基于语义理解的文本倾向性识别机制

徐琳宏，林鸿飞，杨志豪

（大连理工大学计算机科学与工程系，大连 116024）

摘 要： 文本倾向性识别在垃圾邮件过滤、信息安全和自动文摘等领域都有广泛的应用。本文提出了基于语义理解的文本倾向性识别机制。其主要思想是首先计算词汇与知网中已标注褒贬性的词汇间的相似度，获取词汇的倾向性；再选择倾向性明显的词汇作为特征值，用 SVM 分类器分析文本的褒贬性；最后采用否定规则匹配文本中的语义否定的策略提高分类效果，同时处理程度副词附近的褒义词和贬义词，以加强对文本褒贬义强度的识别。

关键词： 程度副词；知网；相似度计算；否定句；Support Vector Machine

Text Orientation Identification Based on Semantic Comprehension

Xu Linhong, Lin Hongfei, Yang Zhihao

(Department of Computer Science and Engineering, Dalian University of Technology , Dalian 116024)

Abstract: At the fields of spam filtering, information security and Automatic Abstracting, text orientation identification is used widely. The paper established the mechanism based on Semantic Comprehension for text orientation identification. Firstly, it acquired the semantic orientation through computing semantic similarity the vocabulary and tagged vocabulary in HowNet, and then used the derogatory or commendatory terms as feature, utilized SVM classifier to identify the text orientation. Finally the article dealed with the negative sentence via matching negative rules, moreover identified the derogatory or commendatory intensity through degree adverb.

Key Words: Degree adverb; HowNet; semantic similarity; negative sentence; Support Vector Machine

1 引言

随着互联网的普及，越来越多的人们从网络获取知识和发布信息，对这些信息的有效处理和过滤已成为一个重要的研究课题。文本倾向性识别可以鉴别用户对某产品、事件和政策等持褒义还是贬义的观点。目前，倾向性识别广泛地应用在许多研究领域，具有极大的实用价值。在企业中，产品评论的褒贬性评估，可以为管理者提供准确而有效的决策信息。在垃圾过滤和信息安全方面，将强烈支持不良观点（如宣传西藏独立）的信息过滤掉。在其他研究领域，如自动文摘提取中，可将褒贬义词汇密集的句子和段落摘出，更好的反映原文的中心思想。

自从上世纪九十年代以来，词汇倾向性的研究在国外得到了普遍的关注，并迅速发展起来。Hatzivassiloglou

基金资助：国家自然科学基金（60373095）

作者简介：徐琳宏（1979），女，辽宁，硕士研究生 qingniao1203@163.com.

林鸿飞（1962），男，辽宁，教授，博士，hflin@dlut.edu.cn.

and McKeown 在 1997 年首先开始了词汇倾向性的研究。他们主要是针对形容词作倾向性分析，利用词汇之间的连词（and, or, but, either-or, 和 neither-nor 等）训练生成词汇间的同义或反义倾向的连接图，然后用聚类的方法将词汇聚成褒义和贬义两类。精确率最低的一组实验也达到 78.08%[1]。

2003 年 Turney and Littman 采用计算基准词对与词汇相似度的方法识别词汇倾向性。他们选择了七对褒贬倾向比较强烈的词汇，计算待定词与每个基准词的 SO-PMI(semantic orientation – pointwise mutual information)值来判定词汇的倾向性[2]。

2004 年 J. Kamps, M. Marx, R. J. Mokken, and M. D. Rijke 利用 WordNet 计算词汇倾向性。先选择基准词，判别待定词与基准词在 WordNet 中是否为同义词，得出词汇的倾向性,计算公式如下[3]：

其中 $d(t_1,t_2)$是词汇 t_1,t_2 在由 WordNet 生成的相似图中的最短路径，bad 和 good 分别代表贬义和褒义基准词

$$SO(t)=\frac{d(t,bad)-d(t,good)}{d(good,bad)} \tag{1}$$

2005年M.J.M. Vermeij利用有倾向性的词汇在产品评论中出现的次数计算用户评论的倾向性，提出了一种按词频加权统计的方法[4]

