信息系统学报
第1卷第1辑：79－91

China Journal of Information Systems
Vol.1 No.1, Oct.2007：79－91

面向互联网评论情感分析的中文主观性自动判别方法研究*

叶强[1,2]，张紫琼[1]，罗振雄[2]
（1 哈尔滨工业大学，哈尔滨 150001，2 香港理工大学，香港）

摘　要　作为非结构化信息挖掘的一个新兴领域，网络评论情感分析引起了人们的极大兴趣。利用对互联网上客户评论信息的挖掘与分析结果，消费者可以了解其他用户的态度倾向分布，做出更好的购买决策。销售商和生产商可以获得用户对其商品和服务的反馈，了解用户对自己和对竞争对手的评价，从而改进产品、改善服务，获得竞争优势。实现情感分析的前提是有效识别文本中表达主观感情、态度和观点的内容，对文本中的主观性成分进行判断。目前，这一领域的研究主要针对英文评论进行。随着中文网络信息的不断发展，如何对中文网络评论进行自动情感分析已经成为一个迫切需要解决的问题。中文评论情感分析中的一个基础性的问题就是如何对中文评论的主观性进行判断。

本文提出了一种根据连续双词词类组合模式(2-POS)自动判断句子主观性程度的方法。在对主观句与客观句的分类试验中，两类语句的查准率和查全率都已接近目前英文同类研究的结果，初步表明了该方法的可行性。

关键词　互联网，评论，中文，主观性判别，N-POS模型

中图分类号　C931.6

1　前言

在互联网络无处不在的今天，网络上信息资源急剧增长。海量的信息虽然为人们带了丰富的信息资源，却也常让人们困惑不已。因为在庞杂的信息资源面前常常无从下手，不知所措。于是如何能快速有效地利用网络信息成为人们关注的焦点。搜索引擎的出现大大提高了互联网上信息搜寻的速度，当人们需要了解一些未知的信息时，就可以利用搜索引擎在互联网上进行搜索。通过采用新技术，未来的搜索引擎将可能提供一个全新的自动分析功能——情感分类(sentiment classification)，通过对客户评论的情感分类，我们将知道，在这些评论文章中，有多少人持正面态度，多少人持负面态度，这将帮助我们了解更多的用户对某种商品的态度倾向的分布，从而做出正确的购买决策。

情感分类主要通过分析客户对某种产品评论的文本内容，挖掘客户的情感倾向，从而自动将该文本判断为正面评论或负面评论。通过对大量客户评论的情感分类，其他客户可以做出是否应该购买某种产品的决策。销售商和生产商则可以利用自动情感分类，获得客户对其商品和服务的反馈信息，从而改进产品改善服务，获得竞争优势。不言而喻，在电子商务大潮席卷世界的今天，能够充分挖掘客户的情感信息，明白客户的喜好偏爱对商家来说具有重要的意义。

作为一个新的研究领域，情感分类目前正日益受到学术界和互联网企业界的关注。一些研究者已将开始对英文客户评论的情感分类进行研究，并取得了一定成果。现有的情感分类技术主要包括

* 国家自然科学基金项目资助(70501009,70771032)；香港理工大学研究基金资助(G-YX93)。
通信作者：叶强，哈尔滨工业大学管理学院副教授，e-mail：yeqiang2006@gmail.com。

机器学习方法及语义(semantic)方法两类。基于机器学习的情感分类方法在使用前需要大量的训练样本对分类模型进行训练,而训练样本集的建立则需要人对这些大量的评论文章进行逐一阅读,这与自动情感分类的目的产生矛盾。因此,许多研究者将情感分类研究的重点集中在对训练样本的需求量较低的语义方法上。情感倾向主要是通过主观句来表达的,因此,在基于语义的自动情感分析方法中,句子主观性的自动判断是一项基础性技术。目前在面向英文的研究中,研究者已经提出了一些自动判断句子主观性的方法,然而由于语言特点的差异,现有的用于英文客户评论情感分类的方法,无法直接用于中文客户评论的情感分类中。本研究将针对这一情况提出基于2-POS模型的中文句子主观性自动判别方法,为解决中文情感分析中的这一基础性问题,提供可用的方法。

该技术除可用于搜索引擎外,还可以用于企业的客户服务系统等需要进行情感分类的应用系统中,帮助企业深入挖掘互联网信息,提升竞争优势。客户可以利用其帮助自己做出正确的购买选择。而商家则可以根据其评价商品的市场反映,制订更加适合的产品策略,从而提升企业的竞争优势。

本文的其他部分结构为:第2部分介绍研究背景,第3部分提出研究方法,第4部分给出研究结果并进行讨论,第5部分是结论。

2 研究背景

随着Internet技术与应用在过去十几年时间中的快速发展,Internet不仅对企业的业务流程带来了巨大的变革,也对消费者的行为模式产生了深刻的影响。据统计,截至2007年1月,全球上网人数已达11.12亿人(Miniwatts Marketing Group,2007)。互联网已经成为人们最重要的信息来源之一(Dellarocas et al.,2005)。DoubleClick Inc.(2005)进行了一项针对美国服装业、计算机硬件设备业、运动与健身产品行业及旅游业网络客户的研究,发现这些行业中都有近一半以上的在线消费者在做出购买决定前会通过各种搜索引擎在互联网上搜索有关产品介绍以及其他客户对商品的评论等信息(图1),而其中互联网上的客户评论对于网络消费者的购买决策有着重要的影响(Ghose et al.,2006)。

