ISSN 1004-9037,CODEN SCYCE4
Journal of Data Acquisition and Processing Vol. 32,No. 4,Jul. 2017,pp. 844－852
DOI:10.16337/j.1004-9037.2017.04.024

http://sjcj.nuaa.edu.cn
E-mail:sjcj@nuaa.edu.cn
Tel/Fax:＋86-025-84892742

一种基于图的情感基准词选择方法

方溢君[1]　何炎祥[1]　刘　楠[2]

(1.武汉大学计算机学院,武汉,430072;2.军事经济学院军需系,武汉,430035)

摘　要:作为文本情感分析的前提和基础,词语的情感极性判别显得尤为重要。现有利用情感基准词进行词语的情感倾向研究中,情感基准词的选择多数基于研究者的人工判别或词语的使用频率。以上方式存在着随机性和主观性的缺陷,并且难以保证对词典中语义关系的全面覆盖。本文提出以候选基准词为顶点,两词间的知网相似度作为边的权重设定参数来构建情感词的无向图。将图中结点的中介性值作为基准词的选择依据,从而保证所选基准词的可靠性。实验证明,通过该方法选取出来的基准词在词的情感倾向分类中具有较高的准确率。

关键词:情感基准词;知网相似度;情感词无向图;中介性值

中图分类号:TP391.1　　**文献标志码**:A

Graph-Based Selection Method for Basic Sentimental Lexicons

Fang Yijun[1], He Yanxiang[1], Liu Nan[2]

(1. Computer School, Wuhan University, Wuhan, 430072; China;
2. Department of Quartermaster, Military Economic Academy, Wuhan, 430035, China)

Abstract: As the premise and basis of text sentimental analysis, the emotion polarity discrimination of lexicons is particularly important. Existing methods of select basic sentimental lexicons in the study of semantic tendency are mostly based on artificial discrimination and lexicons frequency. Those ways suffer the defects of randomness and subjectivity. And it is difficult to ensure the full coverage of the semantic relations in the dictionary. In the paper, we present a method that treats the candidate basic sentimental lexicons as the vertex and the HowNet acquaintance as edge weight to build sentimental lexicons undirected graph. The betweeness-centrality value of nodes in the graph is used as the reference of basic lexicons selecting. Thus we can ensure the reliability of the selected basic lexicons. Experiments show our method has a high accuracy in the classification of emotional tendencies.

Key words: basic sentimental lexicons; Hownet acquaintance; sentimental lexicons undirected graph; betweeness-centrality value

引　言

随着大数据时代的来临,海量的网络资源成为计算机领域工作者面临的机遇和挑战。这些网络数

基金项目:国家自然科学基金(61070083,61303115)资助项目。
收稿日期:2014-09-19;**修订日期**:2017-04-09

据中蕴含着丰富的主观情感。如何从中高效、准确地发掘出有用的信息成为人们研究的热点。这一背景之下,文本的情感倾向性分析愈发受到研究者的关注。作为最小粒度的倾向性分析,词汇的情感分析是对句子、段落和篇章等更大粒度的语义单元进行情感倾向性分析的基础。为了表示词汇的语义倾向,通常的做法是将[-1,1]之间的一个实数作为语义倾向值。通过设定的域值,将语义倾向值大于域值的词作为褒义词,反之则作为贬义词。基于情感词典和基于大规模语料库是目前常用的对词汇进行情感极性分析的方法。基于语料库的统计方法主要利用在大规模语料库中挖掘出来的语言学规则或通过机器学习获得的语言模型来对词汇的情感倾向进行判别。基于词典求语义相似度则主要利用语义相似度和语义相关场功能来计算给定词和基准词之间的相似度,从而得到该词的语义倾向。基于企业事实主题诊断研究则在构建情感本体的基础上,利用条件随机场(Conditional random field algorithm,CRF)挖掘文本中的情感词,大大减轻了人工获取方法的工作量。对于在不同语境中表达不同情感的词汇,利用贝叶斯模型对之进行情感消歧,实验结果显示该方法有较好的实用性。Turney 等[1]以成对出现和不依赖于上下文为标准,挑选了 7 对基准词来判定词汇的情感倾向,实验结果表明,基准词的挑选对词汇情感倾向性的判别有重大影响。朱嫣岚等[2]在思路上同样沿用 Turney 的方法,选择强烈褒贬倾向并且具有代表性的词语作为基准词。选择高频词语集合作为候选基准词集合,再用从 Google 搜索返回的高频词语作为褒贬基准词。实验结果表明,基准词对数目增加,判别效果随之提高。王素格等[3]提出了基于类别区分能力与情感词词表相结合的方法,先计算得出语料库中名词、形容词和动词的类别区分能力,从中选出区分能力较强的词,再将得到的词和情感词词表作交集,计算各自在语料中出现的频率,最后选择出现频率高的词作为基准词。陈岳峰等[4]以知网中的概念作为情感倾向分析的最小单元,通过人工的方法及聚类的方法选择基准概念。彭学仕等[5]则提出应用词聚类的思想,从目标领域中选择初始种子词,经过反复扩展、聚类,最终迭代得出最优基准词。

