情报学报 ISSN 1000-0135
第35卷第5期501-509，2016年5月

JOURNAL OF THE CHINA SOCIETY FOR SCIENTIFIC AND TECHNICAL INFORMATION ISSN 1000-0135
Vol. 35 No. 5, 501-509 May 2016

doi:10.3772/j.issn.1000-0135.2016.005.006

基于关联规则挖掘和极性分析的商品评论情感词典构建[1)]

钟敏娟[1,2] 万常选[1,2] 刘德喜[1,2]

（1. 江西财经大学信息管理学院，南昌 330013；
2. 江西财经大学数据与知识工程江西省高校重点实验室，南昌 330013）

摘要 作为情感倾向性分析的基础性工作，情感词典构建包括情感词的识别与极性判断两大任务。本文以亚马逊网站上的音乐商品评论信息作为数据源，力图构建该领域的情感词典。首先利用关联规则挖掘算法充分挖掘领域主题词和情感词之间的关系，获取体现领域特征的情感词；然后针对每个情感词，引入词项间的混合相关关系，结合PageRank模型构建情感词的量化图模型，获得每个情感词的极性。实验结果表明，本文所提方法能有效地构建音乐领域情感词典，不仅能够识别该领域特征的情感词，同时还能较为准确地判断该情感词的情感原极性。

关键词 情感倾向性 情感词典 关联规则 PageRank 混合相关关系

Opinion Lexicon Construction Based on Association Rule and Orientation Analysis for Production Review

Zhong Minjuan[1,2], Wan Changxuan[1,2] and Liu Dexi[1]

(1. *School of Information Technology, Jiangxi University of Finance and Economics, Nanchang* 330013;
2. *Jiangxi Key Laboratory of Data and Knowledge Engineering, Jiangxi University of Finance and Economics, Nanchang* 330013)

Abstract As fundamental work in sentiment analysis, opinion lexicon construction consists of two task, opinion word identification and orientation computing. This paper tries to build opinion lexicon based on music production review from amazon.com. Firstly, association rule mining algorithm is performed to mine the relation between field key word and opinion word, and opinion words are obtained. After that, the quantification model of opinion word is built to get original polarity, in which PageRank model combined with mixture relevance relation are introduced. Experimental Results show that the proposed method could construct opinion lexicon on music field effectively. It not only obtains the opinion word of field characteristics, but also provides more accurate judgment on their original opinion polarity.

Keywords sentiment orientation, opinion lexicon, association rule, PageRank, mixture relevance relation

1 引 言

随着Web2.0的推广，各种新兴网络媒介（例如博客、微博、论坛等）成为网民发表自己对某些事物的观点、态度及看法的重要场所，形成了巨大的倾向性文本信息，如此大量的情感倾向信息，不论对于普通的网络用户，还是对于互联网公司都有很重要的价值。如何自动地从海量数据中抽取出人们对某一事物的观点和看法，是文本倾向性分析所要解决的

收稿日期：2015年11月6日

作者简介：**钟敏娟**，女，1976年生，博士，副教授，主要研究方向：信息检索，情感分析，数据挖掘，E-mail：lucyzmj@sina.com。**万常选**，男，1962年生，教授，博导，主要研究方向：情感分析，Web信息管理，数据库技术。**刘德喜**，男，1975年生，博士，教授，主要研究方向：情感分析，自动文摘。

1） 基金项目：国家社会科学基金项目（12CTQ042），国家自然科学基金项目（61363039，61562032）江西省高等学校科技落地计划项目（KJLD14035）。

问题[1]。而词汇的倾向性(极性)识别任务作为倾向性分析系统中的基础性工作,更得到了极大关注。

情感词又称为极性词、观点词、评价词语,是表达情感的最小文本单元。情感词的识别和极性分析在情感分析领域创建伊始就引起了人们极大的兴致,进行了大量的研究。纵观目前的研究,情感词的识别和极性分析往往是一个一体化的工作,主要围绕着情感词典构建来进行,方法有两类:基于语料库的方法和基于词典的方法。

(1)基于语料库的方法。主要是利用大语料库的统计特性,观察一些现象来挖掘语料库中的评价词语并判断极性。Hatzivassiloglou 等[2]最早提出了利用句法连接(并列、转折、递进)来识别情感词并判断其极性,认为并列或递进关系连接的两个形容词具有相同的情感极性,而转折关系连接的形容词具有相反的极性。该方法具有一定的局限性,只能处理有限的由连词关联的形容词性的情感词。Kanayama 等[3]对此进行了扩展,认为相同的情感极性往往会在前后相连的句子中表达,而极性的改变则通过转折词 but 或 however 来体现,并基于句间与句内情感一致性的思想来推断情感词在某领域内的极性。基于大规模语料库,Turney 等[4]研究了词的语义倾向,提出了具有较大影响的点互信息(Pointwise Mutual Information,PMI)方法来判断待测词的情感极性。Ding 等[5]提出了一种新的情感词极性判别方法,将情感词与上下文特征联系起来,并据此决定该词的情感极性。Lu 等[6]提出了一个优化框架来学习特征依赖的情感。Mohtarami 等[7]提出了词的情感空间模型从而推理出词的情感相似度,该方法利用概率的方法建立隐藏情感模型,以情感向量构成词的情感相似度,比基于词的总体情感相似度的准确率更高。