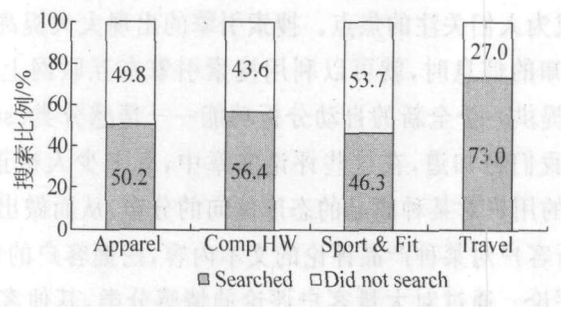

图1 消费者在购买前进行网络搜索的比例(DoubleClick Inc.,2005年)

在互联网络无处不在的今天,网络上信息资源急剧增长。尽管搜索引擎会帮助人们找到有关的信息,然而一般人很难逐一阅读搜索引擎找到的全部评论文章,往往只能阅读其中有限的部分,获得不完全的信息。针对这一问题,近年来以文本情感自动分析为目标的一项非结构化数据分析技术——情感分析(或情感分类,sentiment classification),引起了研究者的极大兴趣。通过对互联网上的客户评论进行自动情感分析,用户将可以了解其他用户的态度倾向分布,从而做出更好的购买决策。同时,对互联网上的这些评论的分析,也为企业获取客户信息提供了新的途径。通过对互联网上非结构化

的客户评论信息的挖掘与分析，可以使企业了解客户对企业及产品的情感倾向，从而为企业提供重要的商务决策信息。网络客户评论情感分析作为非结构化信息挖掘的一个新兴领域，主要涉及对互联网上的商品评论、新闻事件评论、服务反馈意见等非结构化信息的知识挖掘过程，正日益受到人们的关注(Bo Pang et al.，2002；Liu et al.，2005；Chaovalit and Zhou，2005)。目前已有的基于互联网的关于客户情感分析研究包括汽车、电影、股票、旅行目的地、电子仪器等产品评论的客户情感分类(Sanjiv Ranjan Das and Chen，2001；Theresa Wilson，2004；Turney P D，Littman M L，2003)。

作为一个新的研究领域，针对情感分析的研究仍处于起步阶段。一些研究者已经开始对英文客户评论的情感分类进行研究，并取得了一定成果。其中，Sanjiv Ranjan Das 等人(2001)针对 yahoo 网站股票留言板中的评论进行了研究，提取了投资者对其所关注股票的态度。Bo Pang 等(2002)研究者利用机器学习方法研究了文本形式评论的情感分析。Philip Beineke 等(2004)用机器学习和人的注释评论相结合，提高了英文文本情感分析的准确度。日本的 NEC 公司对产品声誉进行了语义抽象和分类的研究(Morinaga et al.，2002)，取得了初步的成功。匹兹堡大学的智能系统计划研究了情感分析中语义强度的提取问题(Wilson，2004)，对该领域研究做出了基础性的工作。Turney 等人(2002)通过基于语义方法的情感分类研究，得到汽车评论的准确率是 84%，电影评论的准确率是 66%，已经具有了初步的应用价值。Wiebe(2000)等从主观句与客观句分类的角度研究了客户情感分类，提出了主观句与客观句的分类方法。2004 年 Fei Zhongchao 利用机器学习方法，针对 sport. yahoo. com 英文体育评论研究了情感分析。

上述这些现有的情感分析技术，主要包括机器学习方法(Liu and Hu，2004；Fei et al.，2004；Beineke et al.，2004；Sanjiv Ranjan Das and Chen，2001)及语义方法(Semantic Orientation)(Turney，2002)两类。基于机器学习的情感分类方法在针对每一种产品使用前，都需要用大量的训练样本对分类模型进行训练，而训练样本集的建立则需要采用人工方法对大量的评论文章逐一阅读甄别，并进行手工标识，这与利用自动情感分类降低人的阅读负担这一初衷还有着一定的差距。因此，近来许多研究者将情感分析研究的重点集中在对训练样本的需求量较低的语义方法上(Chaovalit and Zhou，2005；Turney and Littman，2003；Turney，2002)。Turney(2001，2002)最早提出了基于 PMI-IR 算法的语义情感分类思想，该方法将点互信息(PMI)与信息汲取方法(IR)相结合，借助搜索引擎的后台数据库获得语义倾向信息，从而做出情感判断，该方法的可靠性已经在英文客户情感分类的研究中得到了初步的验证。

2002 年，Turney 针对汽车、银行、电影和旅游目的地等商品和服务的客户评论，用该方法进行了情感分类研究，获得了良好的分类效果。2003 年，Kushal Dave 等利用该方法对亚马逊(Amazon)和 C-Net 等网上商店的客户评论进行了情感分析，再次验证了该方法的性能。2005 年，Zhou 利用电影评论数据对基于语义倾向的情感分类方法和基于机器学习的情感分类方法进行了对比分析，发现语义方法的结果与机器学习方法具有相似性。上述研究均证实了该语义倾向的客户情感分析方法的有效性。除此之外，还有一些学者采用由普林斯顿大学开发的英文词网(WordNet)(Andreevskaia et al.，2006)进行英文语义方法的情感分析，也取得了较好的分析结果。