以往的实验结果表明,基准词的选择对词汇情感倾向性的判别及后续的情感倾向的研究有重大影响。现有研究中,基准词多数来自研究者的人工选择,或简单地根据词性、词频等信息进行判断,存在着随机性和主观性的缺陷且难以保证在词典中对语义关系全面覆盖。本文在现有知网语义倾向性计算方法的研究基础上[2,6],提出了一种基于图的情感基准词选取方法,通过计算词汇和正负情感基准词的平均相似度的差值来确定其情感倾向。和已有方法采用相同的情感相似度计算方式,对比验证得出本文方法提高了情感判别的效果,避免了人工选择的主观性,也确保了基准词的准确性和全面性。

1 知网情感关系图构建

1.1 知网

1.1.1 知网相似度

知网(英文名称为 HowNet)是一个以汉语和英语的词语所代表的概念为描述对象,以揭示概念与概念之间以及概念所具有的属性之间的关系为基本内容的语义常识知识库。知网的基础结构包括概念和义原。概念是对词汇语义的一种描述,每一个词语可以由 1 个或多个概念构成。概念在知网中被用一种叫做 DEF(Definition)的知识描述语言来定义。DEF 由义原组成。义原是用于描述一个概念的最小意义单位。在自然语言处理领域中,关于知网的重要成果之一就是刘群等[6]提出了基于 HowNet 的词语语义相似度计算方法。对两个汉语词语 W_1 和 W_2,如果 W_1 有 n 个概念 $S_1,S_2,\cdots,S_n$,W_2 有 m 个概念:$S_1,S_2,\cdots,S_m$,则 W_1 和 W_2 的相似度为各个概念相似度的最大值,即为:$\mathrm{Sim}(W_1,W_2)=\max\limits_{i=1,\cdots,n;j=1,\cdots,m}\mathrm{Sim}(S_{1i},S_{2j})$,而对于概念 S_1,S_2,它们的相似度可表示为

$$\mathrm{Sim}(S_1,S_2)=\sum_{1}^{4}\beta_i\mathrm{Sim}_i(S_1,S_2) \tag{1}$$

式中:β_i,$1\leqslant i\leqslant 4$ 为可调节的参数,且有 $\beta_1+\beta_2+\beta_1+\beta_4=1$,$\beta_1\geqslant\beta_2\geqslant\beta_3\geqslant\beta_4$;$\mathrm{Sim}_1(S_i,S_j)$,$\mathrm{Sim}_2(S_i,S_j)$,

$Sim_3(S_i,S_j)$,$Sim_4(S_i,S_j)$分别为第一独立义原描述式、其他独立义原描述式、关系义原描述式和符号义原描述式。

1.1.2 知网相似度应用

在传统的基于知网的词语相似度的计算方法基础上,相关研究人员也提出了一些改进的计算方法,并将之应用在词义消岐[7]、数据挖掘、文本分类和信息检索等领域。在进行相似度计算方法改进的研究[8]中,作者详细分析了传统方法的不足之处,并将知网知识和信息量相结合来改善这类缺陷,提出了一种区分度较高的义原相似度计算方法,同时对集合相似度计算和概念相似度计算提出了优化。为了构建一致性的测度平台,文献[9]在进行主题判断研究时先确定了活动事实的主题和特征词,再利用知网语义相似度进行一致性的测量,此方法对于特定领域的一致性测度起到了很好的效果。

1.2 情感关系图构建

1.2.1 图的应用

以往的研究方法大多将重点放在单个词语的词性、情感属性和所属领域等方面。常见的基准词选择方法都是通过词频、互信息和交叉熵[6,10]等方式。而对于情感关系方面则少有体现,这对于所选基准词的准确性及情感覆盖率均有一定的影响。当前对于大数据的处理以及社交网络的研究[11]已经成为自然语言处理领域的重要话题,同时社交网络数据的复杂性给数据挖掘的应用带来了很大的困难。但在这种复杂的关系中,数据之间也有各种紧密的联系,这让图的表达方式成为数据处理领域的重要数据结构。图的各种特点及优势,也使得它能够比坐标向量表示更多的空间。例如在化学信息学研究领域[12]中,用应用图来表示物体,图的顶点表示物体的各个组成部分,图的边表示物体各组成部分之间的关系。在基于图数据挖掘的研究[13]中,作者针对有向图提出了层级度和连结度的特征概念,并设计了一款可以直接并有效操作有向图频繁模式查询的算法,将之应用到频繁子树的识别方法中。同时在大数据环境下针对图数据的高效处理[14-16]也已经成为当前研究急需解决的问题。基于图的以上特点及其在自然语言处理领域的应用,本文选择用图结构来对情感词及其之间的关系进行表示,并通过将情感词的重要性映射到图的相关概念中来实现对情感基准词的选取。

