文章编号：1003-0077(2007)06-0088-07

基于监督学习的中文情感分类技术比较研究

唐慧丰[1,2]，谭松波[1]，程学旗[1]

(1. 中国科学院 计算技术研究所 北京 100080；
2. 解放军外国语学院 河南 洛阳 471003)

摘 要：情感分类是一项具有较大实用价值的分类技术，它可以在一定程度上解决网络评论信息杂乱的现象，方便用户准确定位所需信息。目前针对中文情感分类的研究相对较少，其中各种有监督学习方法的分类效果以及文本特征表示方法和特征选择机制等因素对分类性能的影响更是亟待研究的问题。本文以n-gram以及名词、动词、形容词、副词作为不同的文本表示特征，以互信息、信息增益、CHI统计量和文档频率作为不同的特征选择方法，以中心向量法、KNN、Winnow、Naïve Bayes和SVM作为不同的文本分类方法，在不同的特征数量和不同规模的训练集情况下，分别进行了中文情感分类实验，并对实验结果进行了比较，对比结果表明：采用BiGrams特征表示方法、信息增益特征选择方法和SVM分类方法，在足够大训练集和选择适当数量特征的情况下，情感分类能取得较好的效果。

关键词：计算机应用；中文信息处理；情感分类；文本分类；语言模型；中文信息处理

中图分类号：TP181 **文献标识码**：A

Research on Sentiment Classification of Chinese Reviews Based on Supervised Machine Learning Techniques

TANG Hui-feng[1,2], TAN Song-bo[1], CHENG Xue-qi[1]

(1. Institute of Computing Technology, Chinese Academy of Sciences, Beijing 100080, China;
2. PLA University of Foreign Languages Luoyang, Luoyang, Henan 471003, China)

Abstract: Sentiment classification is an applied technology with great significance. It can solve information disorder and help people locate the required reviews in the Internet. Up to now, most research of sentiment classification is on English reviews, and little work has been done on Chinese reviews. To find an effective way for the task based on supervised machine learning method, and analyze the influence by term expression and term selection, this paper conducted some experiments under distinct environments, including different feature representation, different feature selection, different categorization technique, different size of features and different size of training data, over Chinese text collections. The experimental results show that sentiment classification will obtain high performance, when using bigrams representation, information gain and SVM classifier, enough training data and plenty of features.

Keywords: computer application; Chinese information processing; sentiment classification; text categorization; language model; Chinese information processing

1 引言

随着越来越多的人使用互联网作为一种信息交流的手段，免费可用的在线产品和服务评论也呈现快速上升的势头。对于那些想要获得购物指导的消费者来说，尽管有这么丰富的资源可以利用，但各种评论浩如烟海，且充斥了肯定和否定术语，想判断这

收稿日期：2007-04-20 **定稿日期**：2007-07-18

基金项目：国家"973"重点基础研究发展规划基金资助项目(2004CB318109)

作者简介：唐慧丰(1973—)，男，博士，副教授，研究方向为文本分类、机器学习；谭松波(1978—)，男，博士，助理研究员，研究方向为文本分类、高性能计算与机器学习；程学旗(1971—)，男，博士，研究员，研究方向为智能信息处理、知识检索与算法分析、信息安全、计算语言学等。

些评论的极性(肯定还是否定)，仍然具有相当大的难度。再者，有的消费者只想阅读某种产品的负面评论，以便了解该产品的缺点，而不愿意花时间阅读其正面评论；反之，对于一部影片感兴趣的影迷只想阅读它的正面评论，以便了解该影片的看点。一篇评论的极性是正面还是负面，可通过某种分类方法赋予一个数值来表达，该数值所对应的分类结果被定义为评论的全面观点极性(Overall Opinion Polarity, OvOP)，简称极性。全面观点极性的分类过程称为全面观点极性辨识(Overall Opinion Polarity Identification, OvOPI)，简称观点鉴别[1]。观点鉴别有别于通常所说的自动文本分类，自动文本分类关心的是文档的主题，如文档是属于娱乐类的还是属于体育类的，观点鉴别主要用来辨识自然语言文字中表达的观点、喜好以及与感受和态度等相关的信息，因此有些文献也称其为情感分类(Sentiment Classification)[2]，为了表达的一致性，本文中将其统一表述为情感分类。

