文章编号：1003-0077(2018)08-0032-10

大规模中文实体情感知识的自动获取

卢 奇[1,2]，陈文亮[1,2]

(1. 苏州大学 计算机科学与技术学院，江苏 苏州 215006；
2. 软件新技术与产业化协同创新中心，江苏 苏州 215006)

摘 要：目前中文情感分析的主要资源以情感词典为主，缺乏针对实体或属性的情感知识资源。该文主要研究如何从大规模文本语料中自动获取实体情感知识。在该文方法中，用情感表达组合来表示实体情感知识。首先，基于二部图排序算法对情感表达组合候选集合进行排序。然后，提出了一种基于语义相似的提炼算法对于排序靠后的表达组合进行选择。在提炼选择过程中，充分考虑实体之间和情感词之间的约束。最后，该文在三种大规模不同领域的语料上进行实验，并进行人工评价。评价结果表明，从三个领域数据集上获取的实体情感表达组合正确率均高于 90%。最终我们获得了一个大规模情感知识词典，包括约 30 万对的情感表达组合。

关键词：情感分析；情感词典；情感挖掘；信息抽取

中图分类号：TP391　　**文献标识码**：A

Automatically Building a Large Scale Dictionary of Chinese Entity Sentiment Expressions

LU Qi[1,2], CHEN Wenliang[1,2]

(1. School of Computer Science and Technology, Soochow University, Suzhou, Jiangsu 215006, China;
2. Collaborative Innovation Center of Novel Software Technology and Industrialization, Suzhou, Jiangsu 215006, China)

Abstract: Except for some sentiment dictionaries. There are not sentiment expressions for entities which are very important for analysis. This paper proposes a method of automatically building a dictionary of entity sentiment expressions from large-scale raw text. In our method, we use a sorting algorithm based on a bipartite graph to rank the candidates of sentiment expressions. Then, we present a refining algorithm according to semantic similarity to extract some expressions from the low-rank set. Finally, we conduct the experiments on three datasets from different domains. The experimental results show that the accuracy of the extracted expressions is better than 90%. Totally we obtain a large scale dictionary including about 300K sentiment expressions.

Key words: sentiment analysis; sentiment dictionary; sentiment mining; information extraction

0 引言

情感分析的目的是让计算机理解人类情感，它是自然语言处理(NLP)的一个重要研究方向[1-3]。情感分析可以应用于信息抽取、舆情分析等任务中。目前用于情感分析的资源大都是情感词典，其中英文情感词典有 SentiWordNet、LIWC、ANEW、MPQA 等，中文情感词典有《学生褒贬义词典》、知网的《情感分析用词语集》、台湾大学的《情感词典》[4]、清华大学的《情感词典》[5]以及北京大学的《情绪词典》[6]等。这些情感词典对情感分析提供了帮助，提高了分析性能。但是仅仅利用情感词典收录的情感词进行实体观点的倾向性分析效果通常欠佳。

随着智能互联网时代的到来，人们的需求发生了变化，人们开始大量使用移动设备对自己的购物、旅游等活动发表自己的看法和评论。这些评论的数目增长速度非常快，很多网站的评论数目达到千万量级或者更多。由于数量巨大，用户在浏览时会遇到很大困难。在为用户提供评论的总结性摘要过程中，评论描述的实体、属性或者对象的情感分析变得

收稿日期：2017-10-17　**定稿日期**：2017-12-19

基金项目：国家自然科学基金(61572338)；江苏省高校自然科学研究重大项目(16KJA520001)

尤为重要[7]。为了便于描述，下文统一用“对象”来表示实体、属性。例如，在购物网站上，某款型号电脑的好评率只能给用户一个大致的印象，还不能提供足够信息帮助客户决定是否购买。消费者根据自己的需求，更想了解这款电脑各个方面的具体评价总结。电脑的重要特性包括屏幕、电池、CPU、内存、散热等多个方面的性能都会影响客户判断。从购物网站的大量评论中自动抽取现有消费者对商品各个方面的褒贬观点将为潜在用户提供很大的帮助。因此，从评论中挖掘具体某个对象的消费者观点是一个非常有意义的研究课题。

但是，在评论中一个句子同时存在多个对象和多个情感词，使得自动挖掘对象的消费者观点变得很困难。在挖掘过程中如果有实体或属性情感表达组合词典，将会帮助歧义消解。在本文中，情感表达组合采用二元对的方式表达：对象—观点词，例 1 给出了几个情感表达组合的二元对。

例 1　“价格—高”、“性价比—高”、“食物—精致”、“长城—雄伟”

构建情感表达组合的另一个原因在于：很多特定的观点词只能用于特定的对象。比如“壮观”“壮阔”形容风景类对象，“鲜美”“肥美”形容食物类对象。因此，如果有大量正确的情感表达组合作为情感资源，将对特定对象的情感分析提供很大的帮助。基于这个目的，本文主要研究从语料中提取情感表达组合，建立用于情感分析的词典资源。

本文的工作分为三步：情感表达组合候选集的获取、情感表达组合的排序、情感表达组合的提炼。在候选集获取阶段，我们通过词性匹配来抽取情感表达组合的候选集，并且保留情感表达组合和模式之间的关系结构。由于中文表达的多样性，这些候选集包含着大量错误和噪声。因此本文的难点是如何从大量的候选组合中将正确的情感表达组合挑选出来。针对这个困难，我们从两方面着手：(1)排序：在排序阶段，我们通过二部图排序算法利用情感表达组合和模式之间的关系结构进行排序，同时对模式进行必要的调整。(2)提炼：我们利用排序阶段得到的排序结果靠前的情感表达组合作为参考标准。基于语义的相似性，通过本文提出的提炼算法在排序靠后的结果里进行提炼，获得更多的情感表达组合。实验结果表明，二部图排序算法能很好地对情感表达组合进行高质量排序。同时，提炼过程又弥补了二部图排序算法的一些缺点，进一步成功提取出正确率 90％以上的情感表达组合。