1.2.2 顶点选择

本文旨在将情感词作为图中的顶点来进行情感基准词的选择。所需的基准词来自正情感基准词集和负情感基准词集,因此本文选择了知网义原描述中,属性标注了“良”或“莠”的词语分别作为候选正负情感基准词。其中的词语情感色彩分别如“好”、“称赞”,“凶”、“沉闷”等。相对而言,这些词有比较明确的正负情感含义,是作为候选情感基准词的重要决定因素。候选正负基准词集中词汇的情感极性值分别为 1 和 −1,将它们作为正负知网情感关系图 $G=(V,E)$的顶点。

1.2.3 边的权重设定

在数据结构图的定义中,点与点之间边的权重表示了两点之间的距离,对应到本文的情感关系图中,词与词之间的边的权重刻画了其联系的紧密程度,可通过知网相似度值的大小来进行具体化。通过描述知网词汇间的相似度的概念来说明词语间密切程度的计算方式,这是一种计算图中对应两个顶点间距离的重要参数。为此,对于情感关系图 G 中任意两个顶点 W_i,W_j 结合可计算得到的知网相似度的值 $\mathrm{Sim}(W_i,W_j)$,分别以下面的 3 种方法设定边的权重,并结合实验结果选择最佳的权重计量方式:(1)令两点间边的权重 $W_{ij}=\mathrm{Sim}(W_i,W_j)$,则知网相似度越大的两个词语,对应到图中的两点间的距离越远。(2)令两点间边的权重 $W_{ij}=1/\mathrm{Sim}(W_i,W_j)$,则知网相似度值越大的两个词语,对应到图中的两点间的距离越近,且 $\mathrm{Sim}(W_i,W_j)$值的改变会引起权重值 W_{ij} 的加速变化。(3)令两点间边的权重 $W_{ij}=1-\mathrm{Sim}(W_i,W_j)$,则知网相似度值越大的两个词语,也使得对应到图中的两点间的距离越近,且线性变化的 $\mathrm{Sim}(W_i,W_j)$值会引起权重值 W_{ij} 的线性变化。

在本文的实验 1 部分,分别通过以上 3 种方法设置边的权重,通过实验结果的准确率来确定 W_{ij} 的最佳选择方式。

1.2.4 关系图构造

根据顶点选择的要求与边的权重设定的定义,本文在选择情感基准词的实验 2 部分从知网中各抽取了一定数量的正负情感词分别作为正情感关系图和负情感关系图的顶点,任意两个结点之间边的权重由知网相似度得出。下面以 A、B、C 及集合 S 表示本文中的顶点集,具体说明如何构造关系图。图 1 中的 A、B、C 分别表示正情感词,S 表示候选情感词集中其他的正情感词的集合。分别将这些词语作为图的顶点,并计算彼此之间的距离。通过语义相似度的计算方法可以计算出词 A,B,C 之间的知网相似度 $\mathrm{Sim}(A,B)$,$\mathrm{Sim}(A,C)$,$\mathrm{Sim}(B,C)$,并由 1.2.3 节中 W_{ij} 的计算方法得出两点间边的权重 W_{AB},W_{AC},W_{BC},亦即图中顶点 A、B,A、C 及 B、C 之间边的距离。同理可以算出顶点 A,B,C 与集合 S 中任意一点之间边的权重,以及集合 S 之间任意两点间的边的权重,也就得到了整个图中所有顶点两两之间的距离。

2 情感基准词的选择

2.1 图中的中心性

在图论和网络分析中,边代表着结点之间的关系及关系的紧密程度,结点的重要性可以通过结点的中心性[17]来衡量。在基于图的研究中,主要有 3 种中心性的应用范围比较广,分别是度中心性(Degree centrality)、接近中心性(Closeness centrality)以及中介中心性(Betweenness centrality)。在度中心性的概念中,中心结点指那些拥有与其他结点的链接数目最多、最活跃的结点。假设网络中的结点总数为 n,则在无向图中,结点 i 的中心性的值就是该结点的度。在有向图中,结点的重要性取决于它的出度,经过归一化处理之后可以表示为

