文章编号：1003-0077(2016)05-0136-09

产品评论中领域情感词典的构建

郗亚辉

（河北大学 数学与计算机学院，河北 保定 071002）

摘　要：领域情感词典是情感分析最重要的基础。由于产品评论的数量巨大、领域众多，如何自动构建领域情感词典已经成为近年来的一个研究热点。该文提出了一个两阶段的领域情感词典构建算法。第一阶段，利用情感词间的点互信息和上下文约束，使用基于约束的标签传播算法构造基本情感词典；第二阶段，根据情感冲突的频率来识别领域相关情感词，并根据其上下文约束以及修饰的特征完善领域情感词典。实验结果表明，该方法在实际产品评论数据集上取得了较好的效果。

关键词：情感分析；领域情感词典；上下文约束；基于约束的标签传播算法

中图分类号：TP391　　**文献标识码**：A

Construction of Domain-specific Sentiment Lexicon in Product Reviews

XI Yahui

(College of Mathematics and Computer Science, HeBei University, Baoding, Hebei 071002, China)

Abstract: Domain-specific sentiment lexicon plays an important role in sentiment analysis system. Due to the huge number of the product review in diverse domains , automatic construction of domain-specific sentiment lexicon is a challenging task. This paper proposes a two-phrase automatic construction algorithm of domain-specific sentiment lexicon. In the first phrase, the constrained label propagation algorithm is applied to the construction of base sentiment lexicon by using PMI and contextual constraints. In the second phrase, the domain-specific sentiment words are exacted by the frequency of sentiment conflict, and the domain-specific sentiment lexicon is improved according to the contextual constraints and the product feature modified by the sentiment word. Experiments on diverse real-life datasets show promising results.

Key words: sentiment analysis; domain-specific sentiment lexicon; contextual constraints; constrained label propagation algorithm

1　引言

飞速发展的 Web 技术及电子商务正在极大改变着人们的工作和生活方式，越来越多的人习惯于网上购买商品，网络也成为各种产品的重要销售渠道。为了提高消费者的购物体验，电子商务网站大都允许消费者对其购买的产品发表评论。在这些产品评论中，包含了大量的消费者对产品各方面特征的评价观点信息。这些信息不仅可以帮助消费者全面、综合地了解其他消费者对产品的评价，从而挑选出更适合自己的产品；还可以帮助生产厂商通过评论来了解自己产品的优点以及不足，从而改进产品的设计、获得竞争优势[1-2]。

网络上存在着大量的产品评论，有些热门商品可能包含成千上万的评论。巨大的评论数量使得消费者和生产厂商很难通过人工对产品评论进行分析和处理，获取产品评论中包含的大量有用信息。因此，以获取产品评论中有用信息为目标的非结构化数据挖掘技术——“评论挖掘”，吸引了越来越多学者的关注。

评论观点的情感分析是产品评论挖掘的基本任务之一[1]，其目标是准确识别出消费者对产品不同特征所发表评价观点的情感倾向——褒义或贬义。

收稿日期：2014-12-04　**定稿日期**：2015-06-09

基金项目：国家自然科学基金(61170039)

情感词或词组是人们表达观点的最基本的语言单元，情感词典则是情感分析的基础。近年来，很多学者已经建立了一些情感词典，例如，General Inquirer[3]，Liu 提供的情感词典[1]，SentiWordNet[4]，知网的情感分析用词语集[5]，台湾大学的 NTU 情感词典[6]以及大连理工大学的情感词汇本体库[7]。这些词典主要是以手工或半自动的方式编辑生产，其领域适应性受到限制。

有些情感词在不同领域中具有不同的情感倾向，甚至在同一领域中当修饰不同产品特征时也具有不同的情感倾向。例如，在手机评论中，“高”修饰“价格”时表示褒义，而修饰“屏幕分辨率”时表示贬义。因此，使用通用的情感词典无法准确获取这些领域相关情感词的情感倾向。由于产品评论的数量巨大、领域众多，不可能依靠领域专家人工建立领域情感词典。所以如何自动或半自动地建立领域情感词典已经成为情感分析的重要工作。

本文的主要工作是讨论如何利用情感词的共现信息以及情感词上下文的先验知识来自动建立领域情感词典。我们的工作基于以下两个假设：(1)在产品评论中，情感词之间存在一些固有的先验知识。例如，并列关系的情感词往往具有相同的情感倾向，而转折关系的情感词往往具有相反的情感倾向；(2)领域情感词典中的情感词可以分为领域无关和领域相关的两部分。领域无关的情感词在不同领域中保持相同的情感倾向，而领域相关的情感词在不同领域中可能具有不同的情感倾向。依靠一些先验知识，可以通过上下文中领域无关情感词的情感倾向来推测领域相关情感词的情感倾向。例如，在句子“屏幕大，外观漂亮”中，虽然我们不知道“大”的情感倾向，但我们可以通过“漂亮”来推测“大”的情感倾向。

