基于情感 Ontology 的资源分析模型*

刘 倩 陶县俊 王晓东
(河南师范大学计算机与信息技术学院 新乡 453007)

摘 要 对资源分析方法进行了研究,并提出了一种基于情感 Ontology 的分析方法。首先基于"知网"构建情感 Ontology,然后基于情感 Ontology 抽取资源分析的特征词汇并判断其情感倾向性,最后根据抽取的特征词汇对整篇文本的情感倾向进行分析。实验结果表明,在以人工标注做 Baseline 的基础上,利用情感 Ontology 抽取特征词汇的资源分析方法可以使情感识别的准确率达到 78.87%。

关键词 Ontology 文本倾向性分析 "知网" 词汇相似度
中图分类号 TP391.1

Resource Analysis Model Based on Sentiment Ontology

Liu Qian Tao Xianjun Wang Xiaodong
(Department of Computer Science, Henan Normal University, Xinxiang 453007)

Abstract Methods of resource analysis were studied and present an approach based on Sentiment Ontology. First Sentiment Ontology is constructed with the sentiment words that HowNet has tagged, and then select the features from an undetermined document and estimate their orientation, at last analysis the orientation of the entire document. The experimental results show that the sentiment identification accuracy reached to 78.87% using the features extracted according to Sentiment Ontology compared with the Baseline that identifying the sentiment orientation all by manual work.

Key words Ontology, text orientation analysis, hownet, vocabulary similarity
Class Number TP391.1

1 引言

资源分析的主要技术是文本倾向性分析,该技术在商业产品评论、网络舆情及垃圾邮件过滤等领域中有着广泛的应用,通过判别文本的正负倾向可以指导用户购买某种产品、监控网络舆情等。

常用的文本倾向性分析方法有基于统计的文本倾向性分析方法,其思想是先通过人工对一些文档进行倾向性标注,并将这些文档作为训练集,再通过统计学习的方法构造一个褒贬两类分类器,最后使用构造好的褒贬分类器对待估文档进行分类,即识别待估文档的倾向性。Pang 等人[1]分别使用朴素贝叶斯(Native Bayes)、最大熵(Maximum Entropy)及支持向量机(Support Vector Machines)方法进行文本倾向性研究,并对三种方法作了比较分析,发现 SVM 方法的准确率能够达到约 80%,是这三种方法中最好的方法。基于统计的文本倾向性分析方法在国内也得到较为广泛的研究,例如清华大学夏云庆等人[2]利用 SVM 统计学习方法开发了基于商品的意见挖掘系统,其精确率能达到约 85%。

另一种分析方法是基于语义规则的文本倾向性研究方法,其主要思想是在对能够表达文档倾向性的词汇进行倾向性计算之后,利用语言学中的语义规则对这些词汇的极性进行调整以期望在文本倾向性分析中得到更高的准确率。Hatzivassilo-

* **收稿日期**:2009 年 4 月 10 日,**修回日期**:2009 年 6 月 5 日
基金项目:河南省科技攻关计划项目(编号:082102210007)资助。
作者简介:刘倩,女,硕士研究生,研究方向:语义网络与 ontology。王晓东,男,博士,教授,研究方向:语义网络与 ontology,教育信息技术。

glou 等人[3]使用连接形容词的连词的语言学约束来判断所连接的两个形容词表达的感情是否一致,然后用类聚方法来获得表示情感倾向的两个形容词类。Turney 等人[4]使用 PMI_IR 方法来估计短语与表示情感的两个立场的基准词(如"好"与"坏")的相似度,相似度计算用逐点互信息。

利用上述的方法进行文本倾向性分析已经取得了相对较好的实验结果,但是这些方法在抽取文本特征词汇的时候并没有充分考虑词汇以及词汇之间的语义信息,难以提高文本倾向性分析的效果。本文基于情感 Ontology 研究文本倾向性,在充分考虑词汇以及词汇之间的语义信息的基础上对文本的倾向性进行分析,实验结果表明该方法能够使文本倾向性分析的准确率得到明显的提高。

2 情感 Ontology 的构建

构建情感 Ontology 是为了更充分地表达情感词汇之间所蕴含的语义信息,如词汇的情感倾向性以及词汇间的相似、递进和转折关系等,从而为文本的倾向性分析提供有效的分析依据。我们所构建的情感 Ontology 是以"知网"中的情感词汇和"知网"所提供的词汇相似度计算方法为基础的[5~6]。