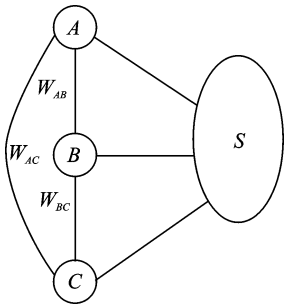

图 1 关系图的构造

Fig. 1 Construction of the relation graph

$$C_D(i)=\frac{d_o(i)}{n-1} \tag{2}$$

式中:$d_o(i)$表示结点 i 的出度。该项指标刻画了结点的活动频繁程度,某一结点的直接连接最多,则可以认为它在网络群中的地位比较突出。一般情况下,某个结点和网络中的其他结点有着更多的联系,则可认为该结点在网络中的地位比较重要,但从另一方面来说,关联的结点越多并不意味着连接了更多的网络范围。所以度中心性还不能全部定义一个结点的重要性。在接近中心性的概念中,接近度或者距离是重要性的决定因素。如果一个结点到其他结点的距离越短,则该结点与其他结点的互动就更加容易,可以认为它在图中的地位比较重要。通过计算最短距离可以得到接近中心性的值:假设在结点总数为 n 的网络中,结点 i 和结点 j 之间的最短距离为 $d(i,j)$,则在无向图中,参与者 i 的接近中心性被定义为

$$C_C(i)=\frac{n-1}{\sum_{j=1}^{n}d(i,j)} \tag{3}$$

在有向图中计算距离时,考虑链接和边的方向即可得出计算公式。图中的某个结点如果处于其他结点间相互联系的路径之中,则该结点可能对其他结点有一定的控制能力。中介中心性(简称中介性)用来度量图中结点对于其他结点的控制能力。如果 i 处在非常多结点的交互路径上,那么 i 就是一个重要的参与者。结合 2.2.2 节中的中介性计算公式可以看出,中介性 C_B 和接近中心性 C_C 之间的差异。C_C 只是做了总体距离的平均,仍然是一种距离;而 C_B 则做了一种比率,比率刻画成一种效率、一种性价比。也就是说 C_C 选出来的重要性结点是由能传播的小范围的大小来刻画的,而 C_B 则考虑一种为了更

远地传播到更远更大的网络中的性价比最高的重要性结点。中介性扮演着“桥”的作用，使原本无关系的结点产生联结。在网络和图中，这样的结点所处的位置十分特殊。对于本文而言，基于 C_B 值选择基准词一方面保证了该类情感词是对其他情感词具有较强控制能力的关键结点，另一方面也可以在最短路径的计算中尽可能多地利用任意两个结点之间的相似度的值，避免个别误差数据带来的影响。

2.2 结点中介性

2.2.1 结点中介性应用

结点中介性在复杂网络及社会学领域进行社交网络分析和路由选择等方面都发挥着重要的作用。在提高复杂网络容量的方法研究中[17]，引入中介性对网络拓扑进行优化和拥塞预测，可以有效平衡中枢结点的负载、缓解拥塞状况和提高网络容量。将有向图的边的中介性分析引入道路网络分析[18]，通过对几个典型城市道路网络进行中介性分析，发现大部分高等级的道路具有高的 C_B 值，研究证实了在城市道路网络中，数学意义上 C_B 值度量层级性与道路所属等级社会意义层级性的相关性。在社会管理及市场营销学中[19]，利用中介性概念可以研究该网络的组成结构，了解每个参与者的主要职责、所做贡献以及影响力等。从而更加针对性地进行人员管理、个性化推送等服务。在对微博影响力个体发现[20]方面，研究者在传统中介性的计算方法上提出一种基于随机游走的中介性的算法，使得该算法不仅能有效地应对海量的微博网络数据，且使得发现结果也明显优于相关的研究。

2.2.2 结点中介性计算

在图 $G=(V,E)$ 中，结点 $v\in V$ 的中介性计算步骤如下：(1)对任一顶点对 (s,v)，$s\in V$，$v\in V$；计算两点之间的最短路径。(2)记录顶点对 (s,v) 的最短路径中，包含的其他顶点。(3)重复步骤(1)和(2)，记录所有顶点对之间的最短路径和包含的其他顶点。(4)计算顶点中介性值为

$$C_B(v)=\sum_{s\neq v\neq t\in V}\delta st(v) \tag{4}$$

式中：$\delta_{st}(v)=\dfrac{\sigma_{st}(v)}{\sigma_{st}}$；$V$ 为所有顶点的集合；σ_{st} 为顶点 s 和顶点 t 之间最短路径的条数；$\sigma_{st}(v)$ 为顶点 s 和顶点 t 之间经过点 v 的最短路径条数。如图 2 所示的有向图中，任意结点的中介性 C_B 值可通过以下方式进行计算：(1) 由图 2 可得，各顶点之间的距离 S 分别为 $S_{AB}=1$，$S_{BC}=5$，$S_{CD}=3$，$S_{DE}=2$，$S_{EA}=4$，$S_{BE}=8$。(2) 通过计算可得任意两个顶点 s，v 间的最短路径 P_{sv} 分别为 $P_{AB}:AB$，$P_{AC}:ABC$，$P_{AD}:AED$，$P_{AE}:AE$，$P_{BC}:BC$，$P_{BD}:BAED$，$P_{BE}:BAE$，$P_{CD}:CD$，$P_{CE}:CDE$，$P_{DE}:DE$。(3) 由中介性计算步骤(4)可得各顶点的中介性值分别为 $C_B(A)=1/1+1/1=2$；$C_B(B)=1/1=1$；$C_B(C)=0$；$C_B(D)=1/1=1$；$C_B(E)=1/1+1/1=2$。

