中西思维模式对于情感倾向性的影响

杨 亮 林鸿飞 王宇轩 许 侃

(大连理工大学计算机科学与技术学院 大连 116024)

摘 要 目前关于情感倾向性的研究均以语言内容为主要对象,然而民族的思维模式影响了语言的表达形式。通过对汉民族"图形式"、"具体性"、"散点视"的思维特点和西方民族"直线式"、"抽象性"、"聚点视"的思维特点进行对比研究,提出了基于中西思维模式的情感倾向性分析方法。针对"图形式"和"直线式"思维特点,给出了位置相关的量化方法;针对"具体性"和"抽象性"思维特点给出了基于词性和语法特征的分析方法;针对"散点视"和"聚点视"思维特点提出了基于视窗的分析方法。实验结果表明,在中英文语料中考虑各自的思维模式特点有助于提高情感倾向性识别的效果。

关键词 情感倾向性分析,中西思维模式,视窗技术

中图法分类号 TP391 **文献标识码** A

Influence of Chinese and Western Thinking Modes on Sentiment Analysis

YANG Liang LIN Hong-fei WANG Yu-xuan XU Kan

(School of Computer Science and Technology, Dalian University of Technology, Dalian 116024, China)

Abstract Sentiment analysis is an important research domain of affective computing, and researchers focus on the context in this field. Because the thinking mode influences the formation of language, this paper researched and compared the Chinese thinking modes and western thinking modes. The former contains "spiral graphic mode", "concreteness" and "scattered view", and the latter owns "straight line mode", "abstractness" and "focus view". A position-related method was proposed to quantify the "spiral graphic mode" and "straight line mode", and then according to the "concreteness" and "abstractness", an approach was implemented based on part of speech and grammar. Finally a view-window method was applied to analyze the "scattered view" and "focus view". The experimental results show that the performance of sentiment analysis including the thinking mode factors is better than that not including.

Keywords Sentiment analysis, Chinese and western thinking mode, View-window

1 引言

情感倾向性分析是情感计算领域中一个重要的组成部分,其主要任务为识别出文本对事物的褒贬态度以及作者自身喜怒哀惧的情感倾向。目前国内外针对情感倾向性分析的研究主要针对文本的内容进行,其中以词汇级和句子级为主,文献[1-4]等做了相关的工作和研究,其中文献[1]认为词汇级的特征是识别分析句子级情感的关键,该文以层级为前提假设,用以划分词汇特征及其相关关系,使用分析工具进行分析以减少特征集合数量达到最优特征集的作用,并在3种分类任务上均达到较好的效果。文献[2]利用条件随机域(CRF)实现了对某一领域或跨领域的评价对象抽取,并在跨领域评价对象抽取方面F值有了较大的提高。文献[3]提出了一种基于bootstrapping的方法,通过对产品的属性及其评价词建立二部图关系,实现了对标准产品评论的倾向性分析,并在实验中得到了很好的验证。在文献[4]中通过识别并分析语料中的资源名词,如"水","石油"等,与消耗类动词建立关联,实现了对自身不含有倾向性的词语在特殊语境下的倾向性分析,实验结果显示该方法也能取得较好的效果。

句子级别的情感分析方法中基于词典的情感计算方法都需要构建相应的情感词典,其中通过人工方法构建的见文献[5-8]。但是这种方法不是被单独使用,其中基于情感词典分析扩充词典的研究有文献[5,7]。文献[5]使用语法规则和语言学规则初步过滤情感句,然后通过抽取特性词、共现词、典型句子特征和对相应类别的一致性分析特征这4类判断句子的极性,最终得到产品的声誉度。而文献[7]通过情感词典和情感模板数据库,对在线的文档抽取主题特征集、情感特征集以及情感主体间关系,目前在产品评论方面有很好的应用。文献[8]采用的是种子词,利用Wordnet等外部资源增加词典内容,该种方法的缺点在于不能有效地解决领域倾向性词汇问题,例如对于电话的"无声"为negative,汽车的"无声"为positive,该文献不能很好地解决。

到稿日期:2011-12-03 返修日期:2012-02-05 本文受国家自然科学基金(60673039,60973068),国家社科基金(08BTQ025),教育部博士点基金(20090041110002)资助。

杨 亮 男,博士生,主要研究方向为情感计算和观点挖掘;林鸿飞 男,博士,教授,博士生导师,主要研究方向为搜索引擎、文本挖掘和自然语言理解,E-mail:hflin@dlut.edu.cn;王宇轩 男,硕士生,主要研究方向为观点挖掘及情感分析;许 侃 男,博士生,主要研究方向为文本挖掘。