由于情感分类可以在一定程度上解决网上各种评论信息杂乱的现象，方便用户准确地定位所需信息，因此，情感分类已成为一项具有较大实用价值的关键技术，是组织和管理数据的有力手段。近年来，相当多的厂家、公司和贸易团体对信息的情感分类有着很强的需求，该领域的研究得到了很多专家的重视。情感分类对于自动处理用户反馈、寻找广告对象和分析消费走势等都能起到相当大的作用。另外，情感分类在电子邮件过滤和博客态度分析上也有着较为普遍的应用。

目前，关于英文情感分类的文献，大多见于在国外召开的国际会议，这些文献采用的研究方法主要可归纳为以下两点：(1)使用有监督的机器学习的方法对英文数据集进行情感分类[2, 10, 17]；(2)使用英文情感词典提取数据集中与情感相关的元素作为情感分类的依据[1, 5~7]。从最新研究进展来看，由于自然语言理解领域还存在一些关键技术尚待研究，方法(2)相比方法(1)，其性能并无明显优势。在情感分类的研究中，英文和中文的分析方法有所不同，如特征提取方法和情感词典构建等方面都存在一定差异。目前国内针对中文情感分类的研究相对较少。

要很好地利用庞大的中文网络评论资源，还有许多亟待解决的问题：(1)各种有监督的学习方法在中文数据集上的情感分类效果孰优孰劣；(2)文本特征表示方法和特征选择机制等因素对中文情感分类的性能将产生什么影响；(3)文档集的哪些情感特征对情感分类的精度具有决定性影响，等等。本文旨在解决前两个问题，通过分析常规分类方法的特点，研究了各种特征表示和特征选择方法对情感分类结果产生的影响，并对实验结果进行了详细对比分析。实验表明，采用 n-Gram 特征表示方法、信息增益特征选择方法和 SVM 分类方法，在足够大训练集和选择适当数量特征的情况下，情感分类能取得较好的效果。

本文是这样组织的：第二部分介绍了和情感分类相关的研究工作；第三和第四部分对文本特征表示和压缩、监督学习分类方法进行了概要的阐述；第五部分列举了不同实验环境下进行情感分类的实验结果，并对其进行了详细的分析；第六部分给出了结论，并对今后的工作提出了展望。

2 相关研究工作

迄今为止，介绍情感分类研究工作的文献以国外刊物和会议为主。这些研究工作可归纳为以下几个领域。

2.1 客观性分类

客观性分类是将 Web 上获取的评论文档按照类型和风格的不同区分为主观和客观两类，这类工作以 Finn 等人的文献为代表[3]，其结论是基于词性标注(POS Tagging)的特征选择方法比词袋(Bag-of-Words)方法效果好。Wiebe 等人的文献[4]中对人工标注的语料从短语、句子和篇章层次进行研究，发现对于不同的标注者，其主观性的判别有较大差异。

2.2 词的极性分类

相比客观性分类，词的极性分类难度更大一些，它是通过分析带有情感词的特征来判断词的极性。Andreevskaia 和 Bergler [5]使用 Sentiment Tag Extraction Program (STEP)方法从 WordNet 中提取情感词，该方法利用了 WordNet 中词的同义、反义、上下位等关系和词的注释。Kennedy 和 Inkpen 使用 General Inquirer (http://www.wjh.harvard.edu/~inquirer/homecat.htm)来查找网络评论中的情感词，并提出了 negations(反转)、intensifiers(增强)和 diminishers(减弱)三种算子来计算情感词的强度[6]。Turney 和 Littman[7]的文献中提出了

一种方法，他们使用 AltaVista 中的 NEAR 运算从 Web 上搜索得到两个词同时出现的次数，以此来决定两个词的相似程度，一个新词归属于正面态度还是负面态度，取决于它和手工选择的正面（或负面）种子词集合中所有词的关系，这类工作和常规的词聚类问题有一定的关联。