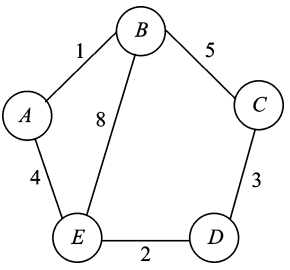

图 2 中介性值的计算

Fig. 2 Value of betweenness centrality

2.3 基准词选择

类似于社交网络中的图的定义，词汇的中介性 C_B 值越大，则表示该词汇出现在越多的词汇间最短路径中，对网络连接的关键作用越明显，符合对基准词的要求。在本文的实验 2 部分，计算候选情感基准词的 C_B 值，选择值大小排名靠前的情感词作为情感基准词。

3 实验及结果分析

3.1 实验环境

本文实验部分主要包括：(1)通过知网相似度的概念计算词语间相似度的值，(2)在情感关系图中，计算每个结点的中介性值 C_B。其中知网相似度的计算通过引入 WordSimilarity 数据包，在 Eclipse 中由 Java 语言完成。中介性值的计算通过引入复杂网络编程包 NetworkX，用 Python 语言完成。

3.2 实验设置

实验内容包括 3 个部分,分别用来进行情感关系图边的权重设置、情感基准词的选择以及本文方法有效性的对比。在知网的概念描述中,属性标明有“良”和“莠”的词有比较明确的正负语义倾向性。同时,有些词语拥有多个概念描述,不同的概念在不同的语境下可能表现出相反语义情感,例如词语“好看”在语句“这件衣服真好看”和“要你好看”中,分别表示正面和负面的语义倾向。为使文章所提方法的效果更准确,将那些概念描述中同时包含“良”“莠”属性的词语从候选基准词中去除,得到正、负候选情感基准词集中的词个数分别为 3 132 和 3 260,并将这些词作为实验 1 和实验 2 中的情感关系图的顶点。

实验 1 边的权重设置

该实验部分用于对 1.2.3 节提出的 3 种权重设置方法进行效果对比,从而选择最佳的权重设置参数。这部分的实验,在不同的权重设置方式下,各选择了 40 对情感基准词对测试集 1 中的情感词进行情感极性判断,其中基准词的选择依据及图的构建方式参考文中 1.2.4 节。实验 1 结果如图 3 所示。

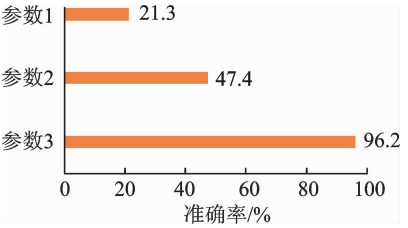

图 3 边的权重不同时的准确率

Fig. 3 Accuracy by different edge weight

实验 2 基准词选择方法的实验对比

朱嫣岚[2]提出的基准词选择方法是在人工挑选正负情感含义较明确的词语基础上,根据词频对词语进行降序排列,分别从正负情感词中挑选出现频率最高的词作为基准词,本文的实验部分将这种选择基准词的方法作为对比方法 1。

应用聚类的思想,选择聚类中心词[5-6]作为基准词是在方法 1 的基础上进行改进的方法。本文将通过 K-medoids 算法对候选基准词进行聚类作为基准词选择的对比方法 2。本文实验将知网概念中的“良”、“莠”属性作为测试集的选取依据。其中,测试集 1 分别选择含有“良”和“莠”属性的词对正情感基准词和负情感基准词进行检验,褒义词和贬义词的数量分别为 3 227 个和 3 355 个。为了使实验效果更具说明性,将测试集 1 中同时包含“良”、“莠”属性的词汇去除,剩下词语的正负情感含义更加明确,将这类词作为测试集 2。测试集 2 中正情感词为 3 132 个,负情感词为 3 260 个。知网中有些词虽然并没有直接包含“良”、“莠”属性,但也有明确的正负情感倾向性。例如词语“伤悲”和“兴高采烈”等在知网中属性并不包含“良”“莠”,但分别有着明显的正负情感,且在基于词典的情感分析中,这类词依然作为选择情感词的一部分。为了说明所选基准词对所有具有正负情感倾向词汇判断的有效性,设置测试集 3。其中共有 1 243 个不包含“良”、“莠”属性的情感词。测试集中的词通过计算其和褒贬义基准词间的平均相似度的差值来确定其语义倾向。本文实验中,以″0″为域值,计算结果大于零的判为正情感词,小于零的判为负情感词,以准确率作为评价标准 。实验 3 结果中的 P'_1,P'_2 分别表示对比方法 2 对正情感词和负情感词极性判断的准确率;P_1 和 P_2 分别表示本文方法对正情感词和负情感词极性判断的准确率。