国内的研究工作集中在词汇级别及句子级别上，主要表现在国内的评测任务设置上，将任务集中安排在了情感词及情感句的识别上，关于篇章级别的任务为探索性任务。例如COAE2008[9]（中文倾向性分析评测）以及COAE2009[10]中，任务要求自动识别出在一定的上下文环境中抽取出的能够明确表达人物情感的词语和句子，并判断该情感词和情感句所属的类别。针对情感词的研究，谭松波[11]针对不同领域的特点，建立了领域内文本的词与词之间的语义网络、词与文档之间的语义网络，以及词与领域外的文档的语义网络，最终得到适用于目标领域的情感词典。

通过对情感倾向性相关工作的调研，发现目前篇章级的倾向性研究较少。有代表性的工作有Turney对电影评论的分类[12]，Turney的方法是将文档中词汇和短语的倾向性进行平均，来判断文章的倾向性。这种方法基于情感倾向性词典，不需要人工标注文本情感倾向性的训练语料。Liu Bing等人基于整体词典的方法进行观点挖掘[13]，其是对传统方法的改进。基于词典的方法过于依赖词典，需要良好的词典做支撑才能有很好的结果。谭松波等人针对不同的领域建设词典[14]，虽然能够提供较好的词典，但针对不同的领域要先建立相应的词典，才能保证一定的准确性，其缺乏灵活性。徐琳宏的基于语义资源的文本情感分析[15]，是利用CRF对文章逐句进行情感标注得到文章的情感链，进而判断文章的倾向性。目前针对篇章级别的情感倾向性工作不是很多，主要是由于篇章级别文本内容丰富，分析效果不理想。但是Web2.0以后，随着用户生成内容的增多，博客和微博成为主要的情感表达的方式，因此对篇章级的情感分析工作要给予关注。

上述的研究均以语言内容为主要研究对象，但是忽略了隐藏在语言形式下的思维模式的影响因素。本文从研究中西方文化的差异出发，着重考虑思维模式差异，根据前人对中西方民族思维的研究，总结出汉民族以及西方民族在思维模式上各自的特点，并针对不同的特点分别给出了相应的量化思维模式的方法。将本文提出的方法与Turney基于词典的方法进行对比，从中英文两个语料集上进行实验得到的结果上看，本文提出的中西思维量化方法均在各自的语料集上体现出各自的优势。本文第2节介绍并分析汉民族思维以及西方民族思维特点及差异；第3节针对汉民族思维和西方民族思维各自的特点，给出了相应的量化方法；第4节介绍实验设计方法以及实验结果；最后为总结。

2 中西思维特点及差异

关于思维是什么，《现代汉语词典》给出了这样的概括："思维，即人脑对客观现实的反映过程。具体的说，它是在表象、概念的基础上进行分析、综合、判断、推理等认识活动的过程。它是人类特有的一种精神活动，是从社会实践中产生的"。从上述定义中可以看到，思维是人脑的一种机能，而语言则是在人们思维过程中不断形成的，所以思维模式决定语言的形式，不同文明决定思维模式，思维模式又决定了语言形式。汉民族和西方民族的文明属于两种截然不同的文明，前者属于大陆文明，而后者属于大洋文明，这两种截然不同的文明下形成的思维模式以及语言形式也存在着不同。下面将具体给出汉民族以及西方民族各自不同的思维特点[16]，及相应的量化方法。

2.1 图形式与直线式

"图形式"是汉民族思维的一个特点，指汉民族在语篇的结构上多为归纳式，先对细节进行阐述，举例说明，然后得出结论，呈现出螺旋形的图形结构，同时语篇文本中的段落结构不紧凑，在归纳总结部分，常缺少一个明确表达文章主旨的语句，文章主题往往需要靠读者在读懂文章后给出相应的归纳总结。在这种思维模式下，中文文本给人一种委婉、含蓄的感觉。相反地，英文文本则给人一种直接、明显的感觉，概括这种思维模式为"直线式"思维模式。该模式下的英文文章经常遵循"三段论"模式，并采用演绎的方法，每个段落都以主题句的形式给出段落的主旨，其他各句围绕着主旨句进行阐述，内容紧扣主题且不含与主题无关的内容，体现了英语写作的严谨性和单一性原则。

目前中国人的思维受西方文化的影响，其语言形式趋向于短小、直接。在简短的文本中，中文文本中更体现出西方民族的思维方式，但是从篇章级别的文本结构中看，汉民族的图形式的思维方式显现得比较明显，因此在对中文篇章级别的文本情感倾向性分析中应该着重考虑"图形式"思维特点。而西方民族的文化未受到其他文化的深刻影响，基本保留着其自身的思维方式，所以在分析英文文本的情感倾向性中，不论文本规模的大小，皆可按照西方民族"直线式"思维方式去分析。