2.1 情感 Ontology 的概念选择

情感 Ontology 中的词汇选自"知网"中已经标注过的情感分析用词语集(中文)[7],正面词汇共有 4566 个,包括正面评价词语 3730 个,正面情感词语 836 个;负面词汇共有 4370 个,包括负面评价词语 3116 个,负面情感词语 1254 个。此外还有表示程度级别的词语 219 个。

2.2 情感 Ontology 的构建

情感 Ontology 中包括能够体现正面倾向和负面倾向的两种词汇,可以形象化的把情感 Ontology 看作一个"地球仪"(如图 1 所示)。"北极"和"南极"分别代表正面和负面两种情感倾向,"赤道"以北均为正面倾向词,"赤道"以南均为负面倾向词,越接近"赤道",词汇的倾向性越不明显,即越接近中性。由于情感 Ontology 中不存在中性词,所以"北半球"和"南半球"实际上不是连接在一起的,所谓的"赤道"上实际并没有词汇。

构建情感 Ontology 首先需要从众多情感词汇中抽象出最能代表正面倾向和负面倾向的词汇分别作为"正极"和"负极",然后按照基于"知网"的词汇语义相似度计算方法分别计算所有情感词与"两极"的相似度,根据相似度的排名分别选出与"两极"相似度较高的若干个词汇作为"第一纬线圈"(即"正极"的下位)上的词汇。同一"纬线圈"上词汇的"经度"可以根据两两词汇的相似度排名来确定,相似度越高则"经度"越接近,反之则越远。"第二纬线圈"(即"第一纬线圈"的下位)以及以后各"纬线圈"上的词汇可以利用相同的算法,根据相似度排名从剩余的词汇中分别选取与"第一纬线圈"相同数量的词汇,再分别计算与前一"纬线圈"上词汇的相似度,按照相似度越大则"经度"越接近的原则使每个词都能在情感 Ontology 中得到定位,且任意两个词汇的"经纬度"都不同,词汇的"经纬度"可以作为其在情感 Ontology 中的唯一索引。

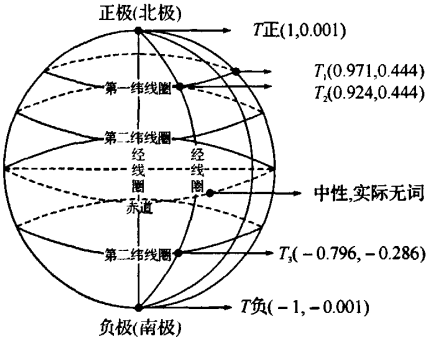

图 1 情感 Ontology 抽象图

分析文本的情感,关注最多的是情感的相似关系、递进关系和转折关系,所以针对情感 Ontology 中词汇的特点,主要考虑了以下三种关系:并列关系、递进关系和转折关系。如果两个词具有并列关系,则表示这两种表达方式在语言文本中相互替代而不改变其意义,因而它们所表达的情感倾向性是一致的;如果两个词具有递进关系,则表示这两种表达方式在语言文本中相互替代会加强或者减弱其意义,相应的它们所表达的情感是逐渐加强的或者逐渐减弱的;如果两个词具有转折关系,则表示这两种表达方式在语言文本中相互替代后意义完全相反,它们所表达的情感倾向性则是相反的。

根据词汇间的关系可以很容易地使词汇在情感 Ontology 中定位,例如,若已知图 1 中 T_2 与 T_1 (0.971,0.444)是同义词即 T_2 和 T_1 具有并列关系,则不必经过计算也可获知 T_2 的"纬度"为 0.444。

另外,表示情感程度级别的词分为 5 个等级,每个等级包含若干词汇。不同等级间词汇具有递进关系,相同等级间词汇具有并列关系。程度级别词汇与情感倾向性词汇之间不建立直接的关系。

2.3 特殊情感(子情感)Ontology

情感 Ontology 可以满足不同人群、不同领域的需求，利用其进行情感分析既可以得到情感的正负倾向，也可以具体到某种特殊情感，即子情感。情感 Ontology 可以分为正、负两个子情感 Ontology，同时，正、负情感 Ontology 又可以进行细分。在“地球仪”上，相邻“经线”上的词汇体现相似的情感倾向，相似度超过某个预设阈值的某几个“经线”上的词汇可以被归为一个子情感 Ontology。例如，负面情感 Ontology 又可以细分为忧郁情感 Ontology、恐惧情感 Ontology 等，图 2 中给出了忧郁情感 Ontology 和恐惧情感 Ontology 中的部分词汇。