1　相关工作

1.1　情感词典构建

本文工作是为了建立用于情感分析的词典资源。目前，情感资源构建工作主要以情感词典为主。Esuli 等[8]和 Baccianella 等[9]以 WordNet 为基础构建了 SentiWordNet。首先从几个褒贬词通过二元关系扩展词语得到种子词集，然后用褒贬种子词集和中性种子词集来训练三分类器，之后通过分类器对 WordNet 中所有词集标记情感倾向，最后通过随机游走模型分别对得到的褒贬词集进行情感倾向性调整，直到最终收敛。这是从词义关系上构建情感词典。Turney[10]利用点对互信息来计算短语的褒贬程度。在此基础上，Banea 等[11]计算新的候选词情感得分后，利用相似性度量、过滤并保留与原始种子集最相似的新词集。这两者都是基于同现关系来构建情感词典。Hatzivassiloglou 等[12]在两千多万篇新闻语料上，以形容词作为候选词，利用连词构建词语间的相互关系，将相近词语聚成簇，通过簇内已知的情感词来判断整个簇的情感倾向。该方法对数据量要求大，利用句法关系构建情感词典，适合大规模语料。Kanayama 和 Nasukawa[13]扩展了 Hatzivassiloglou 的方法，提出了句子内部和句子之间的情感关联性思想。他们认为连续的若干句子往往具有相同的情感倾向，如果其中一个句子含有情感词，那么连接它的句子也会含有情感词，并具有相同的情感极性。这种方法在上下文句子中没有情感词的情况下召回率会大大降低。Qiu 等人[14-15]沿袭了 Kanayama 和 Nasukawa 的工作，他们利用评价词和评价对象的关系抽取情感词并判断其极性，提出了双重传播(double propagation)[16]的思想，这种 bootstraping 的思想联合了抽取评价词和评价对象。他们借助依存句法、POS 标注、parser 结果来分析评价词和评价对象之间的关系，再根据定义的八条规则迭代扩展情感词集。这种方法大大增加了召回率，但在词典扩展的过程中由于引入了噪声导致准确率不够高，另外这种方法不适合处理网络上一些非正式的文本。Agathangelou 等人[17]在总结研究学者的经典方法后构建了一个多步的方法，同时利用连词和双重传播的方法抽取情感词，并利用一些语言学模式进行词语的极性消歧。

此类方法优点是比较简单，针对性强，能够抽取

特征领域的情感词;缺点在于较耗时,人工定义规则也相对有局限性,可扩展性差,在处理网络上的那些非正式文本时利用语法信息往往会产生很多错误。此外,目前情感词典构建工作都局限于构建单一的情感词。没有将对象和情感词作为一个整体的"情感知识"进行这种二元对情感词典构建工作。

1.2 基于对象的情感分析

和本文另一种类似的相关工作属于评价对象和评价词的抽取工作。李智超[18]将形容词作为观点词,利用模式匹配规则抽取属性词(即对象),对未登录词通过"上下文熵"法进行挑选,在特定领域的语料有较好的效果,但是需要人工干预。Popescu[19]构建了一个信息抽取系统OPINE,通过名词和具有一定区分的符号间的点互信息值获得产品特征,利用人工构建的10条规则识别观点词。刘鸿宇等[20]利用句法分析结果获取候选评价对象,结合PMI算法和名词剪枝算法对候选评价对象进行筛选。然后分析情感句句型并归纳分析规则,使用无指导的方法完成评价对象在情感句中的倾向性判断。Zhuang等[21]采用WordNet、电影知识和标注训练数据等生成关键词列表,利用规则获得对象和观点对。Kobayashi等[22]利用文本挖掘技术,提出了一种半自动快速收集评价表达的方法。Somprasertsri等[23]和王素格等[24]在句法信息和语义信息的基础上,提出了一种采用依存关系提取情感表达组合的方法。

基于对象的情感分析结果可以作为本文任务的第一步——构建候选集合。由于缺乏人工标注语料,我们采用无监督方法,即基于pattern方法。目前基于对象的情感词识别任务的识别结果正确率都不是很高,所以本文想通过大规模语料中的不同句子之间的约束关系来寻找不同的对象之间的情感词的差异性。

2 情感表达组合的获取

2.1 情感表达组合候选集的获取

2.1.1 对象—观点对抽取

本文利用词性信息来寻找对象和观点词的候选。对象的词性集合:$N=\{n, ns, vn, nz, s, nr\}$,其中$n$=名词、$ns$=地名、$vn$=动名词、$nz$=其他专名、$s$=处所词,$nr$=人名。观点词的词性选择和其他研究者[10,18]的做法一致,以形容词为判断标准:$S=\{a\}$,其中a=形容词。

抽取实例如图1所示。在抽取时,我们将每句话中词性满足集合N的对象{名气,性价比}添加到列表N_list中,将词性满足集合S的观点词{大,高}添加到列表S_list中。该句中情感表达组合的最大组合数为2×2=4。然后,将对象和观点词的中间词串作为pattern添加到对应情感表达组合的pattern列表里。