实验 3 基准词数量的实验对比

本实验中,分别设置基准词对数为 20,40,60 和 80,以此判断基准词数量对情感极性判断的影响。

3.3 实验结果

通过本文方法选择出的 40 对正负情感基准词分别如表 1,2 所示。对不同测试集的实验结果分别如图 4~6 所示。实验 3 结果如图 7 所示。

表 1 本文方法选择出的 40 个正情感基准词

Tab. 1 Complimentary sense lexicons selected by proposed method

交口称赞	交口称赞	俏皮	保准	厚	可观	嘉	好	好吃	婉	动听	平	快	很	扎实	明	柔和			
机动	清明	灵	炯炯有神	热	爽快	牢靠	特许	狂喜	盛	细	肯定	脆	自尊	自愿	舒缓	艺术			
花	讲究	负责	适	舒畅	风发	飘洒													

表 2 本文方法选择出的 40 个负情感基准词

Tab. 2 Derogatory sense lexicons selected by proposed method

不像话	乱	冷淡	凶	勉强	反	口误	可怜	奸	妖	官僚	官僚主义	强暴	患	惨	拐子	拖泥带水	松松垮垮	死心眼儿	沉闷
油嘴	淡	滑头	窄	脏	苦水	虚浮	蛮	蠢虫	误	豪强	贼	逆	险险恶	难	马大哈	黑	鬼		

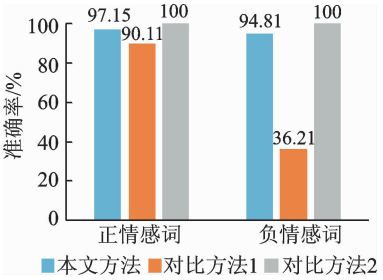

图 4 实验 2 测试集 1 结果

Fig. 4 Result of test set 1 in experiment 2

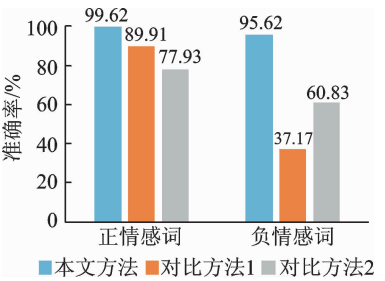

图 5 实验 2 测试集 2 结果

Fig. 5 Result of test set 2 in experiment 2

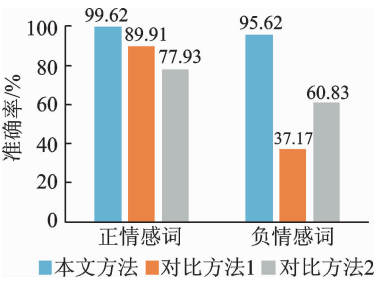

图 6 实验 2 测试集 3 结果

Fig. 6 Result of test set 3 in experiment 2

3.4 结果分析

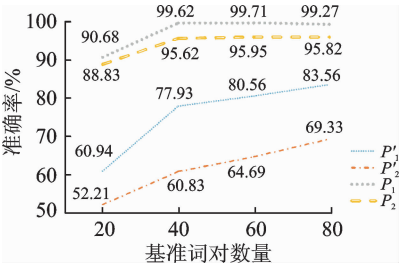

图 7 实验 3 结果

Fig. 7 Result of experiment 3

由图 3 可见，边的权重参数对于情感词情感倾向判断的影响较大。将边的权重参数 W_{ij} 分别设置为 $\mathrm{Sim}(W_i,W_j)$，$1/\mathrm{Sim}(W_i,W_j)$ 和 $1-\mathrm{Sim}(W_i,W_j)$ 分别作为实验 1 的参数 1，参数 2 和参数 3，对于所有情感词整体判断的准确率分别为 21.3%，47.4% 和 96.2%。参数 1 中，两个词之间的知网相似度的值越大，则在图中两顶点的距离越远。在计算 C_B 值时，使得中介性越突出的点出现在最短路径中的次数越少。这与所需选择基准词的要求不符合，因此选择出的基准词准确性差，对于基于基准词计算情感词极性的准确率也较低。参数 2 中，两个词之间的知网相似度值越大，则在图中两顶点的距离越近，且相似度值较小的变化也会使得顶点间距离产生较大的改变。顶点间的距离不能正确地表示信息流通的代价。因此参数 2 表达的有效性虽然较参数 1 有所改善，但仍是一种误差较大的形式。参数 3 中两个词之间的知网相似度值越大，则在图中两顶点的距离越近。对应到社交网络的图中，则可认为，$\mathrm{Sim}(W_i,W_j)$ 值越大的两个结点，其相互间的路径长度越短，信息流动的代价越小。基于上述理念，本文以 $W_{ij}=1-\mathrm{Sim}(W_i,W_j)$ 作为图 $G=(V,E)$ 中结点 W_i,W_j 之间边的权重。两结点联系越紧密，即越相似，则两点间距离 W_{ij} 越小。在求最短路径时，能够尽可能地保证该路径上包含更多语义相似度更大的结点，这点也符合本文对基准词的选择标准。从图 4 可以看出，对比方法 1 所选基准词对正情感词判断有较高的准确率，但对负情感词判断的准确率很低，整体结果不理想。因此仅依靠词频作为基准词的判断依据缺乏完善的科学依据，不能保证语义的覆盖率，导致对不同极性情感词的判断失衡严重。从图 5 可以看到，在经过筛选的测试集 2 中，该方法对贬义词的判断效果有极小的提升，但仍不理想。图 4，5 的结果显示，本文方法对正负情感词的判断都有很高的准确率，在经过筛选的测试集 2 上，实验效果有了更近一步的提升，最高达到了 99% 的准确率。说明本文方法所选基准词有较高的语义覆盖率，克服了传统方法主观性和随机性的缺陷。实验 2 的测试集 3，由于其中类似于“回头”“重”之类的词并没有很明显的正负情感倾向，给利用基准词判断正负倾向带来一定的干扰，使得最终的准确率较测试集