基于上面的假设，本文在文献[8]的基础上提出了一个两阶段的领域情感词典构造方法。第一阶段，利用情感词间的共现信息和上下文约束关系，使用基于约束的标签传播算法构造基本情感词典，为每一个情感词分配固定的情感倾向；第二阶段，识别领域相关情感词，并根据其在语料中的上下文信息对情感词修饰的不同特征分配不同的情感倾向。

2　相关研究

近年来，情感词典的构建已经成为很多学者关注的问题。情感词典构建的方式主要分为两类：基于词典资源的方法和基于语料库的方法。

2.1　基于词典资源的方法

基于词典资源的方法主要利用现有的一些词典资源(例如，英文的 WordNet、GI，中文的 HowNet、同义词词林)中词之间的同义词、反义词等联系以及词的注释来建立情感词典。

Hu 和 Liu[9]人工选取了一些褒义和贬义的形容词作为种子集，并利用 WordNet 的同义词和反义词联系对种子集进行扩展建立情感词典。Kamps 等[10]利用 WordNet 的同义词集构建形容词之间的联系，如果两个形容词是同义词则在它们之间建立一条边，从而构成了一张图。情感词的倾向由其在图中与“good”和“bad”的最短距离决定。Rao 和 Ravichandran[11]利用 WordNet 的同义词、上位词联系来构建词之间的边，从而形成了一张图。同时，给出了一个包含褒义和贬义词的训练集，使用基于图的半监督学习算法 mincuts、randomized mincuts 和 label propagation 将图中的点划分为褒义和贬义两类。Esuli 等[4,12]人工建立了褒义词、贬义词、中性词种子集，利用 WordNet 的同义词联系来扩展种子集，然后利用扩展结果同义词集的注释文本作为训练集建立分类器来判断词的情感倾向。

朱嫣岚等[13]选择了 k 对褒义、贬义的基准词，利用 HowNet 的语义相似度和语义相关场两种计算方法，计算一个词与褒义和贬义基准词集的相似度的差值作为该词的情感倾向分值。路斌等[14]利用同义词词林中的同义词词群，根据褒贬义种子词进行扩展，从而建立情感词典。徐琳宏等[7]结合现有的一些词典、语义网络资源以及情感语料，采用手工情感分类和自动获取强度两种方法构建了情感词汇本体。周咏梅等[15]首先利用 HowNet 获取中文词语对应的各项英文义元；其次使用 SentiWordNet 数据库检索每个英文义元所处的各个同义词集合；接着计算这些同义词集合的平均情感强度值得到每个义元的情感倾向性强度值；最后计算各项义元的平均情感强度值，即得到中文词语的情感倾向强度值。

2.2　基于语料库的方法

基于语料库的方法假设在语料库中共同出现的情感词拥有相同的情感倾向，利用语料中的共现信息、上下文信息等计算情感词的情感倾向。

Turney[16]利用一些特定的语法模式抽取形容词和副词作为候选情感词，然后计算情感词与“excellent”和“poor”之间的点互信息(PMI)的差值来判别其情感倾向。PMI使用搜索引擎AltaVista返回的hits值计算每个词与种子情感词的相似度。Turney和Littman[17]进一步将初始的褒义和贬义词种子集扩展为七个词，并计算词和种子集点互信息的综合值来判断情感词的情感倾向。Hatzivassiloglou等[18]利用大规模语料中的连接词来识别形容词的情感倾向，首先使用对数线性回归模型(log-liner regression model)预测由不同连接词连接的形容词对是否具有相同或相反的情感倾向，然后根据形容词之间的联系利用聚类算法将形容词聚为褒义和贬义的两类。Kanayama和Nasukawa[19]提出了一种无监督的算法建立领域情感词典。首先，他们建立了初始的具有明确情感倾向(词的情感倾向和领域无关)的情感词典，然后通过分析领域相关语料中语句内部和语句间的文本和连接词来获取新词的情感倾向从而扩展情感词典，最终形成特定领域的情感词典。Ding和Liu[20]考虑了即使在同一领域中，修饰不同产品特征时某些情感词也具有不同的情感倾向，利用语句内和语句间的文本和连接词来判断描述特定产品特征的情感词的情感倾向。Lau等[21]不仅利用了情感词之间的上下文关系，而且利用了文档和情感词间的关系来建立领域情感词典。Huang等[8]使用句法分析和主观线索字典抽取情感词，然后根据PMI建立情感词之间的联系图，并抽取语言学规则(例如，un、dis等前缀修饰的词一般和原词表示相反的情感倾向)以及语料中的并列、转折关系作为限制条件。结合情感词间的联系图以及限制条件，利用基于约束的标签传播算法来获取情感词的情感倾向。

王素格，李德玉等[22]在利用PMI计算中文词的情感倾向时，除了考虑一个词和褒义词、贬义词种子集的关系外，还考虑了该词和其同义词集的关系，同时基于词的类别区分能力提出了特定领域中褒义词和贬义词种子集的选取方法。杜伟夫等[23]将词语情感倾向计算问题归结为优化问题，首先利用HowNet相似度和PMI值构建情感词间的无向图，然后利用以“最小切分”为目标的目标函数对该图进行划分，并使用模拟退火算法进行求解。