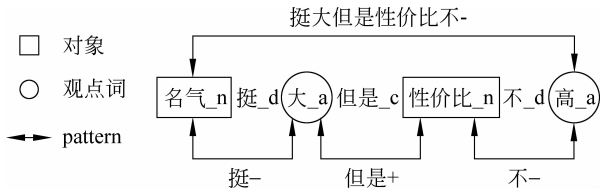

图1 情感表达组合的抽取实例

为了更准确地反映情感表达组合和pattern之间的映射关系,我们做了以下处理。在例2和例3中都有pattern{的}。但例2中的"漂亮"修饰"花朵",而例3中"具体"却不修饰"订单",而是修饰后面的"情况"。在这两句中,"的"的作用不同。例2中句子的形式是"观点词+pattern+对象",而例3却是"对象+pattern+观点词"。若不加以区分,那么这两句的pattern被认为是同样的"的",这和实际情况不符。所以在抽取时,我们用(+,-)代表方向,将"对象+pattern+观点词"形式的pattern记作pattern-;而"观点词+pattern+对象"形式的pattern记作pattern+,对pattern做更细致的区分。

例2 漂亮_a 的_uj 花朵_n

例3 订单_n 的_uj 具体_a 使用_v 情况_n

在整个抽取过程中,我们还统计情感表达组合二元对被pattern匹配的具体次数,即"二元对—pattern—匹配次数"形式。图1中抽取后得到的结构如下,在实际抽取过程中,随着"性价比不高"出现得越多,则其对应的次数也随之增加:

[名气—大]—— 挺- ——1

[名气—高]—— 挺大但是性价比不- ——1

[性价比—大]—— 但是+ ——1

[性价比—高] —— 不- ——1

2.1.2 情感表达组合的调整

例4 简陋_a 的_uj 硬件_n 和_c 粗糙_a 的_uj 服务水平_n

在2.1.1中得到的情感表达组合和pattern的

关系结构中，存在一些常见噪声。例 4 给出了一个典型的情况，该句中有两个对象{硬件，服务水平}和两个观点词{简陋，粗糙}。按照前面的抽取规则，我们会得到以下结构：

[硬件—简陋]—— 的＋ ——1

[硬件—粗糙]—— 和－ ——1

[服务水平—粗糙] —— 的＋ ——1

[服务水平—简陋]—— 的硬件和粗糙的＋——1

由于“和”这样的 pattern 属于很常见的噪声，会影响 2.2 排序算法中的效果。故而，我们将并列连词的 pattern 去除，例如，和，又，而且，而等。并且根据 2.1.1 中给出的实例，由于“观点词＋‘的’＋对象”属于合理情况，而“对象＋‘的’＋观点词”属于较常出现的噪声，因此我们去除了 pattern“的－”保留“的＋”。至于“的硬件和粗糙的”这种出现次数显然较少的 pattern 可以通过排序算法很容易剔除，因此不需要额外处理。另外，考虑到自动分词会导致词性标注中存在些许典型的错误，对文本进行分析后，我们采用了黑名单词典的方法去除了“时候，人，免费，美”这些较频繁的非对象词语。

2.2　情感表达组合的排序

在 2.1 中，我们抽取了大量的情感表达组合作为候选集。本文获取候选集方法比较简单，但是我们的主要目的是尽可能地获取更多的候选集合，为后续排序提供候选。候选集包含对象观点二元对 pair 和 pattern 之间的映射关系。但是，这些映射关系有很多的错误。为此，我们设计了新的排序算法来挑选情感表达组合。排序算法借鉴了 PageRank[25] 算法的核心思想进行以下两个假设：

(1) 如果一个 pair 可以被很多 pattern 多次匹配，那么说明这个 pair 比较重要，其分数相对较高；

(2) 如果一个 pattern 可以被很多分数较高的 pair 匹配，那么该 pattern 的分数也会相应地提高。

2.2.1　二部图排序算法

2.1 中我们获得了情感表达组合和 pattern 之间的映射关系，如图 2 所示，该关系是一种二部图结构。我们可以将这种关系结构转化成矩阵的形，如图 3 所示。Zhang 等[26] 通过服务和应用之间的二部图关系对其进行排序，得到了很好的排序结果。我们借鉴这种排序算法，同样利用这种映射关系对情感表达组合进行排序。