据 2007 年 1 月公布的中国互联网用户最新统计资料显示，截至 2006 年 12 月我国已经有互联网用户 1.37 亿人，排名居世界第二，并仍在以平均每年 23.4%的速度高速增长(CNNIC，2007)，中文用户和中文信息已经成为国际互联网上的一个非常重要的部分。然而，目前尚缺乏对中文客户评论情感分类的相关研究。由于语言结构的差别，现有的面向英文客户评论情感分类的语义方法，也无法直接用于中文客户评论的情感分类。

叶强，李一军等(2005，2006)探索了中文环境下的情感分析理论与方法(Ye et al.，2005a；Lin et

al.,2005; Ye et al.,2005b; Ye et al.,2006),在PMI-IR方法基础上,初步建立了中文语义倾向情感分析方法,并分别将中文搜索引擎www.Google.com和www.Baidu.com提供的API集成于情感分析实验平台中,对手机、图书、电影的中文客户评论进行了情感分析,获得了接近英文同类研究的分析结果,显示出了该方法在中文情感分析上的应用前景。另外,J.Yao等(2006)在研究中提出了使用电子汉英翻译词典结合英文词网的方法,也是对中文评论情感分析的一个有益尝试。

然而现有的这些面向中文评论情感分析的初步研究还存在很多不足。其中,借用英文做中介进行翻译的方法在技术上仍属英文情感分析的范畴,今后的研究方向还应直接针对中文进行情感分析。而叶强等提出的直接针对中文评论的情感分析方法,在情感模式的提取上,采用的仍是英文的情感模式词性组合(表1),还需要进一步针对中文的特点探索发现中文特有的情感模式组合,从而提高情感分类的可靠性。

表1 英文评论情感分析模式

模 式	首 词	尾 词
模式1	形容词	名词
模式2	副词	形容词
模式3	形容词	形容词
模式4	名词	形容词
模式5	副词	动词

在现有的基于语义的方法中,主观性模式的自动识别与判断是一项基础性技术,因为情感倾向主要是通过主观句来表达的。语言的"主观性"(subjectivity)是指在话语中多多少少总是含有说话人"自我"的表现成分。也就是说,说话人在说出一段话的同时表明自己对这段话的立场、态度和感情,从而在话语中留下自我的印记(Lyons,1977)。

Wiebe(2001)等针对英文主观情感识别进行了研究,选择某些词类(代词、形容词、序数词、情态动词和副词)、标点和句子位置作为特征,实现对主观句识别的平均准确率72.17%。Riloff(2003)等人利用boot-strapping算法学习得到了1052个主观性名词,单独使用主观性名词为特征,采用朴素贝叶斯分类器对主观句识别的查准率为77%,查全率为64%;如果加上先前确定的主观线索(来自词典和已有的研究结论)和句子的背景信息,那么分类器对主观句判断的查准率和查全率分别能达到81%和77%。Riloff和Wiebe(2003)进一步提出了从未经过人工标注的文本中自动提取主观句的方法。他们依靠先前研究中确定的主观特征,分别建立了主观分类器和客观分类器,自动从未标注的文本中获得大量主观句(查准率为91.5%,查全率为31.9%)和客观句,再从这些句子中得到更多主观性词语搭配,再用准确性很高的词语搭配更新原始的主观特征。通过重复上述过程进一步提高主观分类器和客观分类器的准确率,最终主观分类器的查准率和查全率分别达到90.2%和40.1%。Yu和Hatzivassiloglou(2003)利用三种统计方法进行主客观句的识别研究,包括相似性方法、朴素贝叶斯分类和多重朴素贝叶斯分类。其中,朴素贝叶斯分类器在原有研究的基础上采用词,2-gram、3-gram和词类,具有情感倾向的词序列,主语和其直接修饰成分等作为特征项,对主观句识别的查准率和查全率达到80%~90%,而客观句的查准率和查全率大约在50%左右。

目前尚缺乏针对中文的主观情感自动识别方法,已经成为中文情感分类研究与应用进一步发展的限制因素,并将影响到对互联网信息的有效应用,本文将对情感分类中的这一基础问题进行初步的探索。

3 研究方法：基于 N-POS 模型的中文主观性判别方法

3.1 N-POS 语言模型

为了使计算机能够自动处理自然语言文本，首先要对自然语言文本利用文本表示模型进行形式化描述。文本表示模型描述自然语言统计和结构方面的内在规律。文本的数学表示模型是文本信息处理的前提和基础。近几十年内，文本信息处理领域出现了许多文本表示模型，如布尔模型、概率模型、向量空间模型和以语料库为基础的统计语言建模 N-Gram、N-POS 模型等。为了简化问题的复杂度，本文提出了一种基于连续次模式的 N-POS 模型。

该 N-POS 模型的基础是将语句中的词按照其语法功能——词性(part-of-speech，简缩为 POS)进行分类，再用语句中连续 N 个词性的顺序组合作为一个项，对文本进行表示。考察一种简单情况 $N=2$，此时的语言模型称为 2-POS 模型。例如：

A：我非常喜欢听这首歌。

分词并词性标注后为“我(代词) 非常(副词) 喜欢(动词) 听(动词) 这(代词) 首(量词) 歌(名词)。(标点符号)”，则该语句的 2-POS 模型是“代词-副词副词-动词动词-动词动词-代词代词-量词量词-名词”，其中，“代词-副词”即为一个 2-POS 项，反映主观情感的 2-POS 项被称为 2-POS 主观模式。