3 算法描述

本文提出了一个两阶段的领域情感词典构造方法。第一阶段，使用情感词间的PMI统计值和上下文约束关系建立情感词间的相似性矩阵，然后利用基于约束的标签传播算法在情感词褒贬义种子集上不断迭代来构造基本情感词典，为每一个情感词分配固定的情感倾向。第二阶段，根据情感词出现情感冲突的频率来识别领域相关情感词，并根据其在语料中的上下文信息对修饰的不同产品特征分配不同的情感倾向。

3.1 领域情感词典

领域情感词典由一系列特定领域中的情感词及其情感倾向构成，我们将领域情感词典的每一个元素定义为一个四元组(D,W,F,P)。其中，D表示情感词典的适用领域；W表示情感词；F表示情感词修饰的产品特征，如果情感词在特定领域中表示相同的情感，则F表示为“ALL”；P表示情感词的情感倾向(褒义为1，贬义为-1)。

3.2 产品特征及其情感词的获取

为了构建领域情感词典，需要抽取产品评论中所包含的产品特征及其对应的情感词。产品特征及其情感的抽取是产品评论挖掘的基本工作之一，很多学者已经提出了各种算法来完成这项工作[1,24-26]。本文利用双向传播算法[26]完成产品特征及其情感词的抽取工作。双向传播算法利用情感词和产品特征之间、情感词之间、产品特征之间的句法依存关系模式抽取产品特征和情感词，不需要标注大量的训练数据，只需要一部分情感词种子，利用特定的句法依存关系模式不断迭代来获取新的产品特征和情感词，并对抽取的产品特征和情感词进行排序以提高准确率。双向传播算法定义了四类规则来抽取产品特征和情感词(表1)。使用规则$R1_i$利用情感词抽取情感词，使用规则$R2_i$利用情感词抽取产品特征，使用规则$R3_i$利用产品特征抽取产品特征，使用规则$R4_i$利用产品特征抽取情感词。

表1中第二列是产品特征和观点之间的句法依存关系模式，第三列是抽取规则的限制条件，最后一列是结果。箭头代表着句法依存关系，例如，“$S\rightarrow S\text{-}Dep\rightarrow F$”表示$S$通过依存关系$S\text{-}Dep$依存于$F$。

表 1　产品特征和情感词的抽取规则

	句法依存关系	限　　制	输　　出
$R1_1$	$S_{i(j)} \rightarrow S_{i(j)}\text{-}Dep \rightarrow S_{j(i)}$	$S_{j(i)} \in \{S\}$， $S_{i(j)}\text{-}Dep \in \{CONJ\}$， $POS(S_{j(i)}) \in \{JJ\}$	$s = S_{i(j)}$
$R1_2$	$S_i \rightarrow S_i\text{-}Dep \rightarrow H \leftarrow S_j\text{-}Dep \leftarrow S_j$	$S_i \in \{S\}$， $S_i\text{-}Dep == S_j\text{-}Dep$， $POS(S_j) \in \{JJ\}$	$s = S_j$
$R2_1$	$S \rightarrow S\text{-}Dep \rightarrow F$ $F \rightarrow F\text{-}Dep \rightarrow S$	$S \in \{S\}$， $S\text{-}Dep \in \{MR\}$，$F\text{-}Dep \in \{MR\}$， $POS(F) \in \{NN, VV\}$	$f = F$
$R2_2$	$S \rightarrow S\text{-}Dep \rightarrow H \leftarrow F\text{-}Dep \leftarrow F$	$S \in \{S\}$， $S\text{-}Dep \in \{MR\}$，$F\text{-}Dep \in \{MR\}$， $POS(F) \in \{NN\}$	$f = F$
$R2_3$	$S \rightarrow S\text{-}Dep \rightarrow H \rightarrow F\text{-}Dep \rightarrow F$ $S \leftarrow S\text{-}Dep \leftarrow H \leftarrow F\text{-}Dep \leftarrow F$	$S \in \{S\}$， $S\text{-}Dep \in \{MR\}$，$F\text{-}Dep \in \{MR\}$， $POS(F) \in \{NN\}$	$f = F$
$R3_1$	$F_{i(j)} \rightarrow F_{(j)}\text{-}Dep \rightarrow F_{j(i)}$	$F_{j(i)} \in \{F\}$， $F_{i(j)}\text{-}Dep \in \{CONJ\}$， $POS(F_{j(i)}) \in \{NN, VV\}$	$f = F_{i(j)}$
$R3_2$	$F_i \rightarrow F_i\text{-}Dep \rightarrow H \leftarrow F_j\text{-}Dep \leftarrow F_j$	$F_i \in \{S\}$， $F_i\text{-}Dep == F_j\text{-}Dep$， $POS(F_j) \in \{NN, VV\}$	$f = F_j$
$R4_1$	$S \rightarrow S\text{-}Dep \rightarrow F$ $F \rightarrow F\text{-}Dep \rightarrow S$	$F \in \{F\}$， $S\text{-}Dep \in \{MR\}$，$F\text{-}Dep \in \{MR\}$， $POS(S) \in \{JJ\}$	$s = S$
$R4_2$	$S \rightarrow S\text{-}Dep \rightarrow H \leftarrow F\text{-}Dep \leftarrow F$	$F \in \{F\}$， $S\text{-}Dep \in \{MR\}$，$F\text{-}Dep \in \{MR\}$， $POS(S) \in \{JJ\}$	$s = S$
$R4_3$	$S \rightarrow S\text{-}Dep \rightarrow H \rightarrow F\text{-}Dep \rightarrow F$ $S \leftarrow S\text{-}Dep \leftarrow H \leftarrow F\text{-}Dep \leftarrow F$	$F \in \{F\}$， $S\text{-}Dep \in \{MR\}$，$F\text{-}Dep \in \{MR\}$， $POS(S) \in \{JJ\}$	$s = S$