1 和测试集 2 有所降低,但本文方法较之对比方法 1 仍有一定的改善。对测试集中有较明显情感倾向的词语,如“喜爱”、“尊重”和“责备”等,本文方法和对比方法 1 所计算最终情感值分别为 0.056,0.078 和−0.0192;−0.006,−0.007 和 0.029。由此可见,本文方法对知网中具有情感倾向的词具有通用性,一定程度上显示了方法的优越性。实验 3 中,分别设置基准词对数为 20,40,60 和 80,以此判断基准词数量对情感极性判断的影响。从图 7 可以看出,总体而言,本文方法对情感词极性判断的准确率随着基准词数量的增加而提升。基准词数量为 20 时,由于覆盖的语义范围不够全面,使得情感词极性的判断产生了一定的误差,准确率较实验 2 的测试集 1 有所降低。由于当基准词数量为 40 时,已取得很高的准确率,设置其为 60 和 80 时,改善的空间有限,但整体维持在一个较高的水平。由此可以看出,当数据量较大时,使用文中方法选择一定规模的基准词即可达到较好的语义覆盖率,对情感词极性判断取得较好的效果,这在一定程度上减轻了大数据环境下的计算量,有一定的使用价值。同样在实验 3 中,当随着基准词数目的增多,对比方法 2 情感极性判断的准确性有着明显的提升,基准词的数量对极性判断的准确率有较大的影响。准确率从基准词数量为 20 时的 64.94%提升到基准词数量为 80 时的 83.56%,情感基准词对整个情感语义的覆盖率会随其数量的增加而显著提高,并且个别误差带来的负面影响也会得到减弱。但当取相同数量的基准词时,本文方法无论在整体的准确率还是在对褒贬义词判断的均衡性方面皆有较大的优越性。文中方法对基准词数量的敏感性较小,在实际应用中更具实用性。

4 结束语

词语的正负情感倾向性研究是文本情感分析的基础,本文尝试通过改善选取方法来获得更具代表性的情感基准词。文章借鉴了前人有代表性的研究和创新点,结合完善的图理论,提出了基于图的情感基准词选取方法。实验结果证明,本文所提方法较之传统方法有较明显的优势,所选情感基准词有较好的语义覆盖率和普遍适应性。

参考文献:

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

[2] 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.
Zhu Yanlan, Min Jin, Zhou Yaqian, et al. Semantic orientation computing based on HowNet[J]. Journal of Chinese Information Processing, 2006, 20(1): 14-20.

[3] 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(9):167-170.
Wang Suge, Li Deyu, Wei Yingjie. A synonyms based word sentiment orientation discriminating[J]. Journal of Chinese Information Processing, 2009, 23(9): 167-170.

[4] 陈岳峰,苗夺谦,李文,等.基于概念的词汇情感倾向识别方法[J].智能系统学报, 2011,6(6):489-493.
Chen Yuefeng, Miao Duoqian, Li Wen, et al. Semantic orientation computing based on concepts[J]. CAAI Transactions on Intelligent Systems, 2011, 6(6):489-493.

[5] 彭学仕, 孙春华. 面向倾向性分析的基于词聚类的基准词选择方法[J]. 计算机应用研究, 2011, 28(1):114-116.
Peng Xueshi, Sun Chunhua. Paradigm words selecting method based onword clustering for sentiments analysis[J]. Application Research of Computers, 2011, 28(1):114-116.

[6] 刘群,李素建.基于《知网》的词汇语义相似度计算[J]. 中文计算语言学,2002,7(2): 59-76.
Liu Qun, Li Sujian. Word similarity computing based on HowNet[J]. Computational Linguistics & Chinese Language Processing ,2002, 7(2):59-76.