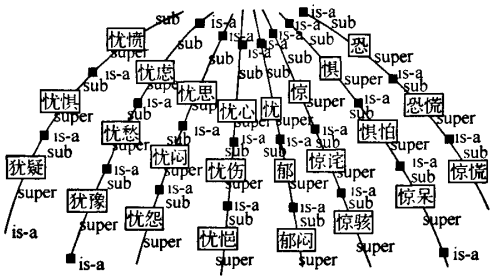

图 2 特殊情感 Ontology 示例

3 词汇倾向性计算

基于情感 Ontology 的语义相似度计算考虑了两个因素：词汇与情感 Ontology 中出现的所有词汇基于知网的语义相似度和词汇在情感 Ontology 中的深度。两者的加权平均和即为词汇基于情感 Ontology 的语义相似度，计算公式如下：

$$Sim=\varepsilon\times Hsim(T_i)+(1-\varepsilon)\times Odis(T_i) \quad (1)$$

其中，$Hsim(T_i)$是词汇与情感 Ontology 中出现的所有词汇基于“知网”的相似度计算，这个体现差异性的值代表了待定词 T_i在情感 Ontology 中的“经度”；$Odis(T_i)$是词汇在情感 Ontology 中的深度，这个体现相似性的值代表了 T_i的“纬度”。ε 是一个可调节的参数，且有 $0\leqslant\varepsilon\leqslant1$。

待定词与正负情感倾向性词的语义相似度是分别计算的，若待定词倾向于正极，则 $Hsim(T_i)$和 $Odis(T_i)$均为正值，相应的 Sim 也为正值；反之，则 $Hsim(T_i)$和 $Odis(T_i)$均取其相反数，相应的 Sim 为负值。$Hsim(T_i)$和 $Odis(T_i)$的具体定义如下：

$$Hsim(T_i)=\max_{j=1\cdots n}(Sim(T_j,T_i)) \quad (2)$$

其中，$1\leqslant j\leqslant n$，n 表示情感 Ontology 中词汇的数量，$Sim(T_j,T_i)$是按照基于“知网”的词汇语义相似度计算方法得到的相似度。

$$Odis(T_i)=\frac{\alpha}{d_i+\alpha} \quad (3)$$

其中，d_i是词汇 T_j在情感 Ontology 中与最高层节点的语义距离，是一个正整数。α 是一个可调节的参数。

4 资源分析模型构建

本文所分析的资源主要指文本资源，并采用向量空间模型来表示文档，单文档可以表示成：$D_i=\{(T_{i1},w_{i1}),(T_{i2},w_{i2}),\cdots,(T_{in},w_{in})\}$，其中 $T_{ij}(j=1\cdots n)$表示文档中的词汇，$w_{ij}(j=1\cdots n)$表示词汇 T_j所对应的权重。进行资源分析时，首先利用情感 Ontology 对文档进行分词，然后将文本中出现的所有词汇与情感 Ontology 中的词汇进行相似度计算初步过滤掉不能够体现情感的词汇，以缩小分析规模。若要判断文本中某个词汇的情感倾向，则首先计算这个词汇与情感 Ontology 中所有词汇的相似度，在情感 Ontology 中找到与待定词相似度最大的那个词作为待定词在情感 Ontology 中的映射，由此待定词在情感 Ontology 中的“经纬度”就可以确定，最后通过计算待定词基于情感 Ontology 的语义相似度和词汇在本文资源分析模型中的权重来分别进行词汇情感倾向判断和特殊情感识别。

4.1 词汇特殊情感识别

词汇在资源分析模型中的权重综合考虑了词频因素以及根据情感 Ontology 计算的词汇倾向性值 Sim。词汇权重分为整数和小数两个部分，二者既是一个整体又各有其特定的意义，其中词频 tf 是一个整数，表示词汇在文档中出现的次数，依据特征词的词频可修改其隶属子情感 Ontology 中的计数器的值；词汇倾向性值 Sim 是一个(0,1]之间的小数，表示特征词与子情感 Ontology 的贴近程度，若其值高于某一个阈值 $\lambda(0\leqslant\lambda\leqslant1)$则可判定该特征词属于某个子情感，否则即判断其不属于某个子情感。基于以上分析，可以得到词汇在本文的资源分析模型中的权重计算公式，如下：

$$w_{ij}=tf_{ij}+Sim_{ij} \quad (4)$$

假定情感 Ontology 中含有 n 个子情感，给每个子情感 Ontology 都设定一个阈值，所有归属到该情感中的特征词的倾向性值 Sim_{ij}都要在这个范围之内，因此可以根据特征词的倾向性值所在的范