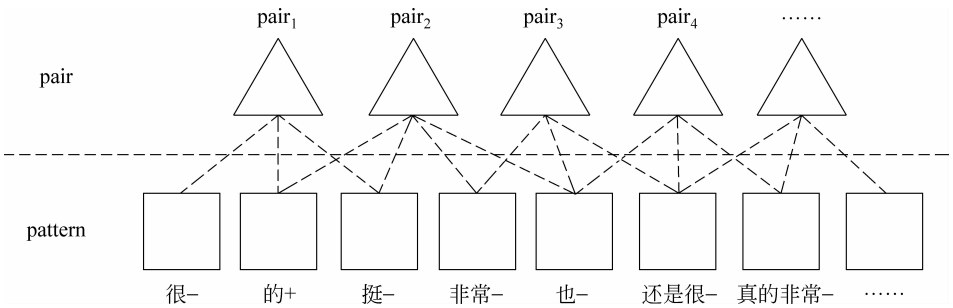

图 2　二部图模型，pair 和 pattern 的关系

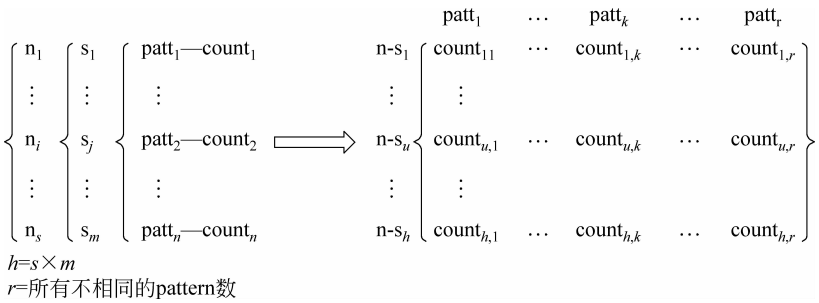

图 3　结构关系转化为矩阵

二部图排序算法的矩阵迭代计算方式，如式(1)所示。

$$\begin{cases} \boldsymbol{A}_i = \boldsymbol{B} \cdot \boldsymbol{C}_i \\ \boldsymbol{A}'_i = \mathrm{norm}(\boldsymbol{A}_i) \\ \boldsymbol{C}_{i+1} = \boldsymbol{B}^{\mathrm{T}} \cdot \boldsymbol{A}'_i \\ \boldsymbol{C}'_{i+1} = \mathrm{norm}(\boldsymbol{C}_{i+1}) \end{cases} \tag{1}$$

式(1)中，$\boldsymbol{B}$ 是图 3 中转化得到的关系矩阵，$\boldsymbol{A}$ 和 $\boldsymbol{C}$ 是一维矩阵。$\boldsymbol{C}$ 代表情感表达组合二元对的分数矩阵，初始化矩阵向量全为 1。$\boldsymbol{A}$ 是 pattern 对应的分数矩阵。期间，每一次矩阵运算结束都要对 $\boldsymbol{A}$ 或 $\boldsymbol{C}$ 的结果进行标准化处理，保证 $\boldsymbol{C}_i$ 和 $\boldsymbol{C}_{i+1}$ 具有相同的总分数。最终，通过式(1)进行迭代运算直至 $\boldsymbol{C}_i$ 和 $\boldsymbol{C}_{i+1}$ 近似收敛，此时可得到每一个情感表达组合的分数以及 pattern 的分数。实际上，把式(1)中两个式子结合起来，迭代形式和 PageRank 算法等效，即：

$$\boldsymbol{C}_{i+1} = \boldsymbol{B}^{\mathrm{T}} \cdot \boldsymbol{B} \cdot \boldsymbol{C}_i \tag{2}$$

norm：

$$X'_j = \frac{X_j}{\sum_{j=1}^{N} X_j} \times N, X \in R^{N\times 1} \tag{3}$$

式(3)中，$\boldsymbol{X}$ 是需要标准化处理的矩阵，N 是 $\boldsymbol{X}$ 矩阵的维度，即标准化后 $\boldsymbol{X}$ 矩阵的总分。式(3)对矩阵 $\boldsymbol{X}$ 进行标准化处理，使得每一维分数按照所占的比例重新分配分数，分数之和为 N。之所以这样标准化处理，是因为实际中矩阵维度很高。如果控制总分数和为 1，则语料之间不具备可比性，因为每一维度平均被分配的分数不同。随着规模越大，每一维被分配的分数会变低。在同一个收敛阈值情况下，维度越高，直观上收敛速度越快，但是收敛越不可靠。

此外，本文中采用的排序算法和 PageRank 算法归一化的方式不同。PageRank 算法是对矩阵 $\boldsymbol{B}^{\mathrm{T}} \cdot \boldsymbol{B}$ 进行归一化，然后循环迭代直至收敛。而本文提出的二部图排序算法是对每一次运算得到的一维矩阵结果进行标准化处理。两种迭代方式最终都可以收敛，本文方法的收敛性在数据分析中给出。在收敛性判断上，当 $\boldsymbol{C}_i$ 和 $\boldsymbol{C}'_{i+1}$ 之间的所有项差值小于 1E-7 时，停止迭代，我们认为情感表达组合此时已经收敛。图 4 中给出了式(1)中 $\boldsymbol{A}_i = \boldsymbol{B} \cdot \boldsymbol{C}_i$ 的示意图。

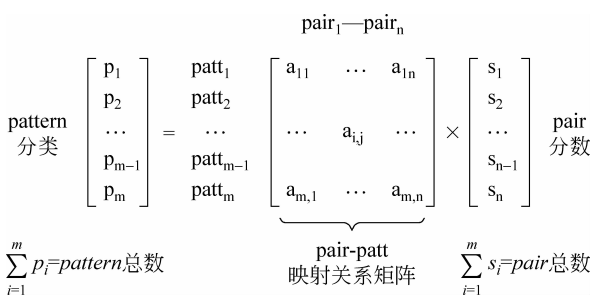

图 4 $\boldsymbol{A}_i = \boldsymbol{B} \cdot \boldsymbol{C}_i$ 示意图

2.3 情感表达组合的提炼

2.2 中通过二部图排序算法得到了情感表达组合的排序结果。经过对结果的采样分析，排序靠前的情感表达组合正确率较高，但是二部图排序算法在召回方面有一定的缺陷。排序完成后，依然存在一些正确的二元对被排在了靠后的位置，比如例 5。