表中，$s(f)$表示抽取的观点(产品特征)，$\{S\}$($\{F\}$)和 $S\text{-}Dep(F\text{-}Dep)$表示已获取的观点(产品特征)以及其句法依存关系，H 表示任意单词。$POS(S)$($POS(F)$)是 $S(F)$的词性信息。$\{JJ\}$和$\{NN\}$、$\{NN, VV\}$是观点和产品特征应满足的词性集。本文抽取形容词作为观点，名词和动词作为产品特征。$\{MR\}$代表产品特征和观点间可能存在的依存关系，例如，SBV，VOB，ATT 等。$\{CONJ\}$表示并列连词依存关系。

3.3　产品评论中情感词的上下文约束

情感词的上下文约束是指情感词和其上下文的情感词间，由于存在并列、转折等关系，从而保持相同或相反的情感倾向。一些学者已经将这些关系运用到情感分析中[8,18-21]。本文提取了以下四种情感词间的上下文约束。

(1) 并列关系

具有并列关系的两个情感词一般具有相同的情感倾向。例如，“外观美丽、大方”。

(2) 转折关系

具有转折关系的两个情感词一般具有相反的情感倾向。例如，“屏幕分辨率虽然比较低，但是显示效果不错。”

(3) 语句内情感关系

产品评论中，经常在同一句话中出现对多个产品特征的评价，这些评价的情感词往往具有相同的情感倾向。例如，“外观大方，屏幕分辨率很高，价格实惠。”

(4) 语句间情感关系

产品评论中，人们经常在相邻的句子中表达相

同的情感倾向。例如,“屏幕分辨率高,色彩鲜艳。电池续航时间长。”

3.4 基本情感词典的构造

3.4.1 情感词联系图

我们定义了一个无向图 $G=(X,A)$来表示抽取的所有情感词间的联系,其中 $X=(x_1,x_2,\cdots,x_{|X|})$表示情感词的集合,$|X|$表示情感词集合的大小,$A$ 表示情感词间的相似性矩阵。我们假设具有较大相似度的两个情感词更有可能具有相同的情感倾向,图中一个词的情感倾向由其相邻词的情感倾向决定。因此,情感词的相似性是在情感词褒贬义种子集的基础上进行情感倾向传播的关键。本文使用情感词的点互信息(PMI)来计算情感词间的相似性,如式(1)所示。

$$\mathrm{PMI}(x_i,x_j)=\log_2\frac{p(x_i,x_j)}{p(x_i)p(x_j)} \tag{1}$$

其中,$p(x_i,x_j)$表示情感词x_i和x_j在语料中特定长度词的窗口中同时出现的概率,$p(x_i)$表示情感词x_i在语料中出现的概率,$p(x_j)$表示情感词x_j在语料中出现的概率。

3.4.2 约束传播

PMI 利用了两个情感词间的共现统计信息,但是没有考虑两个情感词间的上下文语义约束信息(例如,并列、转折关系等)。为了利用情感词间的上下文语义约束,我们提取了四种约束:并列关系、转折关系、语句内情感关系、语句间情感关系。我们将一般具有相同情感倾向的并列关系、语句内情感关系、语句间情感关系定义为正向约束关系,一般具有相反情感倾向的转折关系定义为反向约束关系。

直观上,我们可以定义情感词间约束矩阵 $D=\{d_{ij}\}_{|X|\times|X|}$来表示情感词间的正向和反向约束关系,$d_{ij}$表示情感词$x_i$和$x_j$间的约束关系,$|d_{ij}|$则表示约束关系的置信度,如式(2)所示。

$$d_{ij}=\begin{cases}1, & fs(x_i,x_j)>t_1\\ -1, & fr(x_i,x_j)>t_2\\ 0, & otherwise\end{cases} \tag{2}$$

其中,$fs(x_i,x_j)$表示情感词x_i和x_j间出正向约束关系的次数,$fs(x_i,x_j)$表示情感词x_i和x_j间出现反向约束关系的次数,t_1和t_2表示相应的阈值。