围来识别其情感。

情感 Ontology 可以表示为 $O=\{O_1,O_2,\cdots,O_n\}$,其中 $O_s(s=1\cdots n)$ 代表了第 s 个子情感,将 Sim_{ij} 的取值范围(0,1]分成 n 个等份,其中属于子情感 $O_s(s=1\cdots n)$ 的特征词的倾向性值 Sim_{ij} 的取值范围为 $\left(\frac{s-1}{n},\frac{s}{n}\right]$,对式(4)改进后的特征词权重计算公式如下:

$$w_{ij}=tf_{ij}+Sim_{ij}\times s/n \tag{5}$$

根据式(5)不仅可以识别出特征词对文档的贡献率,而且可以根据词汇权重小数部分的范围很直观的识别该特征词的情感。

4.2 资源特殊情感识别

根据每个子情感 Ontology 中的计数器值的大小可以识别出每篇文档的情感倾向,即若某一子情感 O_s 的计数器 c_s 的值大于其他子情感中的计数器的值,就可以判断该篇文档与情感 O_s 最接近,从而可以识别出这篇文档的情感。

5 基于情感 Ontology 的资源分析实验流程及结果

实验的语料是从网上搜索的校园论坛的帖子,共 479 篇,人工鉴定每个帖子的情感倾向作为 Baseline,其中正面倾向的帖子有 142 篇,负面倾向的帖子有 337 篇;负面倾向的帖子中有 194 篇体现忧郁情感,96 篇体现恐惧情感。特征词汇集划分如下:特征词汇集一选取将情感 Ontology 中的词汇导入分词的扩展词典,利用分词工具对文本进行分词后的所有词汇;特征词汇集二选取经过词汇倾向性计算后初步抽取的能够反映文本情感倾向性的特征词汇;特征词汇集三选取在特征词汇集一的基础上利用情感 Ontology 中表示程度级别的词进行相似度权重调整后抽取的特征词汇。实验具体流程如下:

1)从网络上收集语料,经过预处理后形成纯文本文档,并人工鉴定每篇词汇的情感倾向。

2)将情感 Ontology 中的词汇导入分词的扩展词典,利用 ICTCLAS 的分词工具完成语料的切分等处理工作,以所有词汇作为特征词汇集一。

3)利用知网计算词汇相似度的方法计算文本中所有词汇与情感 Ontology 中所有词汇的相似度,根据词汇的相似度权重抽取特征词汇,构成特征词汇集二。

4)利用情感 Ontology 中的程度级别词调整特征词汇的相似度权重,抽取特征词汇集三。

5)利用本文所提出的资源分析模型对上述三种文本特征词汇集进行"忧郁"和"恐惧"两种情感的识别。

6)评估实验结果。实验结果如表 1 所示,通过分析,实验二和实验三的情感识别的正确率分别比实验一高 17.29%和 20.51%,可见根据情感 Ontology 来抽取文本的特征词汇可以显著提高资源情感识别的准确率。实验三的结果优于实验二,说明利用程度级别词对特征词汇进行调整可以进一步提高情感识别的准确率。Baseline 与采用本文方法进行情感识别的文档归类对比情况如图 3 所示。虽然采用基于情感 Ontology 的资源情感识别方法得到的实验结果并不是非常理想,但是该方法为网络资源分析提供了一种新的思路。

表 1 实验结果

实验序号	特征词汇集	处理方法	情感识别的正确率%
实验一	特征词汇集一	以所有词汇为特征词	58.36
实验二	特征词汇集二	根据相似度权重抽取的特征词	75.65
实验三	特征词汇集三	特征词汇集二经程度词的调整	78.87

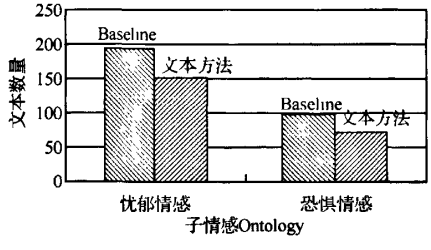

图 3 实验对比

6 结语

本文把情感 Ontology 抽象为一个"地球仪","北极"和"南极"分别对应情感倾向中的正极和负极,"赤道"对应中性,并且可以根据词汇的"经度"和"纬度"计算出词汇与某种特殊情感的相似度,进而可以判断词汇的情感倾向。通过实验对比,利用情感 Ontology 抽取特征词汇比以所有词汇作为特征词汇的情感分析方法的准确率有了明显的提高。我们采用基于情感 Ontology 的资源分析模型对文本的情感进行识别是一种新的尝试,为资源分析提供了一种新的思路,我们需要在此基础上逐步完善,以找到更优的资源分析方法。