目前，在中文词汇倾向性计算方面的研究刚刚起步。主要的方法是选择基准词对，利用知网计算倾向性待定的词汇与基准词汇的相似度，从而判定词汇的倾向性[5]。

2 词汇倾向性计算方法

2.1 知网简介

《知网》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常识知识库。知网的基本思想是：设想所有的概念都可以分解成各种各样的义原，同时应该有一个有限的义原集合，其中的义原组合成一个无限的概念集合 [6]。

2.2 词汇倾向性计算的方法

本文采取的方法是赋予每个词汇一个语义倾向的度量值[5]，它表示词汇与基准词间关联的紧密程度。基准词是指具有强烈褒贬倾向的词汇，本文选择知网中已标注“良”和“莠”的词汇作为标准集，总共 6952 词。其中褒义词 3361，贬义词 3591 个。

首先，在知网中确定词汇的定义，即它的义原组成，然后在标准集中查找它的可能近义词。本文中可能近义词主要通过考察词汇的第二位和第三位上的义原是否相同来确定（若不存在第三位上的义原，则取第一位和第二位的义原，依此类推），这是因为褒贬义词大多数为形容词，而在知网中，形容词在第二位上一定要标注该属性值或数量值所指向的属性或数量特征。通常多数情况下在第三位置上标注该属性值或数量值的具体值[8] ，例如：

美味： DEF=aValue|属性值,taste|味道,good|好

暗淡： DEF=aValue|属性值,circumstances|境况,miserable|惨,undesired|莠

斯文： DEF= aValue|属性值,behavior|举止,gracious|雅,desired|良

可见第二位和第三位上的义原能更好地保留词汇的倾向性。

然后采用知网的语义相似度的计算公式计算可能近义词与待定词汇的相似度。其中定义词汇的义原的相似度采用如下公式：[7]

$$Sim(p_1,p_2)=\frac{\alpha}{d+\alpha} \tag{2}$$

其中p_1和p_2表示两个义原（primitive），d是p_1和p_2在义原层次体系中的路径长度，是一个正整数。α 是一个可调节的参数 [7]。在此基础上计算词汇与可能近义词间第二位和第三位义原的相似度，取其中最大的作为词汇的相似度。

最后将待定词汇与所有可能近义词的相似度求和，获得词汇的倾向性度量值。其中与“良”性词汇的相似度取正值，与“莠”性词汇的相似度取负值。词汇W的语义倾向性计算公式如下：

$$Orientation(W)=\sum_{i=1}^{kp} Sim(WP_i,W) - \sum_{j=1}^{kn} \mathrm{Sim}(\mathrm{WN_j},\mathrm{W}) \tag{3}$$

其中WP_i表示褒义基准词，WN_j表示贬义基准词，W为倾向性待定的词汇，kp和kn分别为褒义和贬义的可能近义词数。

3 文本倾向性的识别

3.1 文本的预处理及特征选择

文本的预处理就是将文本转化为计算机可以识别的格式，本文采用目前应用较广泛的向量空间模型（VSM）来表示文本。权重的计算是使用著名的tf*idf公式：

$$W_{ij}=\frac{tf_{ij}*\log(N/n_i+0.01)}{\sqrt{\sum_{k=1}^{N}[tf_{ik}*\log(N/n_i+0.01)]^2}} \tag{4}$$

其中，w_{ij}表示词i在文本j中的权重，而tf_{ij}为词i在文本j中的词频，N为训练文本的总数，n_i为训练文本集中出现的i的文本数，分母为归一化因子。

构成文本的词的数量一般非常庞大，所以向量空间的维数也非常大，有的甚至达到数万维。这就需要约简文本中的特征项，本文中选择褒贬倾向比较强烈的词作为特征项，这样一方面可以大幅度压缩向量空间的维数，缩短机器学习和分类的时间。另外具有褒贬倾向的词汇能更好的保留原文中作者的观点，提高文本倾向性识别的正确率。