(2)基于词典的方法。主要是使用词典中词语之间的词义联系来挖掘评价词。这里的词典一般是指使用 WordNet 或 HowNet 等。Kamps 等[8]最早提出了基于词的关联关系方法,沿用 Turney 等[4]的 PMI 方法,通过计算 WordNet 中所有形容词与种子褒义词代表 good 和贬义词代表 bad 之间的关联度来判断情感词的情感倾向。由于词语的多义性,该方法的极性判断精度不高,为此 Esuli 等[9,10]使用词典中词语的注释信息来判断情感词的极性,取得了比以往研究更好的准确率。

无论是基于语料库的方法,还是基于词典的方法,其核心思想都是通过词与词之间的联系来分析情感词的情感倾向性。有学者提出利用情感词和主题词之间的关系来识别情感词。Hu 等[11]首先在商品评论中利用抽取的商品特征来辅助情感词的抽取与识别,该方法采用关联规则挖掘找到频繁出现的商品特征,然后从那些包含一个或多个商品特征的句子中提取形容词作为情感词。然而,并不是所有的频繁名词都是特征,因此,这种基于频繁名词和规则的方法会产生很多非特征,根据"特征词-情感词"之间的关联规则,从而造成情感词识别精度降低。Qiu 等[12]提出了一种 Double Propagation 方法同时进行情感词和主题词的识别与抽取,在定义一系列种子情感词的基础上,定义了主题词-情感词、主题词-主题词、情感词-情感词之间的语法规则关系,通过不断迭代将情感词抽取与识别出来,并基于此构建领域相关词典。还有部分学者采用基于图的方法来识别情感词的极性。Takamura 等[13]借鉴物理学中的 Spin 模型来判断词汇的极性。Rao 等[14]也采用基于图的方法来识别情感词的极性,将要分类的词语作为图上的点,利用词语之间的联系形成边来构建图,继而采用各种基于图的迭代算法来完成词语的分类。

综上所述,基于词典的方法获取的情感词规模非常可观,适用于通用的情感词典,但是无法产生与领域以及上下文相关的情感词。在此,本文以亚马逊网站上的音乐商品评论信息作为数据源,构建音乐领域的情感词典。事实证明,它可以较好地描述音乐领域中情感词的语义倾向性。具体来说,本文首先提出了基于关联规则挖掘的情感词抽取与识别,不同于文献[11],本文首先对商品特征进行了语义预处理,不仅较为准确地识别出了特征词(主题词),而且使得识别出的特征词具有领域相关性。然后,结合 PageRank 模型,引入词汇间的混合相关关系来研究词汇的原始情感倾向性(即在通常情况下,情感词表现出的情感极性)。实验结果表明,本文所提方法能有效地构建音乐领域情感词典,不仅能够获得该领域特征的情感词,同时还能较为准确地判断该情感词的情感原极性。

2 基于关联规则挖掘的情感词抽取与识别

抽取领域情感词和修饰对象是文本情感倾向性分析中的一项基础且重要的工作。通常,人在表达各种情感倾向时都是针对某个修饰对象(也称作评

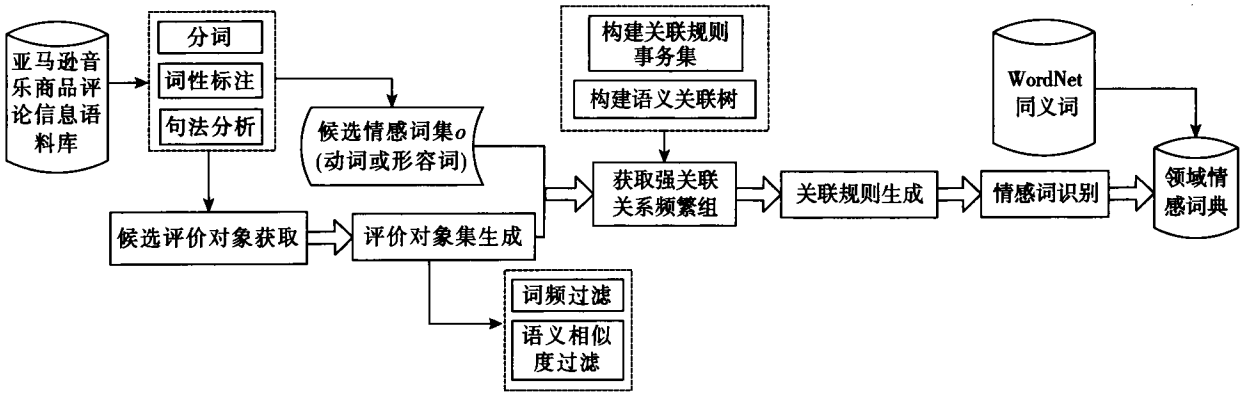

图 1　基于关联规则抽取评价对象和情感词的技术路线图

价对象),该修饰对象一般是表征领域的主题特征词,因此,充分挖掘评价对象和情感词之间的关系,能够获取体现领域特征的情感词。基于此考虑,本文具体采用以下技术路线,如图 1 所示。