2.3 情感分类

2.3.1 基于情感词标注

Subasic 和 Huettner[8] 手工建立了一个基于情感类别相关词的词典，词典中标明了词的强度（表达情感的力度）和向心度（与类别的相关程度）。Liu 等人的文献[9] 提出使用 Open Mind Commonsense 数据库为选择的语言特征赋予情感值，并将其归纳为六个基本类别（高兴、悲伤、愤怒、恐惧、厌恶和惊奇）。他们通过分析带有情感色彩词的特征来判断文档的极性。

2.3.2 基于机器学习

Lin 等人[10] 把观点鉴别问题作为一个分类问题来考虑，提出了一种基于统计模型的学习方法，通过分析词的用法来获取文本所反映的观点。Whitelaw 等人[17] 关注研究带形容词的词组及其修饰语（如“extremely boring”或“not really very good”），他们提取带形容词的词组作为特征，基于这些特征，用向量空间模型表示文档，并采用 Support Vector Machine 进行分类，来区分带有正面和负面评论的文档。Pang 等人[2] 以 Usenet 上的电影评论作为语料进行了研究，采用了不同的特征选择方法和机器学习方法（Naïve Bayes，Maximum Entropy 和 Support Vector Machine）。其实验结果显示，基于 presence-based frequency 模型选择 Uni-Grams 的方法，并采用 Support Vector Machine (SVM)进行分类，能取得最好的分类结果，其准确率为 82.9% 。

3 文本特征表示和特征压缩

3.1 文本特征表示

进行情感分类之前，首先需要把数据集中的文本表示成特征，可以采用反映文本语言学特征的元素来表示特征，如使用词、n-Gram、词组和概念等。向量空间表示模型（VSM）是目前文本表示的主要方法，相关研究集中在以什么语义单元作为项及如何计算项的权重两个问题上，通常以项的出现频率作为基础计算权重[11]，假定集合 $\{f_1, f_2, \cdots, f_m\}$ 是文档 d 中出现的 m 个特征，令 $n_i(d)$ 是特征 f_i 在文档 d 中出现的次数，那么，文档 d 就可以由一个特征向量来表示，记为：

$$\vec{d} = (n_1(d)), n_2(d) \cdots n_m(d)) \quad (1)$$

也有一些文本表示方法希望通过借鉴自然语言处理技术，考虑了被词袋忽略的语义单元间的联系，将词义及短语等复杂的项应用到分类方法的文本表示中。不过这些表示方法在分类效果上还没有明显的优势，而且往往需要比较复杂的语言预处理，在分类时会影响分类器的速度。到目前为止，非 VSM 的表示在理论上的合理性及面对实际应用的可扩展性还需要深入验证。

3.2 文本特征压缩

用文本表示方法生成的特征中可能存在很多噪声，通过特征压缩舍弃一些不太重要的特征，将有效消除噪声的影响，并起到降低向量空间的维数、简化计算、防止过分拟合的作用。

文本特征压缩的研究大体可以分为特征选择[12] 和特征抽取两类[13]。特征选择是根据某种准则从原始特征中选择部分最有类别区分能力的特征；特征抽取是依据某种原则构造从原始特征空间到低维空间的一个变换，从而将原始特征空间所包含的分类信息转移到新的低维空间中来。常用的特征选择与特征抽取算法的效果在不同情况下互有高低，特征选择方法因其复杂度较低而应用更为广泛，而抽取得到的特征更接近文本的语义描述。常见的用于特征选择的衡量标准有文档频率、信息增益、互信息和 CHI 统计等，而在特征抽取方面，主成分分析、线性区分分析、概念索引等方法也先后被引入文本领域。本文实验中使用的均为特征选择方法。特征数量的变化和分类器效果紧密相关，有关文献[14] 的结论表明：合理的特征压缩方法会使多数分类器的性能快速提高并能迅速接近平稳；但若特征数目过大，分类器的性能反而可能出现缓慢降低。

4 基于统计学习的分类方法

本文是将已有的基于监督学习的分类方法与不同的特征选择方法相结合，相对于中文语料寻找较为满意的情感分类方法。这里，我们介绍几种经典的基于统计学习的分类方法。

4.1 中心向量分类方法

中心向量分类方法是一种简单有效的分类法，所有文档都用特征向量来表示，在此基础上，对于所有属于同一类别的文档计算出一个平均向量（即中心向量）。给一个样本向量分类时，只需计算它与各中心向量的相似度，取相似度最大值的中心向量所在类别作为样本的类别即可。