3.2 否定句的处理

在逻辑语义上，否定词是判断主体不具有某种特征或行为的。例如：

演技一点也没进步。

表演极不自信。

其中“进步”和“自信”本来都是褒义词，但是前面加上否定词“没”、“不”，整个句子的语义就转变为贬义了。本文对于上述否定句的处理方法是进行否定规则匹配，被匹配上的词汇褒贬义的性质变反，以正确反映整篇语料的观点。首先从复旦提供的语料库中提取出否定句 242,917 个，在大量的否定句中提炼出高频的否定规则集合。然后将否定规则与语料中否定句匹配，如果否定中心恰好为有褒贬倾向的词汇，则将其用相反意义的词汇替代，以消除否定句对文本观点识别的负面影响。

本文中否定词的获取是通过知网实现的。在知网中选取具有否定意义的义原，从中抽取出包含否定义原的概念，经人工过滤得到 18 个否定词。

3.3 程度副词对语义强度的影响

王力先生在《中国现代语法》里指出:“凡无所比较,但泛言程度者,叫绝对的程度副词。”“凡有所比较者,叫做相对的程度副词” [9]。程度副词会影响句子的语义强度。例如：

他的汉语挺好的

他的汉语说得非常好

他的汉语说得极其好

上述三个例句语义强度依次递增。为了更好的区分作者观点的褒贬义强度，本文对程度副词上下文设置一个观察窗口，观察窗口的大小作为一个参数从训练集得出最佳的选择，这里窗口的大小是按词汇与程度副词切分出的距离来计算，不是两者相距的字数。如果褒贬义词出现在观察窗口内，则按程度副词的量级差别相应增加褒贬义词汇的词频，其中从极量副词到低量副词分别将所修饰词汇的褒贬义强度增加1.5到1.2倍。程度副词的量级差

别做如下划分：[10]

表1 程度副词分类表

Tab.1 Classification table of degree averb

程度副词	相对程度副词	极量	最 最为
		高量	更 更加 更为 更其 越 越发 备加 愈 愈加 愈发 愈为 愈益 越加 格外 益发 还
		中量	较 比较 较比 较为 还
		低量	稍 稍稍 稍微 稍为 稍许 略 略略 略微 略为 些微 多少
	绝对程度副词	极量	太 极 极为 极其 极度 极端 至 至为 顶 过 过于 过分 分外 万分
		高量	很 挺 怪 老 非常 特别 相当 十分 好 好不 甚 甚为 颇 颇为 异常 深为 满 蛮 够 多 多么 殊 特 大 大为 何等 何其 尤其 无比 尤为 不胜
		中量	不大 不太 不很 不甚
		低量	有点 有些

3.4 分类方法

本文采用两种分类方法，即SVM方法和词频加权统计的方法

SVM是一种基于统计的学习方法，它是对结构化风险最小化归纳原则（Structure Risk Minimization Inductive Principle）的近似，其理论基础是统计学习理论[11]。

词频加权统计的文本倾向性识别方法是计算文本中特征项的tf*idf值，并将此值与对应词汇的倾向性度量值相乘（褒义为正值，贬义为负值）。最后将文本中所有特征项的值相加取平均作为该篇文章的倾向性度量值。

4 实验流程及结果

实验的语料是从网上搜索的影评，共499篇。本文采取SVM和词频加权统计两种分类方法。并将这两种方法应用在四类特征集上，特征集划分如下：特征集一选取文本中所有词汇；特征集二只选取有褒贬倾向性的词汇；特征集三在特征集二的基础上加入否定句的处理结果；特征集四是在特征集三的基础上再加入对程度副词的处理。