2.1　评价对象的识别与抽取

在阅读产品评论时,用户往往表现出更加关注评价对象,然而,评论中究竟哪些词语属

于产品评价对象很难给出统一的定义。文献[15,16]中认为,产品评价对象以三种形式出现:产品的整体;产品的某个部件;产品的特性及其外延,它们分别对应产品名称和产品属性。通过对大量真实产品评论文本的观察,我们发现用户对产品名称及其属性的表达往往以名词或名词短语的形式体现。为此,以亚马逊音乐商品评论信息语料库作为基准库,我们采取如下的思路将评价对象提取出来。首先通过预处理方式(分词、词性标注、依存句法分析等)提取其中的名词或名词短语作为候选的评价对象集 T。在此基础上,为了进一步提高评价对象的识别精度以及体现领域相关性,采用词频过滤和语义过滤的两阶段过滤机制对候选评价对象集 T 进行筛选,从而最终将评价对象识别与抽取出来。下面分别对这两项过滤技术进行介绍:

(1)词频过滤

统计语料库中名词和名词短语的出现频度,将出现次数比较少的名词和名词短语过滤掉。

词频过滤主要基于以下考虑:①在基于词项权重的向量空间模型统计方法中,词项的出现次数意味着该词项对文档的贡献程度,与此相类似,语料库中反复出现的评价对象往往也说明了该评价对象属于重要特征,是用户更加关注的方面,相反,那些次要的或者是冷门的特征不被大众所关心,因此较少出现在用户撰写的评论中,对于这些属性特征我们可以忽略。②虽然依据词频信息可能会过滤掉一些低频特征,造成后续情感词识别的查全率性能略微降低,但是相比查全率而言,我们更看重情感词的准确性,而这必须建立在评价对象的正确识别基础上,通过词频的约束条件可以在一定程度上提高评价对象的识别准确率。

(2)PMI 语义过滤

PMI 值能够量化词与词之间的关系,在一定的文本集合中,词 $word_1$ 和词 $word_2$ 的 PMI 值定义如下:

$$PMI(word,word)=\log\left[\frac{P(word_1\ \&\ word_2)}{P(word_1)P(word_2)}\right] \quad (1)$$

其中,$p(word_1\&word_2)$ 表示 $word_1$ 和 $word_2$ 共现的概率。从公式中可以看出,两个单词共现的次数越多,则表明彼此之间的联系越大。为此,我们人工构建了音乐领域的专业术语词表 F,通过 PMI 值计算候选评价对象与词表 F 的语义相似度 $PMI(w_F^i,w_T^j)$,从而将反映领域相关性的评价对象筛选出来。其中,w_F^i 代表音乐领域专用特征词集 F 中第 i 个特征词,w_T^j 代表候选修饰对象集 T 中第 j 个候选修饰对象,值越大,则说明 w_T^j 的领域相关性越强,更可能成为一个评价对象。依据设定的阈值获得最终的修饰对象集 T。后续的实验表明这种方法取得了较好的效果。

2.2　情感词的抽取

基于获取的评价对象集,充分挖掘评价对象与情感词之间的语义关联,从而将反映音乐领域的情感词识别出来。在此,将关联规则挖掘思想引入到情感词的识别与抽取过程中。

(1)构建 Web 商品评论信息的关联规则事务

集。可基于预处理后的亚马逊商品评论信息语料库创建,其中,句子为事务单位,将句子中包含的可能会体现情感特征的形容词或动词提取出来作为词项(item)。与此同时,结合前述的评价对象集,将句中出现的评价对象也提取出来共同构成一个事务集。

(2)对 Web 产品评论信息的事务集进行关联挖掘,获取强关联频繁关系组,并基于此获得评价对象与情感词之间的语义关联关系。

(3)基于最小置信度对所有强关联频繁关系组进行规则产生,并对所产生的规则进行过滤,将只含有一个前件和一个后件的规则提取出来,结合词性判断,最终获得形如 $X \rightarrow Y$ 的规则,其中前件 X 代表某个评价对象 *target_word*,后件 Y 则为领域情感词 *opinion_word*。所有领域情感词可构成情感词典 D_1。