[7] 李辉, 张琦, 卢湖川,等. 基于知网的中文常问问答系统[J]. 计算机工程, 2008,34(23):62-64.
Li Hui, Zhang Qi, Lu Huchuan, et al. Chinese frequency asked questions based on HowNet[J]. Computer Engineering, 2008,34(23):62-64.

[8] 郭勇. 基于《知网》的词语相似度计算研究及应用[D]. 长沙:湖南大学,2012.
Guo Yong. The research of HowNet based word similarity computation and its application[D]. Changsha: Hunan University,2012.

[9] 马续补，郭菊娥. 基于《知网》语义相似度的企业事实主题诊断研究[J]. 情报杂志，2010，29(5)：54-57.
Ma Xubu，Guo Jue. Study of theme's diagnosis based on HowNet word similarity computing[J]. Journal of Intelligence，2010，29(5)：54-57.

[10] 闻扬，苑春法，黄昌宁. 基于搭配对的汉语形容词-名词聚类[J]. 中文信息学报，2000，14(6)：45-50.
Wen Yang，Yuan Chuanfa，Huang Changning. Clustering of Chinese adjectives-nouns based on compositional pairs[J]. Journal of Chinese Information Processing，2000，14(6)：45-50.

[10] 李桃陶，周斌，王忠振. 基于社交网络的图数据挖掘应用研究[J]. 计算机技术与发展，2014(10)：6-11.
Li Taotao，Zhou Bin，Wang Zhongzhen. Research on graph data mining application based on social network[J]. Computer Technology and Development，2014(10)：6-11.

[11] 赵海峰. 基于图的模式识别及其在计算机视觉中的应用[D]. 南京：南京理工大学，2011.
Zhao Haifeng. Graph-based pattern recognition and its applications in computer vision[D]. Nanjing：Nanjing University of Science & Technology，2011.

[12] 周溜溜. 基于图结构的数据挖掘研究及应用[D]. 南京：南京林业大学，2013.
Zhou Liuliu. Research and application of data mining based on graph structure [D]. Nanjing：Nanjing Forestry University，2013.

[13] 罗征，王赛，张帆，等. 面向大数据的图数据处理技术[J]. 情报工程，2015，1(6)：120-125.
Luo Zheng，Wang Sai，Zhang Fan，et al. Graph data processing on big data[J]. Technology Intelligence Engineering，2015，1(6)：120-125.

[14] 丁悦，张阳，李战怀，等. 图数据挖掘技术的研究与进展[J]. 计算机应用，2012，32(1)：182-190.
Ding Yue，Zhang Yang，Li Zhanhuai，et al. Research and advances on graph data mining[J]. Journal of Computer Applications，2012，32(1)：182-190.

[15] 闫朋，高建瓴. 图数据挖掘在社交网络的应用研究[J]. 电子世界，2016(8)：53-55.
Yan Peng，Gao Jianling. Research on application of graph data mining in social networks[J]. Electronics World，2016(8)：53-55.

[16] Opsahl T，Agneessens F，Skvoretz J. Node centrality in weighted networks：Generalizing degree and shortest paths [J]. Social Networks，2010，32(3)：245-251.

[17] 范晶，秦卓琼，张国清. 基于中介中心性提高复杂网络容量的方法[J]. 计算机仿真，2008，25(3)：167-170.
Fan Jing，Qin Zhuoqiong，Zhang Guoqing. A method for improving complex network capacity based on betweenness centrality[J]. Computer Simulation，2008，25(3)：167-170.

[18] 李清泉，曾喆，杨必胜，等. 城市道路网络的中介中心性分析[J]. 武汉大学学报(信息科学版)，2010，35(1)：37-41.
Li Qingquan，Zeng Zhe，Yang Bisheng，et al. Betweenness centrality analysis for urban road networks[J]. Geomatics and Information Science of Wuhan University，2010，35(1)：37-41.

[19] 杨学成，张晓航，等. 社会网络分析在市场营销学中的应用[J]. 当代经济管理，2009，31(6)：25-29.
Yang Xuecheng，Zhang Xiaohang. The application of social network analysis to marketing research[J]. Contemporary Economy & Management，2009，31(6)：25-29.

[20] 朱静宜. 基于中介中心度的微博影响力个体发现[J]. 计算机应用研究，2014，31(1)：131-133.
Zhu Jingyi. Centrality based micro-blog influence entity discovery[J]. Application Research of Computers，2014，31(1)：131-133.

作者简介：

方溢君(1990-)，男，硕士研究生，研究方向：自然语言处理，E-mail：1025759496@qq.com。

何炎祥(1952-)，通信作者，男，教授，研究方向：自然语言处理，可信编译，E-mail：1025759496@qq.com。

刘楠(1983-)，男，博士，研究方向：自然语言处理。