3.2 基于 2-POS 模型的中文句子主观模式提取方法

本研究采用试验方法获取 2-POS 主观模式，该方法由以下 4 个步骤组成：

第 1 步：通过互联网获取大量包含主观句和客观句的中文语料。

第 2 步：采用人工复合标注试验方法筛选和标定主观性与客观性明确的句子。

本研究参考 Bruce 等(1999)对于英文句子主观性标注的方法，提出了针对中文的人工标注方法。通过合并多名参与试验的自愿者对语料句主客观性的判断，筛选出较明确的主观句和客观句。

具体做法是首先以口头结合书面文档的形式向自愿者介绍和讲解主观句和客观句的定义，并展示例句。然后请多位自愿者自行独立判断，将所有语料句标注为主观句或客观句，同时用 0 到 3 分表示这一判断的可靠性程度，3 分表示很肯定，0 分表示很不肯定。

通过上述方法得到主观句集合 Sb 和客观句集合 Ob，试验语料集 $Corp=Sb \cup Ob$

第 3 步：采用中文自动分词与词性标注算法对中文进行分词和词性标注。

利用中文自动分词和词性标注算法，对句子集合 Sb 和 Ob 中的语句进行分词和词性标注，并以句子为单位进行存储。本研究采用中国科学院计算机所软件室研究编写的中文分词工具 ICTCLAS(Institute of Computing Technology，Chinese Lexical Analysis System) (http://mtgroup.ict.ac.cn/~zhp/ICTCLAS/)，对句子进行分词及词性标注。

第 4 步：对每一个主观句子提取全部连续 2-POS 模式。

基于词性标注的结果，构造主观句集合 Sb 的 N-POS 语言模型。本研究提出的方法主要关注连续双词组合模式，并建立语句的 2-POS 描述模型。初始情况下，主观句的双词组合模式集合分别是 Sb 语句集合中出现的所有 2-POS 模式。

第 5 步：提取中文能有效表达情感的 2-POS 主观模式。

本研究提出的方法首先采用卡方公式(1)计算在集合 Sb 和 Ob 里出现的全部 2-POS 模式 χ^2 统计值(Yang，1997)，将所有 2-POS 模式根据 χ^2 值进行排序，选定 χ^2 阈值，从而获得中文 2-POS 主观

模式。

$$\chi^2(\text{pattern}_i, c_j) = \frac{N \times (A \times D - C \times B)^2}{(A+C)(B+D)(A+B)(C+D)} \quad (1)$$

其中，N 表示训练语料中的主、客观句总数，c_j 为某一特定类别，$j \in$ {主观，客观}。pattern_i 表示特定的 2-POS 模式，A 表示属于 c_j 类且包含 pattern_i 的句子频数，B 表示不属于 c_j 类但是包含 pattern_i 的句子频数，C 表示属于 c_j 类但是不包含 pattern_i 的句子频数，D 是既不属于 c_j 也不包含 pattern_i 的句子频数。

第 6 步：获得 2-POS 主观模式的主观性权重。

计算每一个 2-POS 主观模式在语料集 Corp 中的查准率与查全率，并将查准率作为该模式的主观性权重。

3.3 基于 2-POS 模型的中文句子主观程度计算方法

在获得中文 2-POS 主观模式的基础上，对中文句的主观性进行估计，本研究提出的估计方法如下（算法描述见表 2）：

（1）对给定待测试中文句子 S_i，首先应用分析工具进行分词并标注词性，提取该句中全部相邻词的词性组合，建立句子 S_i 的 2-POS 语言模型。

（2）对 S_i 中每一个 2-POS 项，如果它符合某个主观词类组合模式，那么把该模式的主观性权重加到整个句子的总得分 $Subj(S_i)$ 中。

（3）接下来，本文设计了一种句子主观性程度的度量方法，考虑到句子长度对总得分的影响，计算方法是用步骤 2 计算得到的句子总分除以句子 S_i 中 2-POS 的数量 num，其结果作为句子的主观性程度。

（4）给定一个阈值，如果句子的主观性程度大于阈值，那么判断其为主观句；反之，判断为客观句。

表 2 语句主观性程度的计算及主客观句判别

Input: Sentence S_i
Input: Extracted patterns $EP = \{\text{pattern}_j \mid j = 1, 2, \cdots, n\}$
Input: Threshold
Output: Sentence Subjectivity Score Subj(S_i) and Subjectivity Detection

```
Subj(S_i)=0
for i=1 to num
  for j=1 to n
    if 2-pos_i eq uals pattern_j then
      Subj(S_i)=Subj(S_i)+Pr(pattern_j)
    end if
  end for
end for
or Subj(S_i)=Subj(S_i)/num
if Subj(S_i) > Threshold then
  S_i is classified subjective.
else S_i is classified objective.
end if
```

3.4 语料数据

本研究的语料库使用的中文文本数据均来自新华网 http://www.xinhuanet.com/,其中主观性文本主要取自新华网论坛和评论,客观性文本主要来自新华网上的各种报道文章。无论是评论还是报道文章,其内容都涵盖了很多不同主题。研究者初步从评论和论坛上获得了 2133 个句子,从报道性文章中获得了 2462 个句子。四位不同专业背景的成年人作为自愿者参加了本文的研究工作,他们的任务是根据本研究的《标注说明》独立地判断这些句子的主客观性。