例 5 ［杜鹃花—灿烂］—— 盛开的＋ —— 1
［杜鹃花—灿烂］—— 开得很－ —— 1

该情感表达组合有两个 pattern 分别代表了“灿烂盛开的杜鹃花”以及“杜鹃花开得很灿烂”。由于该二元对出自旅游语料，而整个语料中涉及花的评论很少，导致了“开得很”这个 pattern 出现次数较少。经过统计该 pattern 在语料中仅仅出现 16 次，在迭代中获得的分数较低。因此仅包含该 pattern 的二元对分数远低于其他二元对，导致了“杜鹃花 灿烂”排名较低。反之，在一个关于花的评论语料中进行情感表达组合的抽取并排序后，“开得很”这个 pattern 的分数会因为映射到更多的二元对使得分数变高，这样就能成功将和花有关的情感表达组合排序靠前。同时，只要有其他的评价花朵的 pair 识别出来后，如“牡丹—灿烂”或者“杜鹃花—鲜艳”，那么我们自然而然地想到利用二部图排序算法本身得到的高质量抽取结果，并利用 pair 之间的相似性进行提取。

基于这样的原因，我们有必要在排序靠后的结果中进行提炼，进一步将正确的情感表达组合抽取出来。基于此，我们提出两个假设：(1)某个对象的观点词具有一定的语义相似性；(2)某个观点词描述的对象也具有一定的语义相似性。比如对象“长城”，它所拥有的观点词有“雄伟”“壮观”“宏伟”等。同理，观点词“繁茂”一般形容“林木”“灌木”这些。这些对象或观点词语义相似度很高，我们可以利用这一点来进行提炼。在语义相似度计算上，我们使用了 Google 的 word2vec① 模型[27]。

2.3.1 算法描述

算法 1 情感表达组合提炼算法

输入 1：	划分后标记为正确的 OK_list 表，表中元素为 pair(对象 n-观点词 s)
输入 2：	划分后标记为代提炼的 NO_list 表，表中元素同上

① http://radimrehurek.com/gensim/models/word2vec.html

word2vec 模型 model 文件，
输入 3：计算对象 n 或观点词 s 的相似度，如 $sim_{si} = F(s, s_i)$
输入 4：相似度平均分数指标 score
输出：提炼出的情感表达组合列表 Pair_list

```
Step1   For eachpair in NO_list:
        //pair(对象 n-观点词 s):
            SIM_s = SIM_n = []
Step2       //对每一个待提炼的 pair 初始化它们的相似分数组
            For each s_i in OK_list[n]:
Step3       //遍历 OK_list 中 n 的所有观点词组成集合{s1,s2,s3,...}
                sim_si = F(s, s_i)
Step4           SIM_s.append(sim_si)
Step5       score_ns = sum(SIM_s) / len(SIM_s)
            For each n_i in OK_list[s]:
Step6       //遍历 OK_list 中 s 的所有对象组成集合{n1,n2,n3,...}
                sim_ni = F(n, n_i)
Step7           SIM_n.append(sim_ni)
Step8       score_sn = sum(SIM_n) / len(SIM_n)
Step9       If score_ns > score and score_sn > score:
                        Pair_list.append(pair)
Step10  return Pair_list
```

算法 1 中，OK_list 和 NO_list 是根据对排序结果抽样检查进行划分。根据抽样检查结果，我们以前 10%作为合格部分，后 90%作为不合格待提炼部分。对每份数据固定 10%划分可以实现整个抽取过程的自动化。但是由于语料质量的差异，也可以统计正确率后再对语料进行合理的划分。实现算法之前事先利用 word2vec 模型将语料训练成 model 文件。确定相似度分数 score 后开始提炼，将最终 scorens 和 scoresn 都满足 score 的 pair 保留为结果，类似于取交集的过程。

3　实验结果与分析

3.1　实验数据

本实验使用了三种语料：新闻语料来源于 GIGAword①，餐馆语料来自大众点评，旅游语料来自携程，其中餐馆语料和旅游语料是用户评论文本。我们对语料进行预处理：句子切分、分词、词性标注。表 1 是语料的相关统计数据以及候选对抽取的结果统计。从表中可以看出候选对的规模较为庞大，但是经过检查后发现正确率不高。

表 1　语料规模及抽取数量

语料	句子片段数	pair 数量	pattern 数量
新闻	15 887 167	1 198 320	1 772 387
餐馆	13 887 566	935 133	1 884 141
旅游	2 857 252	266 934	332 954

3.2　实验结果

我们的实验过程主要分成三个步骤：情感表达组合候选集的获取；情感表达组合的排序；情感表达组合的提炼。表 2、表 3 分别展示了 2.2 排序和 2.3 提炼两个阶段实验的正确率统计结果。所有统计数据都经过两名研究生参与评价，各抽取样例 50 个，取平均值。若正确率统计结果差距超过 4%，则重新抽取并进行正确率统计。

表 2　二部图排序后正确率统计

语料	0%～10%	10%～20%	20%～30%	30%～40%	40%～70%	70%～100%
新闻	91%	77%	56%	44%	25%	14%
餐馆	92%	71%	63%	46%	31%	15%
旅游	97%	93%	90%	69%	54%	21%

表 2 展示了排序实验结果，其中 $M\%\sim N\%$ 表示排序后的结果分布。实验结果表明，本文提出的二部图排序算法能有效地对情感表达组合进行排序。在表 2 中，三种语料中前 10%的情感表达组合都达到了 90%以上的正确率，随排序往后其正确率也随之降低。由于二部图排序算法效果表现优秀，后 60%正确率才开始大幅度下降，故不再按 10%作为划分标准，而是以 30%为一组进行统计。旅游语料相比较另外两个语料质量较好，在前 30%的结果都保持了较高的正确率，故而在实验 2.3 中只对其后 70%的结果进行提炼，新闻和餐馆语料都是对后 90%的结果进行提炼。