但这些约束关系只能影响与其相关的局部情感词,而不能扩展到整个情感词集[21]。我们将抽取的上下文约束进一步传播,作为先验知识以修正情感词间的相似性矩阵 A,其算法如下:

(1) 基于相似矩阵 A 构造权重矩阵 W 如式(3)所示。

$$W_{ij}=\begin{cases}\dfrac{A(x_i,x_j)}{\sqrt{A(x_i,*)}\sqrt{A(x_j,*)}}, & j\neq i\ and\ A(x_i,x_j)\geqslant 0\\ 0, & otherwise\end{cases} \tag{3}$$

其中,$A(x_i,x_j)$表示情感词x_i和x_j在相似性矩阵 A 中的值,$A(x_i,*)$表示情感词x_i与所有邻居的相似性的和。

(2) 构造矩阵 $S=Z^{-1/2}WZ^{-1/2}$,Z 是对角矩阵,其第 i 行 i 列的值等于 W 第 i 行值的和。

(3) 通过式(4)进行垂直传播,直到收敛。

$$E_v(t+1)=\alpha S E_v(t)+(1-\alpha)D \tag{4}$$

其中,$E_v(0)=D$,α 取值为(0,1)。

(4) 通过式(5)进行水平传播,直到收敛。

$$E_h(t+1)=\alpha E_h(t)S+(1-\alpha)E_v^* \tag{5}$$

其中,$E_h(0)=D$,E_v^* 是步骤(4)的结果。

(5) 输出 $E^*=E_h^*$,表示情感词间约束信息传播的最终结果。

3.4.3 基于约束的标签传播

标签传播算法是一个优秀的基于图的半监督学习算法,具有很好的效率和收敛性[27]。本文结合经过约束传播修正的情感词相似性矩阵 A 和标签传播算法来计算情感词的情感倾向,构造基本情感词典。

定义 $L=\{(x_1,y_1),(x_2,y_2),\cdots,(x_l,y_l)\}$表示已知倾向的情感词种子集,$y_l$表示情感词$x_l$的情感倾向,$y_i\in Y\{y_{pos},y_{neg}\}$,$y_{pos}$表示褒义,$y_{neg}$表示贬义;$U=\{x_{l+1},x_{l+2},\cdots,x_{l+u}\}$表示未知情感倾向的情感词集。因此,$X=L\cup U=\{x_1,x_2,\cdots,x_l,x_{l+1},\cdots,x_{l+u}\}$,$|X|=l+u$。假设,$l\ll|X|$,$Y_L=\{y_1,y_2,\cdots,y_l\}$表示情感词种子集 L 的已知情感倾向,$Y_U=\{y_{l+1},y_{l+2},\cdots,y_{l+u}\}$代表未知情感倾向的情感词集 U 的情感倾向。要解决的问题可以归纳为在 U 和Y_L的基础上估计Y_U。

结合约束传播的结果,对相似性矩阵 A 进行式(6)修正。

$$\widetilde{A}_{ij}=\begin{cases}1-(1-E_{ij}^*)(1-A_{ij}), & E_{ij}^*\geqslant 0\\ (1+E_{ij}^*)A_{ij}, & E_{ij}^*<0\end{cases} \tag{6}$$

$\widetilde{A}$ 由相似性矩阵 A 和约束传播结果E^*导出,表示结合约束信息的情感词相似矩阵。当情感词x_i和x_j间存在正向约束时($E_{ij}^*>0$),增大其相似性A_{ij}

的值。相反($E_{ij}^{*}<0$),则减小其相似性A_{ij}的值。

定义情感词间的迁移概率矩阵$T_{|X|\times|X|}$,迁移概率T_{ij}的定义如式(7)所示。

$$T_{ij}=p(i\rightarrow j)=\frac{\widetilde{A}_{ij}}{\sum_{j=1}^{|X|}\widetilde{A}_{ij}} \tag{7}$$

其中,$p(i\rightarrow j)$表示情感词的情感倾向从x_i传递给x_j的概率。

$f^t=\{f_1^t,f_2^t,\cdots,f_{|X|}^t\}$表示第$i$次迭代的情感倾向向量,$f^0$中褒义情感词种子的值为1,贬义情感词种子的值为-1,未知倾向情感词的值为0。基于约束的标签传播算法如下:

(1) 按如下公式更新情感倾向向量f^t的值,每个情感词的情感倾向都受其相邻情感词情感倾向的影响如式(8)所示。

$$f^{t+1}=Tf^t \tag{8}$$

(2) 将情感词种子集对应的向量元素值复原如式(9)所示。

$$f_l^{t+1}=f_l^0 \tag{9}$$

(3) 重复上述过程直到收敛。

收敛后,可以得到情感倾向向量f^t。如果情感词对应的向量元素的值大于0,则认为其情感倾向是褒义的。如果情感词对应的向量元素的值小于0,则认为其情感倾向是贬义的。

3.5　领域相关情感词的识别

领域相关情感词是指在不同领域中具有不同的情感倾向,甚至在同一领域中当修饰不同产品特征时也具有不同的情感倾向。例如,“大”在修饰“手机屏幕”时表示贬义,而在修饰“通话噪音”时表示贬义。因此,建立领域情感词典必须识别这些领域相关的情感词及其情感倾向。领域相关情感词在领域语料中表达不同的情感倾向,我们利用情感词的上下文约束关系来发现情感词的情感冲突的现象。例如,特定的情感词如果在一些上下文环境中被推测为褒义,而在另外一些上下文环境中被推测为贬义,则该情感词存在情感冲突。本文利用情感词出现情感冲突的频率来识别领域相关情感词,其算法如下。其中,$cf(x_i)$表示情感词x_i在语料中出现冲突的频率,$of(x_i)$表示情感词x_i修饰的产品特征的数量。

(1) 获取所有特征情感词实例集合FO;

(2) 遍历FO中的特征情感词实例对(fo_i,fo_{i+1});

(3) 如果fo_i和fo_{i+1}的情感词间不存在正向和反向约束关系,跳转到(2);

(4) 如果fo_i和fo_{i+1}的情感词x_m和x_n间存在正向约束关系,而且在基本情感词典中的情感倾向不一致,则将x_m和x_n加入候选领域情感词集中,并且$cf(x_m)$加1,$cf(x_n)$加1,跳转到(2);

(5) 如果fo_i和fo_{i+1}的情感词x_m和x_n间存在反向约束关系,而且在基本情感词典中的情感倾向一致,则将x_m和x_n加入候选领域情感词集中,并且$cf(x_m)$加1,$cf(x_n)$加1,跳转到(2);

(6) 遍历候选领域情感词集,如果$cf(x_i)>\alpha$,并且$of(x_i)>\beta$,则将x_i加入到领域相关情感词集DS中。

获取领域相关情感词集DS后,可以根据这些情感词修饰的产品特征进一步修正基本情感词典,从而得到领域相关情感词典,其算法如下。其中,集合OFS是四元组(W,F,Pos,Neg)的集合,W表示情感词,F表示情感词修饰的产品特征,Pos表示褒义倾向的计数,Neg表示贬义倾向的计数。

(1) 获取所有特征情感词实例集合FO;

(2) 遍历FO中的特征情感词实例fo_i;

(3) 如果fo_i的情感词$x_m\in DS$,寻找fo_i的前后实例fo_{i-1}和fo_{i+1};

(4) 将fo_{i-1}和fo_{i+1}中优先级较大的赋予fo',优先级顺序为并列关系、转折关系、语句内情感关系、语句间情感关系;

(5) 如果fo'中情感词为褒义,则在OFS中寻找与fo'的产品特征和情感词对应的四元组(W,F,Pos,Neg),并将Pos的值加1,跳转到(2);

(6) 如果fo'中情感词为贬义,则在OFS中寻找与fo'的产品特征和情感词对应的四元组(W,F,Pos,Neg),并将Neg的值加1,跳转到(2);

(7) 遍历集合OFS,根据Pos和Neg中较大的值来决定其情感倾向并加入到领域情感词典中。

4　结果分析

4.1　实验数据

本文的产品评论数据都取自一些电子商务网站以及评论网站。网络上存在着大量的电子商务网站以及评论网站,经过分析我们选择了亚马逊、京东商城、中关村在线、it168这四个典型的网站作为我们评论数据的来源。电子产品是网络上评论数量最多的一类产品,本文选择了以上网站中的手机、数码相

机这两种典型电子产品的评论来构造实验用的评论数据集。表2给出了数据集中评论和句子的数量。

表2 实验数据集

产 品	评论数量	句子数量
手机	800	5 108
数码相机	800	5 215

4.2 产品特征和情感词抽取及约束关系分析

产品特征及其对应情感词的抽取是构建领域情感词典的基础工作。本文利用双向传播算法同时抽取产品特征及其情感词,抽取结果的准确率和召回率见表3。

表3 产品特征及其情感词的抽取结果

类 别	手 机		数码相机			
	P	R	F1	P	R	F1
产品特征抽取	0.753	0.692	0.721	0.716	0.652	0.683
情感词抽取	0.737	0.788	0.762	0.721	0.781	0.750

情感词在上下文中的并列关系、转折关系、语句内情感关系、语句间情感关系是我们工作的重要基础。我们在这些关系的基础上利用基于约束的标签传播算法建立基本情感词典,并进一步利用这些关系完善了领域情感词典。这四种情感词上下文约束关系在语料中所占的比例以及置信度见表4。

表4 语料中不同上下文约束关系所占的比例及置信度

类 别	并列关系	转折关系	语句内情感关系	语句间情感关系
所占比例/%	12.6	3.7	73.4	10.3
置信度/%	99.1	96.2	89.3	87.6

