_i)$ 表示出现 w_i 的句子数.

例如,假设"物流"为候选属性,包含它的句子数为 10,而这 10 个句子中有 7 个包含了"物流速度",那么"物流"的独立支持度为 10 - 7 = 3.

实验中设定独立支持度阈值为 1,对于候选属性 w,如果 $psp(w) < 1$,则 w 将被从候选属性集中过滤掉.

(4) 邻近度(Proximity) 剪枝

候选属性 w 在任意评论 p 中的邻近度定义为

$$Proximity(w,p) = 1/(distance(w,p) + 1) \quad (2)$$

其中 $distance(w,p)$,表示在评论 p 中 w 与最近的形容词或副词之间的距离,以词的个数来度量. 如果评论 p 中没有形容词或副词,$distance(w,p)$ 的值为 $+\infty$; 如果 w 在 p 中出现多次,则计算 w 每次出现对应的 $distance(w,p)$,取其中最小值作为 $distance(w,p)$.

例如,"屏幕大而且很清晰"经分词后变成"屏幕/n 大/a 而且/c 很/d 清晰/a",其中"屏幕"为候选属性,"大"和"很"为形容词和副词,但"大"离属性"屏幕"更近,因此邻近度为 1/(0 + 1) = 1.

得到每个候选属性在所有评论中的邻近度后,根据预先设定的阈值 δ 对候选属性集进行进一步的过滤,如果任意候选属性 w 的 $\max\{Proximity(w,p_i) \mid p_i \in P\} < \delta$,则从候选属性集中滤掉 w(其中 P 是所有商品评论的集合) . 本文实验时 δ 取 1/3.

3.4　HITS 算法

HITS 算法是 Web 结构挖掘中最具权威性和使用最广泛的算法之一,主要用于对搜索引擎检索到的结果网页进行排序. HITS 算法的主要思想是将网页分为权威页(Authority) 和中心页(Hub) ,根据网页的入度(指向一个网页的超链接数) 和出度(从一个网页指向其他网页的超链接数) 来衡量网页的重要性. HITS 算法隐含的基本假设是: 一个优秀的中心页会指向很多优秀的权威页,而一个优秀的权威页又会被很多优秀的中心页所指向,权威页和中心页之间存在一种相互促进的关系[23].

本文利用 HITS 算法的思想来评估一个候选属性的重要程度,将候选属性看作权威节点,评论看作中心节点,候选属性与评论之间的关系用有向图 $G = (W,P,E,L)$ 来表示(如图 2 所示) ,其中 W 表示候选属性的集合,P 表示评论的集合,E

$=\{e_{ij} \mid w_j \in p_i, p_i \in P, w_j \in W\}$ 为边的集合，e_{ij} 表示属性 w_j 出现在评论 p_i 中，$L=(l_{ij})_{n\times m}$ 是边的权值矩阵，其中 l_{ij} 为边 e_{ij} 的权重，定义为

$$l_{ij}=\begin{cases} Proximity(w_j,p_i), w_j \in p_i \\ 0, w_j \notin p_i \end{cases} \quad (3)$$

用 $A(w_j)$ 表示候选属性 w_j 的权威值，$H(p_i)$ 表示评论 p_i 的中心值，这样，候选属性的权威性与评论的中心性之间相互促进的关系即可表示为

$$A(w_j)=\sum_i^n l_{ij} \cdot H(p_i) \quad (4)$$

$$H(p_i)=\sum_j^m l_{ij} \cdot A(w_j) \quad (5)$$

若用 $\vec{A}$ 表示所有候选属性权威值的列向量，$\vec{H}$ 表示所有评论的中心值列向量，则上式可以表示成如下的矩阵形式:

$$\vec{A}=L^T\vec{H} \quad (6)$$

$$\vec{H}=L\vec{A} \quad (7)$$

由于本文的目的是对候选属性进行排序，进而识别作为评价对象的产品属性，因此只需要计算 $\vec{A}$，计算方式为: 首先将 $\vec{A}$ 初始化为 $\vec{A}^{(0)}=(1,1,\cdots,1)^T$，然后迭代地计算新的 $\vec{A}$，直至收敛. 根据(6)(7)，迭代公式为

$$\vec{A}^{(k+1)}=L^TL\vec{A}^{(k)} \quad (8)$$

其中 $\vec{A}^{(k)}$ 表示第 k 轮迭代得到的 $\vec{A}$.