表 3　对排序结果进行提炼的正确率统计

语料	>0.3 数目	>0.25 数目	>0.2 数目	>0.15 数目	>0.1 数目
新闻	93% 8 004 个	89% 17 198 个	81% 37 374 个	74% 88 234 个	68% 220 599 个
餐馆	91% 7 136 个	88% 20 851 个	90% 52 204 个	81% 112 333 个	75% 225 105 个

① https://catalog.ldc.upenn.edu/LDC2011T13

续表

语料	>0.3 数目	>0.25 数目	>0.2 数目	>0.15 数目	>0.1 数目
旅游	94% 655个	91% 2 013个	90% 5 624个	83% 14 211个	77% 35 858个

表3中第一栏“>K”表示取相似度K以上的值时对应的正确率统计。从结果中可以看出，新闻语料区别于用户评论语料，正确率下降得最快。当相似度平均分低于0.25时，正确率开始有明显的降低，此时提炼出情感表达组合有17 198个。餐馆语料平均相似度在0.2以上保持了较好的正确率，有情感表达组合52 204个。携程和餐馆都属于评论语料，和餐馆提炼效果相似，在均分0.2以上保持了较高的正确率，但是语料规模相对较小，只有5 264个情感表达组合。

我们构建了两种Baseline进行对比：(1)Baseline1：使用情感表达组合候选集直接作为系统结果；(2)Baseline2：按照情感表达组合的出现次数高低进行排序作为系统结果。二部图排序和Baseline2的对比如表4所示。

表4 二部图排序和Baseline2的对比($P@N$)

语料	P@10	P@50	P@100	P@200	P@200-Baseline2
新闻	100%	100%	99%	99%	94%
餐馆	100%	100%	100%	99.5%	93%
旅游	100%	100%	99%	99%	94.5%

由于获得的结果数量较大，我们无法计算召回率和F值。在实验中，我们使用了信息检索的评价方法$P@N$进行评测。我们分别计算$P@10$、$P@50$、$P@100$和$P@200$结果。在表4的结果中，可以很明显地看出二部图排序算法可以提供更好的结果。此外，我们对错误结果进行了分析，餐馆和旅游语料的三处错误是由分词和词性错误引起。而新闻语料是由于其领域的句式复杂度引起，如“基础-平等”。由于新闻语料中较为频繁的出现句式“在xxx的基础上平等地xxx”，导致了{pair：“基础—平等”，patt：“上”}的比例很高，在算法迭代中获得了较高的分数。

接着我们进一步扩大评价的范围。表5展示了Baseline1、Baseline2及二部图排序算法的对比结果，其中Baseline1是对整个候选集随机选取样本进行正确率评估。从旅游语料的三个区间结果上看，基于词频的排序Baseline2在10%之后正确率开始趋于Baseline1，这说明词频和正确率还是存在一定的正相关性。从表中我们还可以看出，二部图排序算法比Baseline2可以更好地进行排序。此外，我们还将二部图排序和Baseline2相应区间的集合进行了比较，来计算它们之间的不同，在表5中用“集合差占比”表示。在“集合差占比”一栏中，在0%～10%区间，两者之间的集合差别在50%左右，这说明二部图排序算法可以将大量正确的低频情感表达组合排在较前位置。

表5 二部图排序和Baseline1、Baseline2的对比($P@N\%$)

语料排序	Baseline1	Baseline2	二部图排序算法	集合差占比
旅游 0%—10%	56%	86%	97%	49.5%
旅游 10%—20%	56%	66%	93%	90.5%
旅游 20%—30%	56%	58%	90%	87.1%
餐馆 0%—10%	43%	80%	92%	49.8%
新闻 0%—10%	39%	70%	91%	65.4%

最终，我们得到了正确率90%以上的新闻情感表达组合的二元对13.7万个，餐馆二元对14.5万个，旅游二元对3.2万个，共计31.4万个。但是从表3提炼的数量上来看，虽然保证了正确率，可是在剩下的结果中依然存在一些情感表达组合未抽取出来。

此外，由于我们的数据量较为庞大，50个抽取样本可能不足以精确地评估二部图排序算法的好坏，我们在携程的数据集上进行了4次抽样统计，来检测正确率偏移情况，结果如表6所示。

表6 旅游语料中随机采样50个样本对正确率统计的影响

次数	0%～10%	10%～20%	20%～30%	30%～40%	40%～70%	70%～100%
1	98%	92%	88%	64%	48%	18%
2	98%	94%	90%	66%	50%	18%
3	96%	94%	92%	70%	54%	20%
4	98%	92%	88%	68%	54%	22%
偏移值	2%	2%	4%	6%	6%	4%

从表 6 可以看出，在前 20％正确率偏移较小，从 20％往后，正确率偏移现象开始逐渐明显，这说明对于排序靠后的结果，50 个统计量不够精确。但是由于我们的目的是获取高正确率的情感表达组合，在前 30％的结果中，50 个统计量足够体现二部图排序算法的排序结果。

3.3　实验分析

3.3.1　算法描述

在排序实验中，pair 分数变化以 1E-7 作为收敛状态值。图 5 中记录了从第一次迭代开始情感表达组合二元对的收敛情况。从中可以看出语料规模越大，收敛速度越快，三个语料的迭代曲线都是单调递减。但在收敛过程中并不是一直平稳地下降，期间三个语料都出现了在趋于稳定时陡然下降的情况。这是因为 pair 和 pattern 的映射关系导致了部分 pair 出现一种“抱团”现象。拥有相同 pattern 的 pair，它们的变化幅度相同，形成了一个小集体。当新的一组 pair 分数变化开始小于 1E-7 后，立刻从不稳定状态变成了稳定状态，这就出现了图中曲线在稳定前发生的骤降现象。