4.2 k-近邻（KNN）分类方法

k-近邻分类方法是一种非常有效的归纳推理方法，直观地讲，k-近邻分类方法就是从测试文档 d 开始生长，并不断扩大区域，直到包含 k 个训练样本点为止，并且把测试文档 d 的类别归为这最近的 k 个训练样本点中出现频率最大的类别。

4.3 感知器分类方法

形式最简单的用于语气挖掘的感知器模型就是关于测试文档集的线性组合。在训练过程中，每出现一个错误的分类样本就反馈更新一次权值，我们可以通过寻找一系列权值使训练集中的所有点都满足上面的条件来实现这种分隔。由感知器模型派生的一个分类算法是 Winnow 算法，Winnow 算法有很多变种，我们在实验中采用的是平衡 Winnow，即对于每一个特征有两个权值 ω^+ 和 ω^- 相对应。

4.4 Naïve Bayes（NB）分类方法

尽管 Naïve Bayes 方法非常简单，并且它的条件依赖假设在实际应用中也往往得不到满足，但使用它却常常能取得令人满意的文本分类结果。实际上，Domingos 和 Pazzani 的文献[15]表明，Naïve Bayes 方法在某些具有强依赖性特征的分类问题上可获得较好的结果，但若条件依赖假设不成立，则通常得不到好的分类结果。

4.5 支持向量机分类方法

支持向量机分类方法（Support Vector Machines，SVMs）是传统分类中非常有效的一种方法，它的分类结果比 Naïve Bayes 方法普遍要好，其目标是给定一个训练集，找到一个具有最大间隔的分隔平面（也称超平面）$\vec{\omega}$，并且间隔越大，得到的分类器也越好。语气挖掘的目的是基于文档特征向量将文档分为正面和负面两类，采用 SVM 方法相当于一个求解一个带约束条件的最优化问题。

5 实验及结果分析

5.1 数据集

实验使用了影视、教育、房产、笔记本电脑（表格中简称电脑）和手机五个主题的中文评论数据，所有数据都是我们从互联网的相关中文评论网站采集获得的。由于同一主题的评论可能出现在不同的评论网站，为防止数据集中出现重复的样本，对于特定的 URL 地址我们指定了特定的采集者。语料采集后，经过抽取，转换成统一的文本格式，并经人工标注极性（正面评论或负面评论），最终得到实验使用的数据集。数据集中的样本情况如表 1 所示：

表 1 实验用数据集样本情况

主题	样本数		
	总数	Negative	Positive
影视	1 980	1 062	918
教育	1 476	1 012	464
房产	1 118	733	385
电脑	901	451	450
手机	992	497	495
合计	6 467	3 755	2 712

从表 1 可见，五个主题的中文评论样本总数将近 6 500 条，评论是有选择性采集的，其表达的态度比较明确，以下是评论中的片段：

例 1：“你一定要看这部纪录片《圆明园》，本年度最好的影片之一，也是中国最好的电影记录片之一。”

例 2：“V703SH 的内屏使用了可视面积为 2.0 英寸的 26 万色 QVGA，它的“四度”（亮度、饱和度、细腻程度、可视角度）表现都非常出色，在当今同类型的手机中堪称极品！”

例 3：“在一些地方，劣质高价的校服流入校园，都与校服采购的黑箱操作及校方拿校服回扣有关。可见，校服由教育主管部门和学校采购，既有悖于市场公平交易原则，也不利于有效监督，很容易造成腐败。”

例 4：“nc4000 最失败的地方就是用了 ATI 的芯片组，对 PM 的 StepSpeed 支持很差！降频只能降到 600MHz，因此待机时间短，散热做的也不好，风扇经常转，加了内存之后背面烫的厉害！”

从上面的例子来看，尽管写法不是很正式，但从例1和例2中能看出作者表达了明显的肯定态度，例3和例4则表达了明显的否定态度。

数据集中教育和房产两个主题的正负面评论数量上存在较大的差异，主要是因为这两个领域的负面评论明显多于正面评论的缘故，其它主题的正负面评论数量相当。各类评论文档的长度各异，影视类评论的平均篇幅最长，约为500个汉字，手机类评论的平均篇幅最短，约为60个汉字。