计算 $\vec{A}$ 后，按权威值大小对候选属性进行排序，选取 Top K 候选属性作为最后识别出的产品属性.

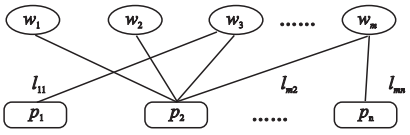

图 2 评论和和产品属性之之间的关系图图

Fig. 2 Relation graph of reviews and product attributes

4 实 验

4.1 实验数据与与评价指标标

本文的实验语料从京东网上商城(http://www.360buy.com)获取，该网站是国内著名的 B2C 购物网站站. 首先从该网站收集了 55 种不同的热门商品的评论，包括 Nokia5233(手机)，Cannon G12(相机)，iPod shuffle(MP3 播放器)，Panasonic S60(DVD 播放器)和 Acer VX275(电脑). 从每种商品的评论中随机抽取了 100 条，人工对其中的产品属性进行标注，形成标准的测试集.

算法的识别效果采用召回率(Recall)、准确率(Precision)和 F 值(F-score)进行评价，它们的定义如下:

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

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

$$\text{F-score}=\frac{2\times\text{Recall}\times\text{Precision}}{\text{Recall}+\text{Precision}}$$

4.2 实验结果

使用 HITS 模型计算候选属性权威值并排序后，需要选取权威值最高的 Top K 候选属性作为最终的产品属性. 为了评价不同的过滤比例(候选属性去除 Top K 剩下的部分)与抽取效果的关系，实验时尝试了不同的过滤比例，结果如图 3 所示，其中横轴表示过滤比例，纵轴表示抽取结果的 F 值，avg 表示平均 F 值.

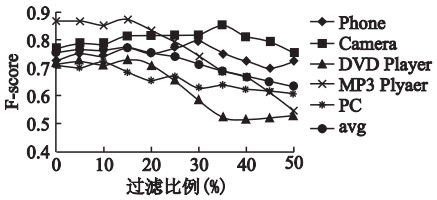

图 3 不同过滤比例下的实验结果

Fig. 3 Results under different filtering ratios

从图 3 很容易看到，当过滤比例为 15% 时，avg 最大，达到 77%. 表 1 给出了这种情况下(过滤比例取 15%)每种产品属性的具体抽取结果，包括 Precision、Recall 和 F-Score.

表 1 产品属性抽取结果

Table 1 Results of product attribute extraction

Prodduct	PPrecision	Recall	F-score
phone	0.6782	0.8696	0.7620
camera	0.6957	0.9697	0.8101
MP3 Player	0.8148	0.9565	0.8800
DVD Player	0.6857	0.7742	0.7273
PC	0.5999	0.7969	0.6845
Avg	0.6949	0.8734	0.7728

由表 1 可知，本文算法对产品属性抽取的平均精确率为 69%，平均召回率为 87%，其中 MP3 player 的属性抽取结果最理想，而 PC 的属性抽取结果最差.

为了进一步验证本文算法的有效性，考虑到 Hu 等[2]提出的方法与相关研究中的其他无监督方法相比具有更高的精确率和召回率，因此，使用本文的测试语料对文献[5]的算法(改进自[2]，用于处理中文商品评论，后文用 Hu + 标识))和文献[6]的算法(后文用 Liu 标识)进行了测试，并与本文算法的结果进行对比，结果图 4 和表 2 所示，其中的精确率和召回率是最大 F 值情况下的精确率和召回率.