研究者向四位自愿者提供了标注说明文档,并简要地为这四位标注人讲解了《标注说明》,要求他们判断每个句子的主客观性,分别用 S(主观)和 O(客观)表示,同时用数字 0,1,2,3 表示对判断的不确定程度,0 表示不确定,3 表示非常确定。

表 3 至表 5 以三种数据组织方式显示了其中两位标注人的标注结果。表 3 是在不考虑确定程度的情况下 Judge1 与 Judge2 的标注结果。表 4 将不确定程度 0 和 1,2 和 3 结合。表 5 列出了与各种不确定程度相对应的标注结果。表格的行是 Judge1 对主客观句的分类,列是 Judge2 对主客观句的分类。n_{ij} 表示 Judge1 判断为 i 而 Judge2 判断为 j 的句子数量。

表 3 Judge1 和 Judge2 的标注结果(两类表格)

		Judge 2		
		Subj	**Obj**	
Judge 1	Subj	$n_{11}=2131$	$n_{12}=385$	$N_{1+}=2516$
	Obj	$n_{21}=262$	$n_{22}=1817$	$N_{2+}=2079$
		$n_{+1}=2393$	$n_{+2}=2202$	$N_{++}=4595$

表 4 Judge1 和 Judge2 的标注结果(四类表格)

		Judge2				
		$Subj_{0,1}$	**$Subj_{2,3}$**	**$Obj_{2,3}$**	**$Obj_{0,1}$**	
Judge1	$Subj_{0,1}$	58	208	57	249	572
	$Subj_{2,3}$	19	1846	5	74	1944
	$Obj_{2,3}$	28	13	1698	62	1801
	$Obj_{0,1}$	163	58	45	12	278
		268	2125	1772	430	4595

表 5 Judge1 和 Judge2 的标注结果(八类表格)

		Judge2								
		$Subj_0$	**$Subj_1$**	**$Subj_2$**	**$Subj_3$**	**Obj_3**	**Obj_2**	**Obj_1**	**Obj_0**	
Judge 1	$Subj_0$	16	21	12	35	6	20	82	110	302
	$Subj_1$	12	9	45	116	17	14	17	40	270
	$Subj_2$	3	10	105	442	0	4	14	45	623
	$Subj_3$	1	5	54	1245	1	0	1	14	1321
	Obj_3	5	12	8	1	484	132	12	21	675
	Obj_2	4	7	4	0	611	471	4	25	1126
	Obj_1	28	12	7	5	28	5	5	2	92
	Obj_0	52	71	31	15	9	3	1	4	186
		121	147	266	1859	1156	649	136	261	4595

表3～表5显示，标注者对“主观性”的理解并不完全一致。在表3中，边缘总和 n_{i+} 和 n_{+j} 显示Judge1比Judge2更倾向于把句子判断为主观，而当把句子判断为主观句时，Judge2比Judge1更加确定(见表4和表5)。我们从原始语料库中去掉那些标注者理解偏差较大的句子，(1)首先去掉既被标注为S(主观)又被标注为O(客观)的句子，标注者对这类句子的主客观性争议较大；(2)在余下的句子中，如果一个句子被标注的不确定性0与1的个数之和大于等于2，这个句子也被去掉。

按照上述方法，我们最终得到四位标注者判断比较一致的1752个主观句和1603个客观句，然后分别随机抽出800个主观句和800个客观句构成训练样本。另外随机选取的800个主观句和800个客观句作为测试样本。对语料库中的句子进行分词和词性标注后，本文分析了语料库的总体统计特征。表6显示了主、客观语料分别包含的句子数量、总词数和平均句长。

表6 语料库总体统计特征

	主观语料	客观语料
句子数	1600	1600
总词数	39039	37040
平均句长	24.399375	23.15

表7列出了两类语料中各词类出现的频率，以及同一个词类在两个语料库中的使用频率是否存在显著差异。本文采用 χ^2 值计算词类频率的偏高程度。结果表明在主观文本中，形容词、副词、代词和语气词的频数明显偏高；在客观文本中，数词、量词、名词、方位词和时间词出现的频率较高。统计结果与不同词类的语义功能是非常吻合的。另外，由于拟声词和前缀出现频次很低，χ^2 值已经失去统计意义。

表7 词类数量及显著性统计

词 类	主观语料	客观语料	χ^2
介词	1386	2061	3.8260
动词	10618	8613	2.3103
形容词	2667	1356	8.5115
数词	913	1825	8.2010
量词	674	1387	6.5277
名词	9515	12033	8.7401
代词	2604	1324	8.3065
副词	4085	2204	10.9314
方位词	491	1026	4.8406
时间词	281	834	6.7848
连词	1301	1080	0.1527
助词	567	767	0.7443
拟声词	3	4	5.0789
后缀	103	69	0.0114
处所词	145	276	0.6646
叹词	8	0	3.0853
前缀	0	1	35.0675
字符串	99	86	0.1276
语气词	495	49	7.5969
状态词	42	26	0.2209
的(助词)	2918	1958	3.1936
地(助词)	53	43	0.2567
得(助词)	71	18	0.0477
标点符号	5999	5148	0.5462

4 结果与讨论

4.1 中文 2-POS 主观模式结果

在主观句训练样本 Sb 中共发现了 254 个 2-POS 模式，按该模式在集合 Sb 和 Ob 中出现的差异性统计量 χ^2 值进行排序后表 8 列出了其中前 30 个模式，并分别计算了在测试样本中对主观句识别的查准率与查全率。