参 考 文 献

[1]Bo Pang, Lillian Lee. seeing stars:Exploiting Class Relationships for Sentiment Categorization with Respect Rating Scales[C]. ACL,2005:115～124

[2]Ruifeng Xu, Yunqing Xia, Kam-Fai Wong, et al. Opinion Annotation in On-line Chinese Product Reviews [C]. the 6th International Language Resources and Evaluation LREC-08, 2008:1625～1632

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

[4]Peter D. Turney. Thumbs Up or Thumbs Down? Semantic Orientation Applied to Unsupervised Classification of Reviews[C]. the 40th ACL,2002:417～424

[5]刘群,李素建. 基于《知网》的词汇语义相似度计算[EB/OL]. http://www. keenage. com

[6]董振东,董强. 知网简介[EB/OL]. http://www. how—net. com

[7]情感分析用词语集(beta 版)[EB/OL]. (2007－10－22). http://www. keenage. com

[8]徐琳宏,林鸿飞,杨志豪. 基于语义理解的文本倾向性识别机制[J]. 中文信息学报,2007,21(1):79,89,99

(上接第 53 页)

模式,能为用户带来更好的应用体验。经实践证明,与传统调查方式相比,用户更愿意在一种对空间或时间的限制较少,相对轻松和从容的气氛中提供信息数据。同时,反向支付方式也以其便捷、人性化的优点被用户认可。下面以某电视台调查收视率为例,结合反向移动电子商务平台的工作流程进行实例分析。

在"信息交易请求"阶段,首先反向移动电子商务平台的数据分析系统对电视台提交的调查问卷进行格式处理,生成题库。然后,从移动用户样本库中提取移动用户数据,通过平台服务号(如:10668888)向已提取的用户手机号码(如:13988888888)发送调查请求短信息。用户如果接受调查请求,则回复指定代码(如:Y)到服务号10668888。

在"信息数据交互"阶段,首先反向移动电子商务平台的业务处理中心选择企业调查题库中的一题通过服务号 10668888 发送给用户 13988888888,例如(您一周中的哪一天收看本台节目较多? a 周一 b 周二 c 周三 d 周四 e 周五 f 周六 g 周日)。用户 13988888888 根据自己的实际情况进行回复,例如用户在周二收看节目较多,则回复字母 b 到服务号 10668888。

反向移动电子商务平台接收到用户 13988888888 的第一题信息反馈后,接着从题库中取出第二题,通过服务号 10668888 发送到用户手机上,例如(您最喜欢收看本台的哪类节目? a 新闻 b 综艺 c 音乐 d 体育 e 电视剧 f 访谈节目)。以此类推,反向移动电子商务平台与用户完成余下的信息数据交互。

在"信息数据支付"阶段,反向移动电子商务平台在确认用户 13988888888 完成所有的信息数据交互流程之后,形成资费支付文件(如本次电视台设定奖励额度为 5 元),提交给移动通信网的小额支付系统。通过移动通信网的支付系统,将报酬以话费等形式充值到用户的个人帐户。

6 结语

反向移动电子商务平台作为一种创新的电子商务模式,其交易模式、商品形式及支付方式均与传统电子商务模式不同。实践证明,反向移动电子商务平台能为用户带来更好的应用体验。作为对传统电子商务模式的补充,反向移动电子商务平台将逐渐被用户所认可,帮助企业加速信息沟通,提高应变能力,增加自己的竞争实力。

参 考 文 献

[1]舒凯. 移动电子商务的信息安全标准研究[J]. 信息技术与标准化,2004,(8):14～17

[2]张璞,文登敏. 基于 J2ME 和 J2EE 的移动电子商务系统的研究[J]. 成都信息工程学院学报,2006,4:504～507

[3]黎星星,黄小琴,朱庆生. 电子商务推荐系统研究[J]. 计算机工程与科学,2004,(5)

[4]刘振滨. 电子商务环境下的逆向物流分析与设计[J]. 经济与管理,2006,(1)

[5]王太成. 电子商务系统结构研究[J]. 通信与信息技术,2005

[6]魏永红. 基于 J2ME 技术的手机信息查询系统的设计与实现[J]. 微计算机信息,2006,(4－3):280～282

[7]余力,刘鲁,罗掌华. 我国电子商务推荐策略的比较分析[J]. 系统工程理论与实践,2004,(8)

[8]崔金红,王旭. 基于 UML 的电子商务系统建模研究[J]. 情报杂志,2004,(2)