5.2 实验结果分析

5.2.1 基于n-Gram的特征表示方法实验结果分析

本实验中，我们分别采用了三种类型的特征表示方法，即UniGrams、BiGrams和TriGrams，其他实验条件相同，即对于每个主题，使用50%的数据作为训练集，剩余50%的数据作为测试集，选取全部特征，使用SVM分类方法。实验结果如下：

表2 基于n-Gram的特征表示方法分类精度比较

	影视	教育	房产	电脑	手机
UniGrams	83.0	97.9	96.1	92.0	97.4
BiGrams	83.0	97.4	96.8	94.2	97.2
TriGrams	81.8	93.1	92.7	91.6	96.6

从上表中可以看出，不同领域的分类精度有较大的差异，其中教育、房产和手机领域的精度较高，而影视领域的分类精度要低得多，降低的幅度达18%左右。这是因为语料文本本身的差别所致，影视类语料使用的描述语言中修辞、比喻较多，且风格各异，对于情感分类精度有很大的影响，从后续的实验结果也都能反映出该问题。n-Gram中各特征表示方法所产生的实验结果都不差，整体而言，BiGrams要略好于另外两种。

5.2.2 基于不同词性的特征表示方法实验结果分析

通过对评论语料进行分析，我们发现，情感分类与其他分类的差别在于，情感的正面表达和负面表达主要以形容词、副词和少数动词和名词的表达为主，因此，我们使用不同词性的词来表示特征，对选取四种词性（名词、动词、形容词、副词）中的一种和选取它们的全部（下表中的nvaa）分别进行了实验，其他实验条件相同，即对于每个主题，使用50%的数据作为训练集，剩余50%的数据作为测试集，选取全部特征，使用SVM分类方法。实验结果如下：

表3 不同词性作为特征表示方法的分类精度比较

	影视	教育	房产	电脑	手机
名词	78.2	95.4	95.7	69.4	88.1
动词	71.1	94.9	94.5	73.9	84.1
形容词	63.0	86.2	79.2	74.9	82.2
副词	58.6	86.2	69.8	73.6	80.2
nvaa	81.5	97.4	96.4	89.8	96.8

从整体实验结果来看，以单个词性为特征的分类精度均比表2中n-Gram为特征的分类精度要差很多，以四种词性为特征的分类精度却能和n-Gram的精度相当，这说明单一词性并不能反映评论语料的整体情感特征，而名词、动词、形容词和副词这四种词性的合集已能近似反映出整个文档的情感特征了，只是对于不同的领域，它的稳定性比n-Gram要稍差一些。对于单个词性而言，各领域中基本都是名词和动词作为特征的分类精度要比形容词和副词的结果好，只有个别领域有不同情况，这与预想中形容词和副词带有绝大部分情感特征的想法有较大差异。一个主要的原因在于网络评论的写法较为灵活，风格与网络语言相近，因此表达方式与常规方法有所差异。因此，整合更多用于情感表达的特征来进行分类，将可能大幅提高分类精度。

5.2.3 基于不同的特征选择方法实验结果分析

我们分别采用了互信息（MI）、信息增益（IG）、CHI统计量（CHI）和文档频次（DF）四种不同的特征选择方法进行了实验，其他实验条件相同，即以BiGrams作为特征表示方法，对于每个主题，使用50%的数据作为训练集，剩余50%的数据作为测试集，使用SVM分类方法。实验结果如下：

表4 不同特征选择方法的分类精度比较

	MI	IG	CHI	DF
影视	61.2	75.7	62.9	67.5
教育	83.3	97.2	83.7	92.4
房产	73.1	96.6	93.7	93.7
电脑	72.1	93.3	90.2	91.6
手机	63.7	97.3	94.2	95.8