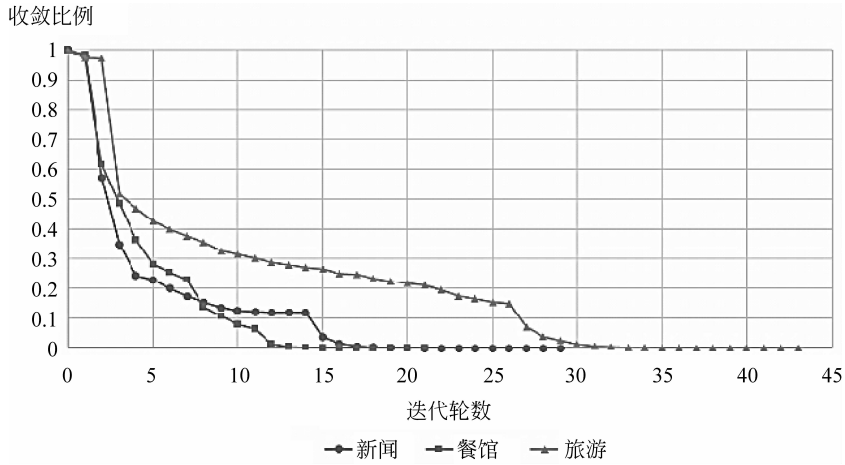

图 5　迭代过程中收敛示意图

3.3.2　Pattern 排序结果展示

图 6 给出了三个数据集排序后前 20 个 pattern 的对比。从图中可以看出，两种用户评论语料（餐馆、旅游）经过排序后，前 20 个 pattern 的排序结果非常相似。两种用户评论的语料展现了很强的相关性，它们之间相同的 pattern 在图中用线标示出来。在前 20 个 pattern 中，有 13 个相同。由于新闻语料的风格和前两种用户评论的语料不一致，相同的 pattern 只有“的＋”、“不－”和“是－”三个，在图中用方框圈出。

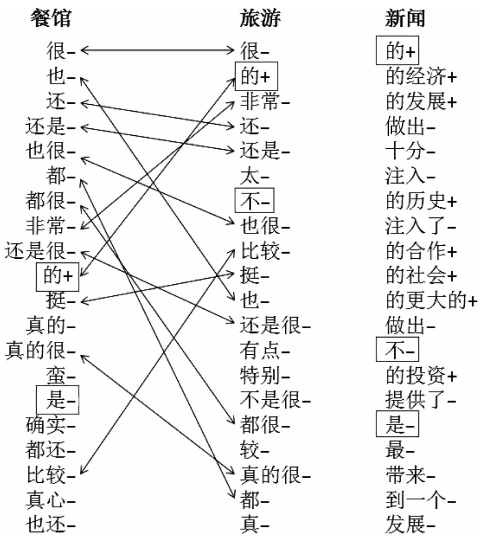

图 6　pattern 最终收敛后的前 20 个结果

3.3.3　抽取结果展示

从表 7 抽取出的情感表达组合中，分别展示了三个数据集上的对象样例各 20 个。描述这些对象的观点词修饰正确，将对象的主要特点都成功体现出来。当然，其中也存在些许错误，比如“洪水 高”。一般“高”形容水位，形容洪水的量词以“大”为主。“洪水 高”排名高的原因是语料中出现多次“洪水水位高”，从而导致水位被作为分数较高的 pattern 使得“洪水 高”这个 pair 获得了较高的分数。

表 7　情感表达组合抽取样例

旅游—“东方明珠”	餐馆—“川菜”	新闻—“洪水”
东方明珠　高	川菜　不错	洪水　大
东方明珠　好	川菜　一般	洪水　罕见

续表

旅游—“东方明珠”	餐馆—“川菜”	新闻—“洪水”
东方明珠 不错	川菜 好	洪水 深
东方明珠 壮观	川菜 有名	洪水 凶猛
东方明珠 著名	川菜 精致	洪水 严重
东方明珠 近	川菜 难吃	洪水 湍急
东方明珠 漂亮	川菜 合适	洪水 最大
东方明珠 高大	川菜 够味	洪水 巨大
东方明珠 耀眼	川菜 常见	洪水 混浊
东方明珠 低	川菜 很辣	洪水 汹猛
东方明珠 绚丽	川菜 最爱	洪水 猛烈
东方明珠 很棒	川菜 很好	洪水 恶劣
东方明珠 雄伟	川菜 辣	洪水 特大
东方明珠 有名	川菜 香	洪水 疯狂
东方明珠 灿烂	川菜 不辣	洪水 高
东方明珠 靓丽	川菜 挺辣	洪水 猖獗
东方明珠 绚烂	川菜 贵	洪水 平静
东方明珠 宏伟	川菜 蛮好	洪水 狂暴
东方明珠 繁华	川菜 红	洪水 污浊
东方明珠 很好	川菜 容易上火	洪水 猖狂

4 小结

本文提出了一种从大规模文本语料中自动获取情感知识词典的方法。在本文方法中，我们通过二部图排序算法可以获得较高正确率的二元对，再通过语义之间的约束进一步提取更多的表达组合。该方法的主要特点是：可以基于语料自动生成，不需要人工干预设置种子词或者 pattern 就可以获得正确率很高的情感表达组合。基于对象的情感分析往往依赖大量的外部资源以及人工制定的抽取规则，而本文所提的方法以语料为单位，通过一个完整的大规模语料得到的对象情感词结构关系进行有效的排序以及提炼，并且不需要任何外部资源和人工干预，自动化构建情感知识对。实验结果表明，本文所提方法能有效地获取情感表达组合。本文得到的结果已经放在 Github① 上。