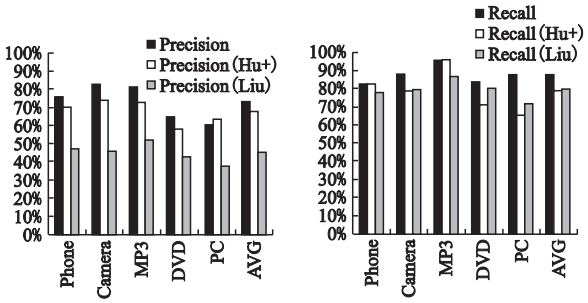

图 4 产品属性抽取结果对比

Fig. 4 Comparison of attribute extraction for diflerent products

从图 4 和下页表 2 可以看到，本文提出的算法与 Liu 和 Hu + 相比，F 值分别有 21.8% 和 6.7% 的提高. Liu 虽然使用了 PMI 值过滤，但实际效果并不理想. Hu + 采用关联规则的方法，单纯依据频率信息对候选进行过滤，导致低频产品属性无法有效识别. 而本文使用 HITS 算法，综合考虑了候选属性的出现频率和包含候选属性评论的权威性，据此对候选属性

进行综合评估. 这样,一个低频产品属性也可能被识别出来,只要它出现在重要的评论中.

表 2　不同方法结果对比

Table 2　Comparison betueen diflerent methods

	Hu +	Liu	本文
Precision	0.679	0.453	0.732
Recall	0.787	0.794	0.875
F-measure	0.728	0.577	0.795

5　结束语

本文提出了一种无监督的产品属性自动抽取算法,该算法综合利用模板、频率和 HITS 对候选属性进行评估,在此基础上从商品评论中识别抽取产品属性. 充分的实验结果表明,所提出的无监督算法对产品属性识别的 F 值可以达到 77.3%,优于文献中提出的其他类似算法. 对本文工作进一步完善的方向是把本文算法与态度挖掘方法相结合,实现面向产品属性的态度倾向判断,同时分析本文算法对面向属性态度挖掘的作用和效果.

References:

[1] Zhao Jun, Xu Hong-bo, Huang Xuan-jing, et al. Overview of Chinese opinion analysis evaluation [C]. Proceedings of the Chinese Opinion Analysis Evaluation 2008, 2008: 1-20.

[2] Hu Min-qing, Liu Bing. Mining opinion features in customer reviews [C]. Proceedings of the Association for the Advancement of Artificial Intelligence, 2004: 7552-760.

[3] 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, 2005: 339-346.

[4] Etzioni O, Cafarella M, Downey D, et al. Un-supervised named-entity extraction from the web: an experimental study [J]. Artificial Intelligence, 2005, 165(1): 91-134.

[5] Li Shi, Ye Qiang, Li Yi-Jun, et al. Mining product features and sentiment orientation from Chinese customer reviews [J]. Application Research of Computers, 2010, 27(8): 3016-3019.

[6] Liu Hong-yu, Zhao Yan-yan, Qin Bing, et al. Comment target extraction and sentiment classification [J]. Journal of Chinese Information Processing, 2010, 24(1): 84-88.

[7] Wei Jin, Hung Hay Ho. A novel lexicalized HMM-based learning framework for Web opinion mining [C]. Proceedings of the 26th International Conference on Machine Learning, New York, 2009: 465-472.

[8] Zhang Shu, Jia Wen-jie, Xia Ying-ju, et al. Research on CRF-based evaluated object extraction [C]. Proceedings of the Chinese Opinion Analysis Evaluation 2008, Peking, 2008: 70-76.

[9] Zhang Li, Qian Ling-fei, Xu Xin. Comment target extraction based on nuclear sentences and syntactic relations [J]. Journal of Chinese Information Processing, 2011, 25(3): 23-29.