4.3 实验结果

本文提出了一个两阶段的领域相关情感词典构造方法。为了验证该方法的有效性,我们在表5中对比了几种不同算法的结果。其中,HowNet代表文献[12]中基于HowNet语义相似度的方法;Cilin代表文献[13]中基于同义词词林的方法;PMI代表在当前语料库中基于PMI的方法;ChConsLP代表针对中文语料改进后的文献[21]中方法,但只使用了并列和转折两种关系;ImChConsLP代表本文提出的方法。为了便于比较,表5的结果都是基于领域内选择的十对褒贬义种子集。

表5 实验结果

方 法	手 机		数码相机			
	P	R	F	P	R	F
HowNet	0.836	0.987	0.905	0.863	0.986	0.920
Cilin	0.911	0.991	0.949	0.913	0.992	0.951
PMI	0.921	0.905	0.913	0.923	0.905	0.914
ChConsLP	0.937	0.973	0.955	0.941	0.967	0.954
ImChConsLP	0.963	0.985	0.974	0.961	0.981	0.973

从表5的结果中可以看出,本文提出的方法在两个领域中都取得了最好的F-measure值。HowNet和《同义词词林》都是手工编制的词典,包含了大量词汇,因此HowNet和Cilin这两种方法的召回率都比较高。但这两种方法都没有考虑领域情感词的情感倾向,因此准确率较低。PMI利用语料上的点互信息统计值来计算情感词的情感倾向,相对于HowNet和Cilin来说具有更高的准确率。但对于一些语料中出现频率较少的情感词存在数据稀疏的问题,因此召回率较低。ChConsLP和ImChConsLP相对于PMI在准确率和召回率上都取得了更好的效果,证明了情感词的上下文约束关系和标签传播算法的有效性。同时,ImChConsLP比ChConsLP在准确率和召回率上都有所提高,证明了本文提出方法的有效性。

本文使用了并列关系、转折关系、语句内情感关系、语句间情感关系来建立基本情感词典以及修正领域情感词典。表6对比了使用不同上下文约束关系以及修正领域情感词典的效果。其中,ChConsLP使用了并列关系和转折关系,ChConsLP1在ChConsLP的基础上增加了语句内情感关系和语句间情感关系,ImChConsLP在ChConsLP1的基础上利用四种上下文约束关系以及情感词修饰的特征进行了领域情感词典的修正。

表6 使用不同上下文约束关系及修正领域情感词典的结果

方 法	手 机		数码相机			
	P	R	F	P	R	F
ChConsLP	0.937	0.973	0.955	0.941	0.967	0.954
ChConsLP1	0.953	0.985	0.969	0.957	0.981	0.969
ImChConsLP	0.963	0.985	0.974	0.961	0.981	0.973

从表6的结果中可以看出,加入了语句内情感关系和语句间情感关系后准确率和召回率都有所提

高，证明了语句内情感关系和语句间情感关系能有效提高情感倾向计算的效果。ImChConsLP 利用四种上下文约束关系计算情感冲突频率来识别领域相关情感词，并利用情感词在语料中的上下文信息对其修饰的不同特征分配不同的情感倾向，从而进一步提高了准确率。但由于修饰不同产品特征时具有不同情感倾向的情感词在整个语料中所占比例较小，因此准确率的改善较小。

表 7 对比了褒贬义种子数量对实验结果的影响。从结果中可以看出：(1)随着种子数量的增长，准确率和召回率都有所提高，但对召回率的影响较小。(2)当种子由五对变成十对时，在手机语料中准确率提高了 1.7%，在数码相机语料中准确率提高了 1.2%。当再增加更多种子时，准确率提高并不显著。因此，本文提出的算法使用较小的种子集就可以得到较好的效果。

表 7　褒贬义种子数量对实验结果的影响

种子集	手机		数码相机			
	P	R	F	P	R	F
5 对褒贬义种子	0.946	0.985	0.965	0.949	0.980	0.964
10 对褒贬义种子	0.963	0.985	0.974	0.961	0.981	0.973
15 对褒贬义种子	0.967	0.985	0.976	0.968	0.981	0.974
20 对褒贬义种子	0.968	0.986	0.977	0.970	0.982	0.976
30 对褒贬义种子	0.970	0.987	0.978	0.972	0.982	0.977

5　结论和进一步的工作

情感词典是进行情感分析的基础。但有些情感词在不同领域中具有不同的情感倾向，甚至在同一领域中修饰不同产品特征时也具有不同的情感倾向。因此，建立领域情感词典能更好地识别情感词的情感倾向。本文提出了一个两阶段的领域情感词典构建方法，并在手机和数码相机两种电子产品评论语料集上验证了该方法的有效性。同时，使用较小的种子集就可以取得理想的准确率和召回率。本文只判别了情感词的情感倾向，如何判断情感倾向的强度将是今后工作的一个重要问题。

参考文献

[1] M HU, B LIU. Mining and summarizing customer reviews[C]//Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2004: 168-177.

[2] A M Popescu, O Etzioni. Extracting product features and opinions from review[C]//Proceedings of the Human Language Technology Conference and the Conference on Empirical Methods in Natural Language Processing, 2005: 339-346.