本文方法还可以从多个角度进行改进。首先由于使用的三个语料缺少极性标注，所以抽取的情感表达组合没有标注对应的情感倾向极性，此项内容可以作为下一阶段的工作继续研究。基于对象的情感词分析已经有了大量的有监督方法研究，后续我们也会人工标注相关语料构建有监督系统来改进候选集合获取方法。其次，抽取的实体或属性之间具有一定关系以及情感观点词之间的网络关系，为建立一个专门用于情感分析的情感知识图谱提供了可能。

参考文献

[1] Pang B, Lee L. Opinion mining and sentiment analysis [J]. Foundations and Trends in Information Retrieval, 2008, 2(1-2): 1-135.

[2] Pang B, Lee L, Vaithyanathan S. Thumbs up? Sentiment classification using machine learning techniques [C]//Proceedings of the ACL-02 Conference on Empirical Methods in Natural Language Processing-Volume 10. Association for Computational Linguistics, 2002: 79-86.

[3] 宗成庆. 统计自然语言处理[M]. 北京：清华大学出版社，2008:1-475.

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

[5] Li J, Sun M. Experimental study on sentiment classification of Chinese review using machine learning techniques[C]//Proceedings of International Conference on Natural Language Processing and Knowledge Engineering, 2007. Nlp-Ke. 2007:1-12.

[6] Xu G, Meng X, Wang H. Build Chinese emotion lexicons using a graph-based algorithm and multiple resources. [C]//Proceedings of COLING 2010, International Conference on Computational Linguistics, Proceedings of the Conference, 23-27 August 2010, Beijing, China. 2010:1209-1217.

[7] 刘知远，崔安颀. 大数据智能[J]. 信息安全与通信保密，2016，2：066.

[8] Esuli A, Sebastiani F. SentiWordNet: A publicly available lexical resource for opinion mining[C]//Proceedings of LREC, 2006, 6: 417-422.

[9] Baccianella S, Esuli A, Sebastiani F. SentiWordNet 3.0: An enhanced lexical resource for sentiment analysis and opinion mining [C]//Proceedings of LREC, 2010, 10: 2200-2204.

[10] Turney P D. Thumbs up or thumbs down?: Semantic orientation applied to unsupervised classification of

① https://github.com/rainarch/SentiBridge

reviews[C]//Proceedings of the 40th Annual Meeting on Association for Computational Linguistics. Association for Computational Linguistics, 2002: 417-424.

[11] Banea C, Wiebe J M, Mihalcea R. A bootstrapping method for building subjectivity lexicons for languages with scarce resources[C]//Proceedings of International Conference on Language Resources and Evaluation, Lrec 2008. DBLP, 2009: 2764-2767.

[12] Hatzivassiloglou V, McKeown K R. Predicting the semantic orientation of adjectives[C]//Proceedings of the Eighth Conference on European Chapter of the Association for Computational Linguistics. Association for Computational Linguistics, 1997: 174-181.

[13] Kanayama H, Nasukawa T. Fully automatic lexicon expansion for domainoriented sentiment analysis [C]//Proceedings of Conference on Empirical Methods in Natural Language Processing, 2006:355-363.

[14] K Qiu G, Liu B, Bu J, et al. Expanding domain sentiment lexicon through double propagation[C]//Proceedings of International Jont Conference on Artifical Intelligence. Morgan Kaufmann Publishers Inc. 2009:1199-1204.

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

[16] K 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.

[17] Agathangelou P, Katakis I, Kokkoras F, et al. Mining domain-specific dictionaries of opinion words [C]//Proceedings of Web Information System Engineering, 2014:47-62.

[18] 李智超. 面向互联网评论的情感资源构建及应用研究[D]. 北京：清华大学博士学位论文,2011.

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

[20] 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.

[21] Zhuang L, Jing F, Zhu X Y. Movie review mining and summarization [C]//Proceedings of the 15th ACM International Conference on Information and Knowledge Management. ACM, 2006: 43-50.

[22] Kobayashi N, Inui K, Matsumoto Y, et al. Collecting evaluative expressions for opinion extraction [C]//Proceedings of International Conference on Natural Language Processing. Berlin Heidelberg: Springer, 2004: 596-605.

[23] Somprasertsri G, Lalitrojwong P. Mining feature-opinion in online customer reviews for opinion summarization[J]. J. UCS, 2010, 16(6): 938-955.

[24] 王素格，吴苏红. 基于依存关系的旅游景点评论的特征-观点对抽取[J]. 中文信息学报，2012，26(3)：116-122.

[25] Page L, Brin S, Motwani R, et al. The PageRank citation ranking: Bringing order to the web[J]. Stanford Digital Libraries Working Paper, 1999: 9(1): 1-14.

[26] Zhang R, Zettsu K, Kidawara Y, et al. Context-sensitive web service discovery over the bipartite graph model[J]. Frontiers of Computer Science, 2013, 7(6):875-893.

[27] Mikolov T, Chen K, Corrado G, et al. Efficient estimation of word representations in vector space[J]. arXiv preprint arXiv:1301.3781, 2013.

卢奇(1992—)，硕士研究生，主要研究领域为命名实体识别。
E-mail: luqibhf@qq.com

陈文亮(1977—)，通信作者，博士，教授，主要研究领域为信息抽取和知识图谱。
E-mail:wlchen@suda.edu.cn