SVM实验取其中185篇为训练集，314篇作为测试集。特征词加权统计的实验不需要训练集，但为了便于比较，采用SVM全部测试集314篇作为语料。实验具体流程如下：

（1） 利用网页抓取程序，从网上搜索语料，并人工鉴定每篇词汇的褒贬倾向性。

（2） 将知网中的词汇导入分词的扩展词典，利用哈工大的分词完成语料的切分等预处理工作。

（3） 根据知网的相似度计算公式编写程序，从语料中提取具有褒贬义的词汇，生成词汇表，并计算其倾向性的度量值。

（4） 分别以全部词汇和（3）步中的词汇表为特征项，生成文本分类的特征集一和特征集二。

（5） 从复旦语料中抽取高频否定规则。

（6） 将否定规则与语料中的否定句匹配，将匹配上的褒贬义词性质变反，在文本分类特征集二的基础上生成特征集三

（7） 将程度副词应用于（6）中的文本分类特征集三，生成特征集四

（8） 利用SVM在上述四种文本分类特征集上进行褒贬义分类。

（9） 用词频加权统计的方法计算文本的倾向性

（10） 评估结果的正确率和召回率。

表 2　实验的结果

Tab.2 The result of experimentation

实验序号	文本分类特征集	处理方法	SVM分类的F-Score%	词频加权统计的正确率%
实验一	特征集一	以所有词汇为特征项	76.62	无意义
实验二	特征集二	以褒贬义词汇为特征项	81.48	65.61
实验三	特征集三	特征集二加入否定句的处理	84.28	69.75
实验四	特征集四	特征集三增加程度副词的处理	85.58	69.43

实验一用所有词汇为特征项，实验二用褒贬义词为特征项，结果表明实验二比实验一的F-score值提高5%左右。增加否定句的处理，又使结果在实验二的基础上提高了大约3%。可见加入褒贬义词和处理否定句对文本的倾向性识别有一定的帮助。程度副词的处理虽然对结果的正确率提高不大，但它尝试为文本褒贬义强度的分析提供一种新的思路，即除了判别作者的观点的倾向性，还要更好地判定这种倾向性的大小。词频加权统计方法不区分训练集和测试集，不进行机器学习，只是对结果做简单得求和取平均运算，在三类特征集上的正确率都低于SVM的分类方法。

5　结束语与进一步改进

本文所提出的方法是对文本倾向性识别的一个初步尝试，在词汇倾向性计算的基础上对文本褒贬义倾向性进行分类。利用知网计算语义相似度，获取具有褒贬倾向的词汇的度量值，以这些有强烈倾向性的词汇作为特征项，采用目前分类效果较好的SVM对文本的倾向性分类。在文本的倾向性识别方面主要是采用词频向量空间模型，处理文本中的否定句，以正确反映整篇语料的观点。对程度副词修饰的褒贬义词汇增加其词频，以更好的反映文本的褒贬义强度。目前中文在倾向性研究方面的语料还不够丰富，本文的语料是作者手工收集和整理，语料的丰富和校验工作还需进一步进行，同时词汇倾向性度量值采用最简单的义原间距离计算，并且未进行词汇与概念间的词义消歧。另外，处理否定句时只考虑了语义否定，而没有考虑语用否定的情况。对篇幅较大的文章应通过中心句和中心段的识别来提高分类精度[12]。以上情况都有待作进一步细致的研究。

参考文献：

[1] V. Hatzivassiloglou, K. R. McKeown. Predicting the semantic orientation of adjectives[A]. In Proceedings of ACL-97, 35th Annual Meeting of the Association for Computational Linguistics[C], Madrid, ES, 1997: 174–181

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

[3] J. Kamps, M. Marx, R. J. Mokken, and M. D. Rijke. Using WordNet to measure semantic orientation of adjectives[A]. In Proceedings of LREC-04, 4th International Conference on Language Resources and Evaluation[C], Lisbon, 2004 . 1115 - 1118

[4] M.J.M. Vermeij. The Orientation Of User Options Through Advers, Verbs And Nouns[A]. 3rd Twente Student Conference on IT , Enschede, 2005

[5] 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

[6] 董振东. 语义关系的表达和知识系统的建造[J]. 语言文字应用, 1998, 27(3): 76-82.

[7] 刘　群, 李素建. 基于《知网》的词汇语义相似度计算. http://www.keenage.com

[8] 董振东, 董　强. 《知网》 http://www.keenage.com

[9] 王　力. 中国现代语法[M] . 北京:商务印书馆,1985. 131-132

[10] 蔺　璜 ,郭姝慧. 程度副词的特点范围与分类[J]. 山西大学学报(哲学社会科学版), 2003,26(2): 71-74

[11] Vapnik V. The Nature of Statistical beaming Theory. New York;Sprfinger-Verlag,1995

[12] 金　珠, 林鸿飞, 赵　晶. 基于 HowNet 的话题跟踪及倾向性分类研究[J]. 情报学报, 2005, 24(5): 555-561