一般来说，使用BiGrams作为特征，特征空间维数将会很高，在相同规模训练语料条件下，更高的

维数必然导致更多的低频词出现。这种情况下，使用 MI 和 CHI 进行特征选择，由于它们对低频词的倚重，必定会将更多的低频词作为特征使用，从而导致了分类效果不如 DF[13, 16]。从实验结果来看，DF 的分类精度明显高于 MI 和 CHI，这与上述相关文献的结论相吻合，只是 DF 的分类精度比 IG 低一些。这种现象产生的主要的原因在于，网络评论的用语多为一些网络流行语言，甚至是一些表情符号，且不同的网络评论作者可能使用不同的表达方式，而每篇网络评论一般较为短小（我们使用的语料集中，70%以上的网络评论不超过 100 个汉字），这些网络用语在同一篇网络评论中重复的几率相对较低，因此这些特征的 DF 值也相对较低。DF 通过设置阈值去掉了低频词，当低频词为噪音时，的确可提高分类效果，但低频词也可能带有很大信息量，这时直接去掉低频词会损失一部分特征，影响分类效果；而 IG 不但考虑了类别信息，而且考虑了低频词对分类结果的影响，因此分类效果最好。

5.2.4　基于不同分类方法实验结果分析

以 BiGrams 作为特征表示方法，对于每个主题，使用 50%的数据作为训练集，剩余 50%的数据作为测试集，在选取全部特征的情况下，我们分别采用了中心向量、KNN、Winnow、NB 和 SVM 五种不同的分类方法进行了实验。实验结果如下：

表 5　不同分类方法的分类精度比较

	中心向量	KNN	Winnow	NB	SVM
影视	79.5	80.4	74.6	78.2	83.0
教育	95.0	96.6	93.2	95.8	97.4
房产	94.0	95.5	88.7	95.2	96.8
电脑	92.5	88.9	82.0	89.4	94.2
手机	95.8	92.1	87.9	96.2	97.2

在以上几种分类方法中，相比而言，SVM 的分类效率较低，且需要大量的存储资源和很高的计算能力，但它的分隔面模式有效地克服了样本分布、冗余特征以及过拟合等因素的影响，具有很好的泛化能力，因此，SVM 分类方法的分类精度明显高于其他方法。

5.2.5　基于不同特征数量实验结果分析

采用信息增益选择的特征，按照特征权重值大小降序排列，选取权重值靠前的一定数量（500 个，1 000 个，…，10 000 个）特征进行实验，其它实验条件相同，即以 BiGrams 作为特征表示方法，对于每个主题，使用 50%的数据作为训练集，剩余 50%的数据作为测试集，使用 SVM 分类方法。精度随特征数量变化的情况如下表所示：

表 6　不同数量特征的分类精度比较

特征数量	影视	教育	房产	电脑	手机
500	62.8	94.2	90.5	86.0	93.8
1 000	65.8	94.7	91.8	88.9	95.6
2 000	71.3	96.5	93.4	88.5	96.8
3 000	73.6	97.2	95.2	91.6	94.8
4 000	75.8	97.3	96.1	92.0	95.8
6 000	73.3	97.7	95.0	93.3	96.0
8 000	75.3	98.1	95.9	93.8	97.6
10 000	75.7	97.2	96.6	93.3	97.4

从实验结果来看，各领域的分类精度都随特征数量的增加而增大，当特征数量达到一定值时（表 6 中为 8 000），分类精度达到最佳，这说明对于一定的分类数据集，并非选择的特征数量越多越好，当特征数量达到一定值时，分类精度将趋于平稳，特征数量大于该值时，分类精度反而会有不同程度的降低。

5.2.6　基于不同规模训练集的实验结果分析

以上各实验中，对于每个领域，都使用了 50%的数据作为训练集，剩余 50%的数据作为测试集。以下实验用来考察不同规模的训练集对分类精度的影响，我们分别选取了训练集的全部、1/2、1/3…直到 1/10，在全部测试集上进行实验，其他实验条件相同，即以 BiGrams 作为特征表示方法，选取全部特征，使用 SVM 分类方法。实验结果如下：

表 7　不同规模训练集的分类精度比较

规模	影视	教育	房产	电脑	手机
全部	83.0	97.4	96.8	94.2	97.2
1/2	80.0	95.7	95.5	90.2	96.6
1/3	78.1	93.5	94.3	87.4	94.8
1/4	73.7	92.0	91.6	84.9	93.5
1/5	75.7	90.4	90.2	84.5	95.2
1/6	71.6	87.9	88.6	81.4	86.9
1/7	68.8	85.9	82.8	84.3	93.0
1/8	70.1	82.8	80.9	80.3	91.0
1/9	69.4	82.8	84.6	80.9	91.6
1/10	69.6	82.4	77.8	79.8	91.6