(4)为了获取覆盖面更广的情感词,以 D_1 作为种子情感词典,基于 WordNet 进行同义词扩充,形成情感词典 D_2。将 D_2 与 WordNet 原有的情感词词典 D_3 合并从而形成最终的情感词典 D。

3 基于 PageRank 模型和混合相关关系的情感词极性判别

在网页评级中,一个网页被链接的越多(即被其他网站投票越多)意味着该网页越重要,而来自重要的网页的投票也被认为具有较高的价值。与此相类似,一方面,如果一个情感词与其它持“支持”(“反对”)观点的词紧密相关(链接多),则它将持“支持”(“反对”)观点的概率也越大;另一方面,来自种子词或置信度较高的情感词的“投票”将会被认为具有较高的价值。因此,本文借鉴网页评级中随机游走模型,利用情感词本身的极性强弱(种子情感词)以及情感词与情感词的混合相关关系(直接相关关系和间接相关关系)来推断未知情感词的情感倾向性,构建情感词的情感倾向量化模型。

3.1 问题定义

已知种子情感词向量 $S=\{s_1,s_2,\cdots,s_n\}$,其对应情感值向量 $Y_S=\{y_1,y_2,\cdots,y_n\}$;未知倾向性的情感词向量 $W=\{w_{n+1},w_{n+2},\cdots,w_{n+m}\}$,其对应情感值向量 $Y_W=\{y_{n+1},y_{n+2},\cdots,y_{n+m}\}$。目标是利用 S 的情感值向量 Y_S 来估计 W 的情感值向量 Y_W。

3.2 情感词的情感倾向性量化框架

为了解决该问题,我们在随机游走模型方法基础上充分考虑词项间的混合相关关系,构建情感词的情感倾向性量化框架。具体步骤如下:

(1)种子情感词生成。统计情感词典 D 中所有情感词在语料库中出现的频率并排序,选取 *Top - N* 的情感词作为种子情感词,我们认为这些种子情感词具有较高的置信度,人工标注其情感极性,并设正向情感词取最大情感值 +1,负向情感词取最小情感值 -1。

(2)构建基于随机游走模型的情感图,反映情感词之间的混合相关关系。拟采取以下方式进行具体构建:

Step1:统计情感词典 D 中所有情感词在语料中的出现情况,建立无向图模型 $G(N,E)$。其中,情感词作为 G 中的节点,N 为节点的集合,$|N|=|S|+|W|$,$|S|$为种子情感词数目,$|W|$为未知情感倾向性的情感词数目;E 为边的集合,两个情感词 w_i 和 w_j 若共同出现在一定长度窗口单位内,则 w_i 和 w_j 之间有一条边相连。

Step2:构建连接矩阵 M,连接矩阵 $M(|W|\times|N|)$描述未知情感倾向性的情感词与种子情感词节点之间的无向图连接关系,M 可分解为 $|W|\times|S|$ 的子矩阵 U 和 $|W|\times|W|$ 的子矩阵 V 两部分。u_{ij} 表示未知情感倾向性的情感词 w_i 和种子情感词 s_j 之间的语义相似度,v_{ij} 表示未知情感倾向性的情感词 w_i 和情感词 w_j 之间的语义相似度。

通常衡量词项间语义相似度有互信息、共现频率等方法,该类方法认为如果两个词项在一定长度的窗口单位内共同出现,则说明两个词项之间存在直接相关性,且共现的概率越大,两个词项之间的相关度越高。因此,如果词项在一定长度的窗口单位内没有共同出现,则词项被认为是不直接相关,两者语义相似度值为0。显然完全依赖该方法会产生片面的结果。一方面,实际中由于语料规模受限,一些没有共同出现的词对之间可能也会有相关性;另一方面,共现是以一定长度的窗口单位为限制的,长度选取不同,会造成不同的结果。因此,在计算词项之间的语义相似度时,不仅仅从直接相关关系角度获得,还要考虑间接相关关系。所谓间接相关关系,我们认为包含如下两种情况:一种是在指定长度的窗口下没有共同出现的词项对;另一种是词项对之间并没有直接联系,而是借助第三方词项隐含导出。

假设在语料统计中,$w_{i1},w_{i2},\cdots,w_{im}$ 与词 w_i 有直接相关关系,$w_{j1},w_{j2},\cdots,w_{jn}$ 与词 w_j 有直接相关关系,如图2所示,分析图中词项 w_i 和 w_j 之间的间接

相关关系。

在图2中，我们认为 w_i 和 w_j 之间的间接相关关系与图中 w_i 到 w_j 的途径路径有关。一方面，两个词之间存在的一条路径上中间节点数越少，两者之间的相关度越高；另一方面，两个词之间的连接路径数越多，则两者的相关度越高。