表 8 主观词类组合模式及例句

编号	首词	尾词	主观语料	客观语料	CHI	查准率	查全率
1	副词	形容词	272	94	112.2442	0.743169	0.34
2	副词	动词	603	438	74.85578	0.579251	0.75375
3	动词	代词	274	125	74.12703	0.686717	0.3425
4	形容词	的(助词)	242	103	71.39812	0.701449	0.3025
5	动词	副词	224	96	64	0.7	0.28
6	动词	形容词	231	105	59.81013	0.6875	0.28875
7	动词	语气词	77	10	54.56465	0.885057	0.09625
8	形容词	语气词	59	4	49.98399	0.936508	0.07375
9	代词	动词	275	153	47.47536	0.642523	0.34375
10	名词	语气词	51	2	46.85392	0.962264	0.06375
11	代词	副词	160	67	44.40067	0.704846	0.2
12	代词	形容词	75	19	35.44404	0.797872	0.09375
13	副词	副词	181	97	30.71866	0.651079	0.22625
14	的(助词)	副词	44	11	20.50485	0.8	0.055
15	代词	的(助词)	119	62	20.23992	0.657459	0.14875
16	代词	代词	33	6	19.15932	0.846154	0.04125
17	动词	量词	39	10	17.7055	0.795918	0.04875
18	形容词	副词	34	8	16.52913	0.809524	0.0425
19	代词	名词	190	127	15.61404	0.599369	0.2375
20	的(助词)	动词	178	117	15.4649	0.60339	0.2225
21	动词	得(助词)	26	5	14.50688	0.83871	0.0325
22	动词	动词	534	469	11.28941	0.532403	0.6675
23	连词	代词	71	40	9.303057	0.63964	0.08875
24	形容词	形容词	38	16	9.276029	0.703704	0.0475
25	得(助词)	副词	9	0	9.050911	1	0.01125
26	副词	助词	13	2	8.143007	0.866667	0.01625
27	得(助词)	形容词	13	2	8.143007	0.866667	0.01625
28	连词	副词	85	53	8.1207	0.615942	0.10625
29	副词	名词	27	10	7.995712	0.72973	0.03375
30	副词	代词	18	5	7.454992	0.782609	0.0225

4.2 模式分类效果比较

文本的“主观性”实际上是一个程度问题，于是主观句与客观句的分类就是一个确定分类阈值的问题。

本研究通过实验，讨论了选取不同分类阈值对主观句识别效果的影响。实验中，本研究根据卡方值选取了前20个2-POS主观模式建立了文本分类器。图2给出了处在不同主观性程度区间内的测试语句数量。图3表示随着分类阈值的改变，主客观句子分类的查准率和查全率情况。根据不同需要人工设置分类阈值，通过人工改变阈值可以影响分类器的输出结果，从而满足不同需求。