从实验结果来看，各领域的分类精度基本都随训练集规模的减少而迅速下降，当使用全部训练集时，分类精度达到最佳。实验结果说明，通常情况下，足够大的训练集对于较高的分类精度具有决定性作用。

6 结论和展望

通过实验表明，在中文情感分类中，语料的语言风格对分类结果有一定的影响，DF 特征选择方法相对于 MI、CHI 和 IG 等特征选择方法并不占优；n-Gram 特征表示方法能产生良好的结果，而单一词性的词并不能反映网络评论的整体情感特征，整合更多情感表达的特征来进行分类，才可能提高分类精度；和其他分类方法比较，SVM 分类方法在精度方面具有明显的优势；在数据集一定的情况下，特征空间的维数并非越多越好，分类精度将在一定的维数达到最大值；通常情况下，足够大的训练集对于较高的分类精度具有决定性作用。总之，采用 n-Gram 特征表示方法、信息增益特征选择方法和 SVM 分类方法，在足够大训练集和选择适当数量特征的情况下，情感分类能取得较好的效果。

然而，我们也应看到，本文实验所使用的语料是从互联网上有选择性地获取的，其态度比较明确，用词较为规范，语言朴实简单。要将情感分类做到实用，其实还面临许多困难，例如：

1. 训练集标注的不一致。训练集文档中相类似的定性描述可能会标注成不同的类别，不同的标注者对同一评论所持的观点也不一致。

2. 测试集文档中存在正反态度并存现象以及对比描述手法。有一些评论中使用了带有否定态度的特征词，然而在结论句中却表达了较含糊的肯定态度，另外的一些评论在肯定一种产品的同时又否定另一种产品，因此很难区分它的评价核心，这给情感分类带来了噪音。

3. 数据的稀疏性。许多网站的评论都写得非常短，如笔记本电脑和手机评论一般都不超过 80 个汉字，这使得特征矩阵非常稀疏，选择特征的方法和质量对于分类的精度起着至关重要的作用。

4. 文档整体和文档段落的情感差异问题。影视类评论通常都写得较长，而其中的各个段落分别对某部电影各个方面进行了评价，这些评价往往有褒有贬，这也是本文实验中，影视领域分类精度较低的原因。

以上这些因素将成为情感分类问题的难点。为了改善分类的精度，还有很多问题亟待研究，如：准备一个标注更详细的训练集，训练集的标注最好能细到句子或词组这个层次，而不仅仅是标注某个文档为正面或负面；增大数据集的规模，目前可用的训练集和测试集规模都相对较小，只有在更大的数据集上训练并测试，才能更加接近实用的效果；尝试去除重复特征，寻找对于情感和产品属性选择更为有用的特征；把整个评论文档的类别和评论文档中段落的类别区分开，从每个文档段落的情感信息来判断整篇文档的情感。

参考文献：

[1] Franco Salvetti, Stephen Lewis, Christoph Reichenbach. Automatic Opinion Polarity Classification of Movie Reviews[J]. Colorado Research in Linguistics, 2004, Volume 17, Issue 1.

[2] Bo Pang, Lillian Lee, and Shivakumar Vaithyanathan. Thumbs up? Sentiment classification using machine learning techniques[A]. In: Proceedings of the 2002 Conference on Empirical Methods in Natural Language Processing (EMNLP), pages 79 86.

[3] Aidan Finn, Nicholas Kushmerick, and Barry Smyth. Genre classification and domain transfer for information filtering[A]. In: Fabio Crestani, Mark Girolami, and Cornelis J. van Rijsbergen, editors, Proceedings of ECIR-02, 24th European Colloquium on Information Retrieval Research, Glasgow, UK. Springer Verlag, Heidelberg, DE.

[4] Janyce Wiebe, Rebecca Bruce, Matthew Bell, Melanie Martin, and Theresa Wilson. A corpus study of evaluative and speculative language[A]. In: Proceedings of the 2nd ACL SIGdial Workshop on Discourse and Dialogue, 2001.