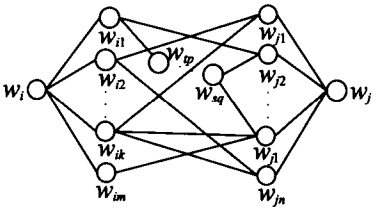

图2 词项间相关关系图

$$id\_sim(w_i,w_j)=Max\left(\frac{C}{I(w_i)*I(w_j)*Num(w_i,w_j)}\sum_{1\le k\le m,1\le l\le n}d\_\mathrm{sim}(w_{ik},w_{jl})\right) \tag{2}$$

$$d\_sim(w_{ik},w_{jl})=MI(w_{ik},w_{jl})=p(w_{ik},w_{jl})*\log\left[\frac{p(w_{ik},w_jl)}{p(w_{ik})\times p(w_{jl})}\right] \tag{3}$$

其中，$C(0<C<1)$ 为衰减系数，表示相似度在传递过程中的衰减速度。$I(w_i)$ 表示 w_i 连接的边数。$Num(w_i,w_j)$ 表示 w_i 和 w_j 之间存在的一条路径上的中间节点数。$d\_sim(w_{ik},w_{jl})$ 表示 w_{ik} 和 w_{jl} 直接相似度，可以采用互信息的方法来计算。由于 w_i 到 w_j 的途径路径有多条，因此取最大值作为最终的间接相似度。

Step3：对 U 和 V 进行联合归一化：

$$\left.\begin{aligned}u_{ij}&=\frac{u_{ij}}{\left(\sum_{j=1}^{|S|}+\sum_{j=1}^{|W|}v_{ij}\right)}\\ v_{ij}&=\frac{v_{ij}}{\left(\sum_{j=1}^{|S|}+\sum_{j=1}^{|W|}v_{ij}\right)}\end{aligned}\right\} \tag{4}$$

(3)基于随机游走模型和混合相关关系的情感词情感值量化计算。根据随机游走模型 PageRank 算法，迭代计算情感词的情感值，借鉴文献[17]提出的迭代公式：

$$y_W^{(n)}=(1-\beta)UY_S+\beta VY_W^{(n-1)} \tag{5}$$

其中，$Y_W^{(n-1)}$ 表示第 $(n-1)$ 次迭代结果，β 为加权系数 $(0<\beta<1)$，对 Y_W 进行归一化操作，并不断重复，直到 Y_W 的值收敛为止。返回结果向量 Y_W，最终获得情感词典中所有情感词的情感值，情感值的正、负分别表示情感极性为“正向”或“负向”。

4 实验与分析

4.1 实验准备

本实验选用伊利诺伊大学芝加哥分校 Hu 和 Liu 提供的亚马逊网站的产品评论信息作为语料[17]。该语料库收录了亚马逊网站上四个大类产品，包括 Book、Music、DVD/VHS 和 mProducts（主要指电子和信息产品）的用户评论信息。该语料集总共包含 6700 万个产品，5800 万评论，整个文件大小为 5.498G。我们首先对该数据集文件进行了如下预处理：

Step1：文件分割。由于整个的评论信息汇总在一个大的文本文件里，不便于后续的处理，因此，我们首先对该文件进行了分割。分别以产品编号和评论编号（即用户编号）为关键字，获得每个产品的所有评论信息和每个用户所发表的所有评论信息，以独立的文本文件形式保存。两种分割方法产生 120 万个产品编号文件和 240 多万个评论编号文件。

Step2：数据的分类。语料集并没有明确每条评论是针对哪类产品进行评价，因此还需将上述文件分割结果对号入座到相应的类别信息里。为了更加精准的获取每个产品编号对应的产品类别，我们采用以下方法：在浏览器中输入网址：http://www.amazon.com/dp/0000000868，其中 0000000868 表示待标注的文件名，也即产品编号，在网页结果中得到该产品编号的类别信息。

Step3：基于上述类别标注的结果，选取 Music 类部分数据作为实验对象（7705 个产品，约 78 521 条评论）进行后续情感词典的构建。

4.2 实验设置与评价

为了衡量本文所提方法的有效性，我们选取了志愿者的判别方法，即首先对获取的情感词进行判别标注（即判断识别出的词语是否为情感词）和相应词的极性（即情感词的正向与负向）标注，然后对参与实验的评论文本进行极性标注。标注均采取少数服从多数的原则，若两个志愿者的判断一致，则以该判断结果作为最终结果，若两个志愿者的判定结果不一致，则以第三人的判断为基准，选取三人中较多一方所体现出来的判定结果作为最终结果，并进行实验评价。

同时，在情感词极性计算实验中，为了避免与未知情感词相似度值过低而造成迭代过程不收敛，本文对反映未知情感倾向性的情感词 w_i 和情感词 w_j 之间的语义相似度矩阵 V 进行了处理，将矩阵 V 中第 i 行元素 V_i 按值由大到小排序，以排名第 k 位元素的值为阈值。对于 V 中任何一值，小于该阈值的元素值设为0，否则，其值保持不变。