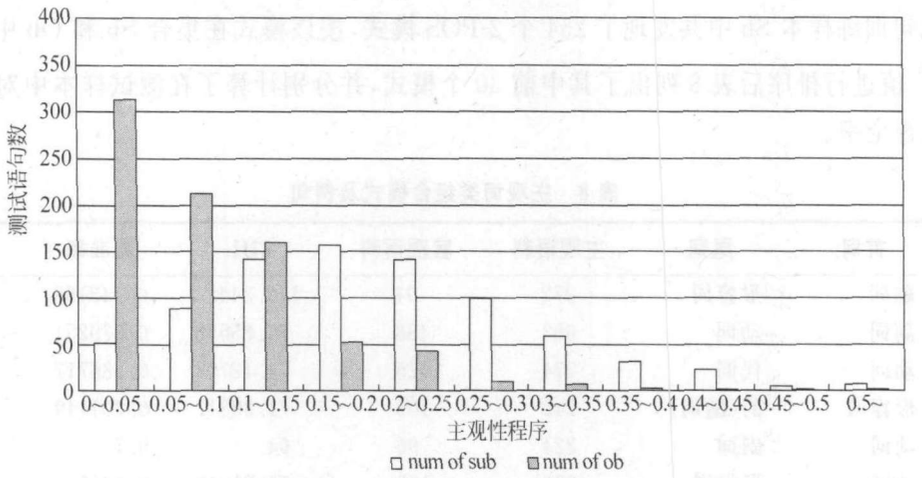

图2 处在不同主观性程度区间内的测试语句数量

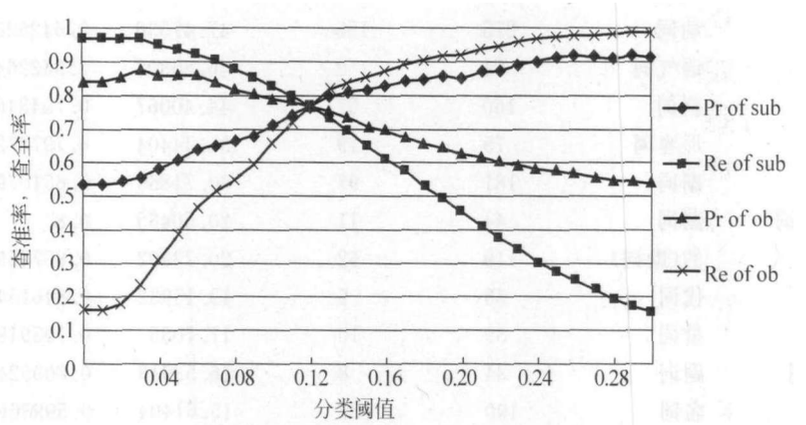

图3 随着分类阈值的改变，主客观句子的查准率和查全率情况

在阈值设定为0.12时主观文本的分类查准率和查全率均达到了76%，已经接近了英文同类研究的结果，表明了该方法的适用性。

5 结论

中文文本主观成分的自动判断，是面向互联网评论的情感分析中的一项基础性关键技术。尽管这一问题在英文中已经开展了一定的研究，然而在针对中文情感分析的研究目前仍很不足。本文从词类角度对中文文本的主观特征进行了一定程度的概括，初步探索了中文文本的主观性自动识别的方法。

本研究首先在N-POS语言模型的基础上，利用CHI统计方法提取中文主观文本词类组合模式，提出了主观文本词类组合模式提取方法，建立了中文双词主观情感词类组合模式2-POS模型，初步完成了中文文本情感分析中的一项基础工作。

本研究基于所建立的2-POS模型进一步提出一种利用加权后的主观词类组合模式计算语句主观性程度的新算法，实现主观与客观文本的自动分离。在语料实验中，采用20个2-POS模式的分类器对主观句判断的查准率和查全率都达到了76%左右，从而表明了本文提出的语句主观程度度量方法是可行的。

该方法一方面可以用于评论性与非评论性互联网信息的自动分离，从而实现互联网评论的自动获取与分析。另外，本文提出的方法也可以实现文本主观特征的提取，从而根据这些集中反映主观情感的文本成分对文本的情感倾向进行判断。

对今后的研究将可以从以下3个方面展开：(1)进一步探索词类组合模式主观性程度权重的设定。本文采用主观模式的准确率作为模式的权重，可以进一步探索其他设置方法。(2)对词类标注结果进行修正。现有分词工具的词类自动标注结果还存在许多错误，词类标注结果对于词类组合的提取及句子主观程度的判断会产生较大影响，最好依照某个标准在自动标注的基础上进行人工修正。(3)对词类进行更细致的划分，以便表示更精确的主观情感词类组合模式。如代词可划分为三个子类人称代词、指示代词和疑问代词，其中疑问代词表达的主观性更强。

参考文献

[1] Alina Andreevskaia, Sabine Bergler. Mining Word Net for Fuzzy Sentiment: Sentiment Tag Extraction from Word Net Glosses. Proceedings of the 11th Conference of the European Chanpter of the ACL (EACL'06), April, 2006: 209-216.

[2] Bin Lin, Tao Lu. Qiang ye. Opinion Classification for Chinese Movie Reviews. Proceeding of 12th International Conferecn on Management Scince and Engineering, July, 2005, Korea.

[3] Bing Liu, Minqing Hu. Mining. summarizing customer reviews. In: Proceedings of the 2004 ACM SIGKDD international conference, 2004.

[4] Bo Pang, Lillian Lee, Shivakumar Vaithyanathan. "Thumbs up? Sentiment Classification using Machine Learning Techniques," presented at the 2002 Conference on Empirical Methods in Natural Language Processing (EMNLP'2002), 2002: 79-86.

[5] Chevalier, J., D. Mayzlin. The effect of word of mouth online: Online book reviews. Journal of Marketing Research, forthcoming. 2006.

[6] Dave K, Lawrence S, Pennock D M. Mining the Peanut Gallery: Opinion Extraction and Semantic Classification of Product Reviews. In Proceeding of 12th international conference on World Wide Web. Budapest, Hungary: ACM Press, 2003: 519-528.

[7] Dellarocas, C., N. Awad, M. Zhang. Using online ratings as a proxy of word-of-mouth in motion picture revenue forecasting. Working Paper, University of Maryland, 2005.

[8] Ellen Riloff, Janyce Wiebe. Learning Extraction Patterns for Subjective Expressions. Proceedings of the 2003 Conference on Empirical Methods in Natural Language Processing, EMNLP-03, 2003.

[9] Fei Zhongchao, Liu Jian, Wu Gengfeng. Sentiment Classification Using Phrase Patterns. In: Proceedings of the Fourth International Conference on Computer and Information Technology (CIT'04). WuHan, China: IEEE, 2004: 1-6.

[10] Ghose, A, P. Ipeirotis. Towards an Understanding of the Impact of Customer Sentiment on Product Sales and Review Quality. Proceedings of the Workshop on Information Technology and Systems (WITS), Milwaukee, December 2006.

[11] Hong Yu, Vasileios Hatzivassiloglou. Towards Answering Opinion Questions: Separating Facts from Opinions and Identifying the Polarity of Opinion Sentences. In: Proceedings of the 2003 Conference on Empirical Methods in Natural Language Processing, 2003.

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

[13] Jianxin Yao, etc. Using Bilingual Lexicon to Judge Sentiment Orientation of Chinese Words. Proceedings of The Sixth IEEE International Conference on Computer and Information Technology (CIT'06).

[14] Liu B, Hu Minqing, Cheng Junsheng. Opinion Observer: Analyzing and Comparing Opinions on the Web. In: Proceedings of the 14th international World Wide Web conference (WWW-2005). Chiba, Japan: ACM Press, 2005: 10-14.