[5] Alina Andreevskaia and Sabine Bergler. Mining WordNet For a Fuzzy Sentiment: Sentiment Tag Extraction From WordNet Glosses [A]. In: Proc. EACL-06, Trento, Italy, 2006.

[6] Alistair Kennedy and Diana Inkpen. Sentiment Classification of Movie Reviews Using Contextual Valence Shifters[J]. Computational Intelligence, 2006, 22(2): 110-125.

[7] P. D. Turney and M. L. Littman. Unsupervised learning of semantic orientation from a hundred-billion-word corpus[D]. Technical Report ERB-1094, National Research Council Canada, Institute for Information Technology, 2002.

(下转第 108 页)

tion of ACM 18(11): 613-620 (1975).
[2] Willian B. Frakes, Ricardo Baeza-Yates. Information Retrieval Data Structures & Algorithms[M]. Prentice Hall PTR, New Jersey, 1992.
[3] 李雪蕾,张冬茉. 一种基于向量空间模型的文本分类方法[J]. 计算机工程,2003,29(17).
[4] 刘少辉,等. 一种基于向量空间模型的多层次文本分类方法[J]. 中文信息学报,2002,16(3): 8-14.
[5] 刘群,李素建. 基于《知网》的词汇语义相似度的计算[A]. 第三届汉语词汇语义学研讨会,台北,2002.
[6] 董振东,董强. "知网", http://www.keenage.com [DB/OL].
[7] 朱嫣岚,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.
[8] 王大亮,等. 基于HowNet构造语义场的方法[J]. 清华大学学报(自然科学版),2005,44(1).
[9] 金珠,等. 基于HowNet的话题跟踪及倾向性分类研究[J]. 情报学报,2005,25(5).
[10] 孙建涛,等. 网页分类技术[J]. 清华大学学报(自然科学版),2004,44(1).
[11] C. I. Barnes, L. Costantini, S. Perschke. Automatic Indexing Using the SLC-II System[J]. Information Processing and Management, 1978, 14 (2): 107-119.
[12] 韩客松,王永成. 一种用于主题提取的非线性加权方法[J]. 情报学报,2000,19(1).
[13] 秦兵,刘挺,等. 基于改进的贝叶斯模型的中文网页分类器[A]. 全国第六届计算语言学联合学术会议[C]. 2001.7.

(上接第94页)

参考文献:

[8] P. Subasic and A. Huettner. Affect analysis of text using fuzzy semantic typing[A]. IEEE-FS, 9:483 496, Aug. 2001.
[9] Hugo Liu, Henry Lieberman, and Ted Selker. A model of textual affect sensing using real-world knowledge[A]. In: Proceedings of the Seventh International Conference on Intelligent User Interfaces[C]. 2003. 125-132.
[10] Wei-Hao Lin, Theresa Wilson, Janyce Wiebe and Alexander Hauptmann. Which Side are You on? Identifying Perspectives at the Document and Sentence Levels[A]. In: Proceedings of the 10th Conference on Computational Natural Language Learning (CoNLL-X)[C]. New York City: June 2006. 109-116.
[11] Sebastiani F. Machine learning in automated text categorization[A]. ACM Computing Surveys, 2002, 34 (1):1? 47.
[12] 周茜,赵明生,扈旻. 中文文本分类中的特征选择研究[J]. 中文信息学报, 2004, 18(3): 17- 23
[13] Yang, Y. and Pederson, J. O. A comparative Study on Feature Selection in Text Categorization[A]. ICML 1997[C]. 412-420.
[14] Forman G. An extensive empirical study of feature selection metrics for text classification[J]. Journal of Machine Learning Research, 2003, 3 (1): 1533? 7928.
[15] P. Domingos and M. Pazzani. On the optimality of the simple Bayesian classifier under zero-one loss[J]. Machine Learning, 29:103-130, 1997.
[16] 代六玲,黄河燕,陈肇雄. 中文文本分类中特征抽取方法的比较研究[J]. 中文信息学报, 2004, 18(1): 26-32
[17] Casey Whitelaw, Navendu Garg and Shlomo Argamon. Using appraisal groups for sentiment analysis [A]. In: Proceedings of CIKM-05, 14th ACM International Conference on Information and Knowledge Management[C]. Bremen, DE. 625-631.