4.3 实验结果与分析

4.3.1 情感词的识别

情感词的识别是情感词典构建的首要步骤。本文中，利用关联规则挖掘算法充分挖掘音乐领域主题特征词（评价对象）和情感词之间的关系，从而将情感词抽取与识别出来。实验评价指标采用被同类实验普遍使用的准确率和召回率进行评价。

情感词的正确识别依赖于主题特征词的识别。因此，本文首先对主题词的识别精度进行实验测试，实验效果的评价主要在于观察识别出的主题词是否具备领域相关性，即抽取出的主题词是否体现音乐领域的特征，相应的实验结果如表1和表2所示。

从表1和表2的数据可以看出，识别的主题词不仅具有较高的识别精度，而且领域相关性也比较强。同时我们也观察到召回率性能略低，分析原因，我们认为主要是以下两方面原因导致：

（1）词频过滤使得一些出现次数较低的主题词不能被挑选出来；有些主题词，例如，bassist，中文意思是低音电吉他手，该词属于音乐领域，但是并不大众化，所以在评论中出现次数较少。

（2）PMI语义过滤性能很大程度上依赖于大规模数据集，理论上讲，评论数目越多，则统计效果越明显，PMI值也越准确。本文中，选取的数据规模不算太大，因此造成了PMI值的部分偏差，从而影响领域主题词获取的数量，最终造成主题词的识别召回率较低。

根据主题词和情感词间的关系，我们从Music商品评论信息中最终识别出情感词，其识别准确率表3所示。表3的数据显示，情感词的识别准确率较高，具有较好的性能。这主要源于主题特征词的正确抽取和关联规则挖掘算法的有效实施。在主题特征词具有较高领域相关特性的前提下，结合关联规则挖掘算法能有效的把领域相关的情感词识别出来。同时实验中，我们也观察到有部分情感词没能识别出来。究其原因，由于自然语言表达的复杂性和上下文特性，我们发现有些评论语句在表达情感时并不显式针对某个具体对象特征，而事实上该特征在上一句话中会显式的提及，对于这种情况，关联规则挖掘算法无能为力。

表1 音乐领域主题词识别结果

语料库中不同词项个数	候选主题词个数	词频过滤后主题词个数	PMI值过滤后主题词个数	准确率	召回率
7705	3223	1975	449	0.90	0.79

表2 音乐领域的部分主题特征词结果

Feature	Feature	Feature	Feature	Feature	Feature	Feature
musician	songwriter	tongue	rock	composers	listener	concert
channels	classic	tapes	musicals	chords	piano	solo
chord	player	artistry	tunes	voices	orchestra	bands
composition	synthesizer	albums	tone	rhythm	bass	hazy
cd	keyboards	percussion	melodies	orchestra	vocals	violin
masterpieces	artist	glory	vocalists	styles	singers	audience
choirs	voice	drums	concerts	jazz	performer	disk
mode	harmonica	market	popularity	techno	episode	winners
opera	recordings	disc	instruments	pianist	pop	ballad
producer						

表 3 情感词的识别性能

测试性能	数据
语料中人工判别情感词的总数(个)	1112
算法抽取与识别的情感词个数(个)	907
算法抽取出的情感词中人工判别为情感词个数(个)	834
正确率	0.92
召回率	0.75

4.3.2 情感词的极性分析

基于上述识别出来的情感词,本文对其极性进行了倾向性分析。在进行分析中,我们依次考察了种子情感词数目对情感词极性判断的影响、参数 β 和 k 的取值对算法性能的影响以及混合相关关系在情感词情感极性计算中的效用,实验评价依然采取准确率来衡量。

我们首先以 $k=700$ 为依据,在不同种子集词数和不同 β 值的条件下,进行了敏感程度的实验。由公式(5),$\beta=0$ 表示算法中不考虑未知的情感词之间的语义相似度。实验结果分别如表 4、表 5 和表 6 所示。

表 4 情感词的极性判断结果
(种子集 $=30,\beta=0,k=700$)

实际极性	算法判定极性正向负向		查准率	整体准确率	整体误差率
	正向	负向			
正向	328	104	0.55	0.52	0.48
负向	268	74	0.42		

表 5 情感词的极性判断结果
(种子集 $=40,\beta=0,k=700$)

实际极性	算法判定极性正向负向		查准率	整体准确率	整体误差率
	正向	负向			
正向	359	79	0.59	0.56	0.44
负向	253	63	0.44		

表 6 情感词的极性判断结果
(种子集 $=50,\beta=0,k=700$)

实际极性	算法判定极性正向负向		查准率	整体准确率	整体误差率
	正向	负向			
正向	312	127	0.57	0.51	0.49
负向	233	62	0.33		