[15] Miniwatts Marketing Group. Internet Usage Statistics-The Big Picture World Internet Users and Population Stats. http://www.internetworldstats.com/stats.htm2007/3/1).

[16] Philip Beineke, Trevor Hastie, Shivakumar Vaithyanathan: The Sentimental Factor: Improving Review Classification Via Human-Provided Information. Proceedings of ACL 2004: 263-270.

[17] Pimwadee Chaovalit, Lina Zhou. Movie Review Mining: a Comparison between Supervised and Unsupervised Classification Approaches. In: Proceedings of the 38th Annual Hawaii International Conference on System Sciences, 2005.

[18] Qiang Ye, Yijun Li, Yiwen Zhang. Semantic-Oriented Sentiment Classification for Chinese Product Reviews: an Experimental Study on the Reviews for Books and Cell Phones. Tsinghua Science and Technology, 2005, 10(10).

[19] Qiang Ye, Bin Lin, Yijun Li. Sentiment Classification for Chinese reviews: a Comparison Between SVM and Semantic Approaches. The 4th International Conference on Machine Learning and Cybernetics ICMLC2005 (IEEE), Aug, 2005.

[20] Qiang Ye, Wen Shi, Yijun Li. Sentiment Classification for Movie Reviews in Chinese by Improved Semantic Oriented Approach. In: Proceedings of the 39th Annual Hawaii International Conference on System Sciences, 2006.

[21] Sanjiv Ranjan Das, Mike Y. Chen. Yahoo! for Amazon: Sentiment parsing from small talk on the web. In Proceedings of the 8th Asia Pacific Finance Association Annual Conference, 2001.

[22] Satoshi Morinaga, Kenji Yamanishi, Kenji Tateishi, Toshikazu Fukushima. "Mining Product Reputations on the web." Proceeding of the 8th ACM SIGKDD international conference on Knowledge discovery and data mining, Edmonton, Alberta, Canada, 2002.

[23] Theresa Wilson. Just how mad are you? Finding strong and weak opinion clauses. Proceedings of AAAI, 2004: 761-769.

[24] Turney P D, Littman M L. Measuring Praise and Criticism: Inference of Semantic Orientation from Association, ACM Transactions on Information Systems, 2003, 21(4): 315-346.

[25] Turney P D, Thumbs Up or Thumbs Down? Semantic Orientation Applied to Unsupervised Classification of Reviews. Proceeding of Association for Computational Linguistics 40th Anniversary Meeting. Philadelphia, PA, USA: ACL, 2002: 417-424.

[26] Turney, P. D. Mining the Web for synonyms: PMI-IR versus LSA on TOEFL. In: Proceedings of the Twelfth European Conference on Machine Learning. Berlin: Springer-Verlag, 2001: 491-502.

[27] Wiebe J M. Learning Subjective Adjectives from Corpora. In: Proceeding of 17th National Conference on Artificial Intelligence. Menlo Park, California: AAAI Press, 2000: 735-740.

[28] Yang Yiming, Pederson J O. A Comparative Study on Feature Selection in Text Categorization[A]. Proceedings of the 14th International Conference on Machine Learning[C]. Nashville : Morgan Kaufmann, 1997: 412-420.

[29] E. Riloff, J. Wiebe, T. Wilson. Learning Subjective Nouns using Extraction Pattern Bootstrapping. Proceedings of the Seventh Conference on Computational Natural Language Learning (CoNLL-03). 2003.

[30] Rebecca Bruce, Janyce Wiebe. Recognizing subjectivity: A case study in manual tagging. Natural Language Engineering. 1999, 5(2): 187-205.

[31] J. Lyons. Semantics. Cambridge: Cambridge University Press. 1977.

[32] 杨惠中. 语料库语言学导论. 上海外语教育出版社, 2002: 159-161.

[33] 中国互联网信息中心(CNNIC). 中国互联网络发展状况统计报告. 北京: 2007.1: 30-31.

Automatically Measuring Subjectivity of Chinese Sentences for Sentiment Analysis to Reviews on the Internet

YE Qiang[1,2], ZHANG Ziqiong[1] & Law Rob[2]
(1. Harbin Institute of Technology, Harbin, 150001
2. Hong Kong Polytechnic University, Hong Kong)

Abstract As a new domain of unstructured data mining, online sentiment analysis has aroused great interest recently. Through automatically mining online reviews to certain products, consumers could know the distribution of attitudes of other consumers to this product before making a buying decision. Meanwhile, manufacturers and retailers would get consumers' feedback about their products or services, as well the opinion of customers to their competitors, which is useful for them to improve the products or services and gain competitive advantages. A crucial step before sentiment analysis is to identify subjective expressions in the context. Subjective sentences are usually the parts expressing sentiment, attitude or opinions in text. The existing researches on sentiment analysis mainly focus on English reviews. Few studies have been conducted to Chinese texts. As Chinese information has increased dramatically in cyber space, how to automatically analyze opinions of reviews in Chinese on the Internet has become urgent. One basic and important task is to establish a method to identify subjective expressions in Chinese reviews. This paper proposed an approach to measure subjective strength of Chinese sentences using patterns of continuous two words combination, 2-POS model. In experiments of subjectivity classification to Chinese sentences, both subjective and objective sentences achieved high precision and recall. The results show that the performances of the proposed approach are comparable to existing studies in English.

Key Words Internet, Review, Chinese, subjectivity detection, N-POS model