[10] Niklas Jakob, Iryna Gurevych. Extracting opinion targets in a single- and cross-domain setting with conditional random fields [C]. Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing (EMNLP 10), Stroudsburg, USA, 2010: 1035-1045.

[11] Wei Jin, Hung Hay Ho, Rohini K Srihari. OpinionMiner: a novel machine learning system for Web opinion mining and extraction [C]. Proceedings of Kokusai Denshin Denwa´09, NY, USA, 2009: 1195-1204.

[12] Bing Liu, Minqing Hu, Junsheng Cheng. Opinion Observer: analyzing and comparing opinionson the Web [C]. Proceedings of the 14th International Conference on World Wide Web (WWW ´05), NY, USA, 2005: 342-351.

[13] Qiu Guang, Liu Bing, Bu Jia-jun, et al. Opinion Word expansion and target extraction through double propagation [C]. Proceedings of the 2nd Workshop on Computational Approaches to Subjectivity and Sentiment Analysis, PA, USA, 2011: 125-131.

[14] Chih-Ping Wei, Chen Yen-ming, Chin-Sheng Yang, et al. Understanding what concerns consumers: a semantic approach to product feature extraction from consumer reviews [J]. Information Systems and E-business Management, Springer, Heidelberg, ALLEMAGNE, 2010, 8(2): 149-167.

[15] Li Shi, Ye Qiang, Li Yi-Jun, et al. Research on mining product features in Chinese customer reviews [J]. Journal of Management Sciences in China, 2009, 2(2): 142-152.

[16] Li Shi, Li Qiu-shi. Research on pruning algorithm of product feature mining in Chinese review [J]. Computer Engineering, 2011, 37(23): 43-45.

[17] Li Chun-liang, Zhu Yan-hui, Xu Ye-qiang. Research of attribute word extraction method in chinese product comment [J]. Computer Engineering, 2011, 37(12): 26-28.

[18] Ivan Titov, Ryan McDonald. Modeling online reviews with multi-grain topic models [C]. Proceeding of the 17th International Conference on World Wide Web (WWW-2008), NY, USA, 2008: 21-25.

[19] Samuel Brody, Noemie Elhadad. An unsupervised aspect-sentiment model for online reviews [C]. Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics, Los Angeles, California, 2010: 804-812.

[20] Wayne Xin Zhao, Jiang Jing, Yan Hong-fei, et al. Jointly modeling aspects and opinions with a MaxEnt-LDA hybrid [C]. Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing (EMNLP 10), Cambridge, Massachusetts, 2010: 56-65.

[21] Yohan Jo, Alice Oh. Aspect and sentiment unification model for online review analysis [C]. Proceedings of the Fourth ACM International Conference on Web Search and Data Mining, 2011.

[22] Lin Cheng-hua, He Yu-lan. Joint sentiment/topic model for sentiment analysis [C]. Proceedings of the 18th ACM Conference on Information and Knowledge Management, NY, USA, 2009: 375-384.

[23] Bing Liu. Web data mining [M]. New York: Springer, 2009.

附中文参考文献:

[1] 赵　军,许洪波,黄萱菁,等. 中文倾向性分析评测技术报告 [C]. 第一届中文倾向性分析评价论文集,2008: 1-20.

[5] 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.

[6] 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.

[8] 张　姝,贾文杰,夏迎炬,等. 基于 CRF 的评价对象抽取技术研究 [C]. 第一届中文倾向性分析评价论文集,2008: 70-76.

[9] 张　莉,钱玲飞,许　鑫. 基于核心句及句法关系的评价对象抽取 [J]. 中文信息学报, 2011,25(3): 23-29.

[15] 李　实,叶　强, 李一军,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, 2009,2(2): 142-152.

[16] 李　实,李秋实. 中文评论中产品特征挖掘的剪枝算法研究 [J]. 计算机工程, 2011,37(23): 43-45.

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