从以上数据可以看出,当种子情感词的个数 $=40$ 时,情感词的情感极性判断准确率最大。同时,我们也观察到正向情感词的极性判断准确率要高于负向情感词。分析原因,我们认为可能有以下两方面因素造成:一方面,本文的数据集采用亚马逊网站上的 Music 类商品评论信息,该数据集大部分都是对产品的正向评论,因而本身正向情感词所占比例较大。相应情感词的正向极性判断的概率要高于负向情感词,所以相比负向情感词的判定,表现出更高的准确率;另一方面,种子情感词的质量也会对算法性能产生影响,算法中未知情感词的极性是通过种子情感词推断出来,因此,情感词的极性判断会受到种子情感词的质量影响。

我们依次变换 β 值,得到不同的情感极性判别结果,如图 3 所示。

正如算法所述,参数 β 取值在 0 至 1 之间,端点 $\beta=0$ 表示不考虑未标注的代测情感词之间的语义相似度,端点 $\beta=1$ 表示不考虑种子情感词的情感极性。从图 3 中可以看出,随着 β 值从 0 开始逐渐增大,词语极性的判别准确率也随之提高,一直到 β 取值 0.4 为止,情感词的极性判断准确率达到峰值 0.74,随后随着 β 值的继续增大准确率反而降低,一直到 β 取值为 1 时倾向性判别结果降到最低值,这充分说明了代测情感词之间的语义相似度和种子情感词对整个情感词的极性判断具有较强的指导作用,影响着最终的情感极性判别结果。

最后,基于种子情感词个数 $=40,\beta=0.4$,我们也对 k 值进行了敏感性测试,实验结果如图 4 所示。

从图 4 中可以看出,k 值对算法的性能有较大影响,k 值取值为 700 时算法性能最佳。从性能曲线可以清晰地看出,随着 k 值不断增大,情感词的情感极性判断准确率在不断增加,一直到 k 取值 700 为止,然而,随着 k 值的进一步增大,极性的判断准确率反而呈下降趋势。分析原因,我们认为当 k 值取值较大时,会加入一些连接权重较低的情感词,使其对极性判别造成影响,导致准确率大幅度下降。

4.3.3 混合相关关系在情感词极性判断中的影响分析

在计算情感词极性时,我们考虑了情感词之间的混合相关关系。为了验证混合相关关系在情感词极性判断中的作用,我们在相同参数条件下进行了对比实验。比较考虑混合相关关系与未考虑混合相关关系条件下所获得情感词的情感极性准确率,实验结果如表 7 所示。

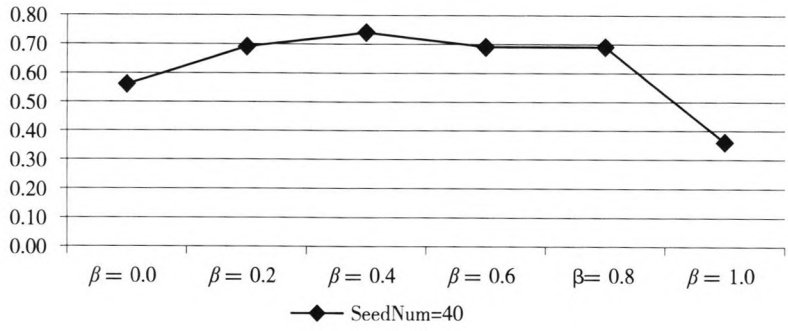

图3 参数 β 值的实验结果图

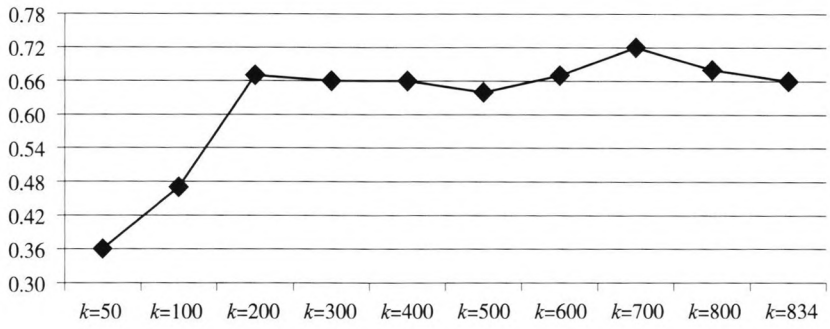

图4 k 值的敏感性分析实验图

表7 混合相关关系影响分析结果(种子集=40,β=0.4,k=700)

比较对象	实际极性	算法判定极性		查准率	整体准确率	整体误差率
		正向	负向			
考虑混合相	正向	465	71	0.79	0.74	0.26
关关系	负向	126	92	0.56		
不考虑混合	正向	428	137	0.77	0.65	0.35
相关关系	负向	126	63	0.32		

从表7中的数据可以看出,考虑了混合相关关系所获得的情感词极性判断正确率明显要高于未考虑混合相关关系,性能提高了14%,说明充分考虑混合相关关系能有效提高情感词的情感极性判断准确率。