[3] P Stone, D Dunphy, M Smith, et al. The General Inquirer: A Computer Approach to Content Analysis [M]. Cambridge: MIT Press, 1966.

[4] S Baccianella, A Esuli, F Sebastian. SENTIWORDNET3.0: An Enhanced Lexical Resource for Sentiment Analysis and Opinion Mining[C]//Proceedings of the Seventh Conference on International Language Resources and Evaluation, 2010: 2200-2204.

[5] 知网，董振东：http://www.keenage.com/[EB/OL].

[6] L W Ku, H H Chen. Mining Opinions from the Web: Beyond Relevance Retrieval[J]. Journal of the American Society for Information Science and Technology. 2007, 58(12): 1838-1850.

[7] 徐琳宏，林鸿飞，潘宇，等. 情感词汇本体的构造[J]. 情报学报, 2008, 27(2): 180-185.

[8] S Huang, Z Niu, C Shi. Automatic Construction of Domain-specific Sentiment Lexicon Based on Constrained Label Propagation[J]. Knowledge-Based Systems, 2013, 56: 191-200.

[9] M HU, B LIU. Mining Opinion Features in Customer Reviews[C]//Proceedings of 9th National Conference on Artificial Intelligence, 2004: 755-760.

[10] J Kamps, M Marx, R J Mokken, et al. Using Wordnet to Measure Semantic Orientations of Adjectives [C]//Proceedings of the 4th International Conference on International Language Resources and Evaluation, 2004: 1115-1118.

[11] D Rao, D Ravichandran. Semi-supervised Polarity Lexicon Induction[C]//Proceedings of the 12th Conference of the European Association of Computational Linguistics, 2009: 675-682.

[12] A Esuli, F Sebastiani. Sentiwordnet: A Publicly Available Lexical Resource for Opinion Mining[C]//Proceedings of the 5th Conference on Language Resources and Evaluation, 2006: 417-422.

[13] 朱嫣岚，闵锦，周雅倩，等. 基于 HowNet 的词汇语义倾向计算[J]. 中文信息学报, 2006, 20(1): 14-20.

[14] 路斌，万小军，杨建武，等. 基于同义词词林的词汇褒贬计算[C]. 第七届中文信息处理国际会议论文集. 武汉，中国：电子工业出版社, 2007: 17-23.

[15] 周咏梅，杨佳能，阳爱民. 面向文本情感分析的中文情感词典构建方法[J]. 山东大学学报(工学版), 2013, 43(6): 27-33.

[16] P D Turney. Thumbs Up or Thumbs Down?: Se-

mantic Orientation Applied to Unsupervised Classification of Reviews[C]//Proceedings of the 40th Annual Meeting on Association for Computational Linguistics, 2002: 417-424.

[17] P D Turney, M L Littman. Measuring Praise and Criticism: Inference of Semantic Orientation from Association[J]. ACM Transaction on Information System, 2003, 21(4): 315-346.

[18] V Hatzivassiloglou, K R McKeown. Predicting the Semantic Orientation of Adjectives[C]//Proceedings of the 8th Conference on European Chapter of the Association for Computational Ling, 1997: 174-181.

[19] H Kanayama, T Nasukawa. Fully Automatic Lexicon Expansion for Domain-Oriented Sentiment Analysis [C]//Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing, 2006: 355-363.

[20] X Ding, B Liu. The Utility of Linguistic Rules in Opinion Mining[C]//Proceedings of the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, 2007: 811-812.

[21] R Y K Lau, C L Lai, P D Bruza, et al. Leveraging Web 2.0 Data for Scalable Semi-supervised Learning of Domain-specific Sentiment Lexicons[C]//Proceedings of the 20th ACM International Conference on Information and Knowledge Management, 2011: 2457-2460.

[22] 王素格，李德玉，魏英杰，等. 基于同义词的词汇情感倾向判别方法[J]. 中文信息学报，2009,23(5): 68-74.

[23] 杜伟夫,谭松波,云晓春,等. 一种新的情感词汇语义倾向计算方法[J]. 计算机研究与发展，2009，46(10): 1713-1720.

[24] G Qiu, B Liu, J Bu et al. Expanding domain sentiment lexicon through double propagation[C]//Proceedings of the 21st International Joint Conference on Artificial Intelligence,2009: 1199-1204.

[25] L Zhang, B Liu, S H Lim, et al. Extracting and ranking product features in opinion documents [C]//Proceedings of the 23rd International Conference on Computational Linguistics,2010: 1462-1470.

[26] Y Xi. 产品评论特征及观点抽取研究[J]. 情报学报，2014,33(3): 326-336.

[27] F Wang, C Zhang. Label Propagation through Linear Neighborhoods[C]//Proceedings of the 23rd International Conference on Machine Learning, 2006: 985-992.

郁亚辉(1977－)，副教授，主要研究领域为文本挖掘、信息检索。
E-mail:xiyahui@hbu.edu.cn