5 结 论

情感词典的构建是情感分析领域的基础性工作,包含情感词的识别及其极性判断两大任务。本文提出了一种基于关联规则挖掘和极性分析的情感词典构建方法,其主要贡献可概括为以下两方面:

(1)本文提出了基于关联规则挖掘的情感词识别与抽取方法。与以往的工作不同,为了能较为准确的把情感词识别出来,我们对情感表达的主题词进行了前期预处理,使得获取的主题词能反映领域特征,依据主题特征词与情感词之间的上下文语义关系,从而保证后续情感词识别的准确性。

(2)本文构建了情感词的量化模型。不同于其他人的工作,该模型在网页评级随机游走模型的基础上充分考虑了情感词之间的混合相关关系,这种混合相关关系不仅仅表现为直接相关关系,更多的则是间接相关关系,并通过这种混合相关关系来体现词项间的语义相似度,从而推断未知情感词的情感倾向性,取得了较好的实验效果。

在今后的工作中,我们会进一步扩充实验数据集,进一步验证本文所提方法的有效性,对实验中的参数进行优化工作。同时,针对情感词在不同语境下表现出的不同极性,进行更加细粒度的情感倾向性计算。

参 考 文 献

[1] Na S, Lee Y, Nam S, et al. Improving Opinion Retrieval

based on Query - Specific Sentiment Lexicon [C]// Proceedings of ECIR '09. Toulouse: [s. n], 2009: 734-738.

[2] Hatzivassiloglou V, McKeown K R. Predicting the Semantic Orientation of Adjectives[C]//Proceedings of the 35th Annual Meeting of the Association for Computational Linguistics (ACL), Madrid, Spain, 1997: 174-181.

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

[4] Turney P D, Littman M L. Measuring praise and criticism: inference of semantic orientation from association [J]. ACM Transactions on Information Systems, 2003, 21 (4): 315-346.

[5] Ding X W, Liu B, Yu P S. A Holistic Lexicon-based Approach to Opinion Mining[C]//Proceedings of First ACM International Conference on Web Search and Data Mining, 2008.

[6] Lu Y, Castellanos M, Dayal U, Zhai C. Automatic Construction of a Context-aware Sentiment Lexicon: an Optimization Approach[C]//Proceedings of International Conference on World Wide Web (WWW), Hyderabad, India, March 28 - April 1, 2011.

[7] Mohtarami M, Lan M, Tan C L. From Semantic to Emotional Space in Probabilistic Sense Sentiment Analysis [C]//Proceedings of the 27th AAAI Conference on Artificial Intelligence (AAAI), Bellevue, Washington, USA, July 14-18, 2013: 717-717

[8] Kamps J, Marx M, Mokken R J, et al. Using WordNet to Measure Semantic Orientation of Adjectives [C]// Proceedings of 4th International Conference on Language Resources and Evaluation (LREC), 2004: 1115-1118.

[9] Esuli A, Sebastiani F. Determining the Semantic Orientation of Terms through Gloss Analysis [C]//Proceedings of ACM Conference on Information and Knowledge Management (CIKM), 2005: 617-624.

[10] Esuli A, Sebastiani F. Determining Term Subjectivity and Term Orientation for Opinion Mining [C]// Proceedings of the 11th European Chapter of the Association for Computational Linguistics (EACL), 2006: 193-200.

[11] Hu M Q, Liu B. Mining and Summarizing Customer Reviews[C]//Proceedings of the 10th ACM International Conference on Knowledge Discovery and Data Mining (SIGKDD), 2004: 168-177

[12] Qiu G, Liu B, Bu J, et al. Opinion Word Expansion and Target Extraction through Double Propagation [J]. Computational Linguistics, 2011, 37(1): 9-27.

[13] Takamura H, Inui T. Latent Variables for Semantic Orientation of Phrases [C]//Proceedings of the 11th Conference of the European Chapter of the Association for Computational Linguistic (EACL), Trento, Italy, 2006: 201-208.

[14] Rao D and Ravichandran D. Semi-supervised Polarity Lexicon Induction[C]//Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics (EACL). Athens, Greece, 2009: 675-682.

[15] 宋晓雷,王素格,李红霞. 面向特定领域的产品评价对象自动识别研究[J]. 中文信息处理学报,2010, 24(1):89-93.

[16] 李荣军, 王小捷, 周延泉. PageRank 模型在中文情感词极性判别中的应用[J]. 北京邮电大学学报, 2010, 33(5): 141-144.

[17] Nitin Jindal, Liu Bing. Opinion Spam and Analysis [C]//Proceedings of First ACM International Conference on Web Search and Data Mining (WSDM-2008), Feb 11-12, 2008, Stanford University, Stanford, California, USA.

(责任编辑 马 兰)