2.2 具体性与抽象性

汉语侧重于形象思维，是一种直觉思维，侧重具体思维，多借助其他事物来表达作者的意图，汉民族的形象思维使得汉语偏重用直观、动态的动词，例如："望梅止渴"、"画饼充饥"用来表达"用空想安慰自己"的意思。而西方民族侧重逻辑思维，抽象思维则使英语倾向于大量使用抽象名词和介词或其它词类、短语或结构，善于抽象思维，因此在英语句中大量使用抽象名词。这些词语大多是通过虚化手段从其他词类派生而来的，如：表示"性质、状态"的后缀有"-ness"，"-tion"，"-sion"，"-ance"，"-hood"等。从上述分析可以看出汉语的动词优势和英语的名词优势。如：Where there is a will, there is a way，即有志者，事竟成。中文使用动词来表达，而英文使用两个名词来表达该含义。

这两种思维模式特点不限于文章的篇幅，短到一个句子，长到一篇文章，都能够体现汉语动词优势以及英语名词优势，故本文认为在针对不同的语言的文本进行情感倾向性分析时，无论文章的长短，均需要相应地考虑"具体性"和"抽象性"的特点，以帮助提高情感倾向性分析的效果。

2.3 散点视与聚点视

从本体论角度比较，中国人受《易经》哲学的影响，注意整体统一，强调从多归一的思想[17]。故中文文本中体现出"散点视"的思维特点，即在汉语表达中为使用多个表达同样含义的词汇，甚至是多个名词、形容词或数词充当谓语，例如在中文中常使用"风声鹤唳，草木皆兵"两个成语来表达受惊吓的程度。由于中文的表达极其丰富，因此在抒发自己的情感或者评论的时候，若单一的词语不足以表达作者的情绪，则采用多词语重叠使用的方式来加强对情感的表达。与中文表达不同，英文则侧重于分析原则，注重逻辑推理，强调由一到多的思想，倾向于简明地表达作者的情感或者是观点，所以在需要表达作者情感或者观点时，其使用单一的词汇来表达一句话

的含义，体现出“聚点视”的思维特点。

“散点视”中的“散”并不是指使用的表达情感的词汇均匀地出现在句子或者篇章中，而是体现为在需要表达情感的位置上使用多个词汇来表达同一个意思。而英文的“聚点视”则表现为集中使用一到两个词在某个位置上来表达情感。

本文总结了部分汉民族以及西方民族在思维模式上的各自的特点；在了解了各自的思维模式的特点后，针对上述的3个思维特点给出相应的量化方法。

3 中西思维模式计算方法

3.1 外部资源

本文在研究文本情感倾向性分析时，发现民族思维模式因素对情感倾向性分析存在辅助作用，故本文以基于词典的基本方法为基础，针对英汉民族思维模式的不同特点，给出了相应的量化计算方法，旨在说明在基于词典的情感倾向性分析的方法中考虑思维模式的因素，有助于提高该类方法的情感倾向性分析效果。

本文以基于词典的方法为基础，对情感分析方法提出相应的改进。为了满足词汇的丰富性，本文借鉴了大连理工大学的情感词汇本体库[18]，文献[7]中提及的情感分类方法是从认知心理学的角度定义了7大类情感和20小类情感。由于近些年的国内的中文评测任务，例如：COAE（Chinese Opinion Analysis Evaluation），将情感分为喜、怒、哀、惧4类，因此本文将文献[7]中的7大类情感进行了修改，将其分为4类表示情感的词，即喜、怒、哀、惧，以及两大类表示评价的词，即褒、贬。确定了6大类的分类标准，同时为了保证词汇情感表达的准确性，人工选取了具有明显情感倾向的16585个词汇，以该词典作为本文使用的词典。在对英文文本进行情感分析时，使用的词典为HowNet，其中情感分为6类，分别为anger，disgust，fear，joy，sadness，surprise。

3.2 基于中西思维共性情感计算

需要指出的是，即便中西思维存在差异，中西方的情感表达中仍存在着共性的部分，例如：中西方都将否定词用在情感词或者观点词的前面，来表达与情感词以及观点词相反的意思；在使用程度副词时，同样将副词置于形容词前，程度副词也会增加或者减少词的情感强度；当文本中出现转折词时，文本的情感倾向性则趋同于转折词后的文本的情感。本文研究中西方思维模式差异，是以中西方思维的共性为基础的。故在给出量化中西方思维差异的方法前，需要给出中西方思维的共性的量化方法。

能够满足上述共性现象的最小的级别为句子。句子可以构成段落和文章，成为段落和文章的基本组成单位，故本文以句子作为最小的分析对象进行研究。为实现中西方思维共性的计算方法，本文采用修饰词窗口策略，将用于修饰情感词的否定词、程度副词以及转折词存放到该词的修饰窗口中，当计算当前情感词对句子的情感影响时，根据情感词的基本情感值以及修饰词的作用来得到句子的情感分类以及得分。具体算法如下：

Input：Sentence(s)，s={w_1，w_2，w_3，…，w_n}

Initialize：Lexicon；notList(否定词表)；butList(转折词表)；adverbList(程度副词表)；window(修饰词窗口)；score(句子情感得分)，kind(句子情感类别)

While (i<n)

  if (lexicon. contain(w_i) == true)

    While (window is not null)

      comp(window，score，kind) //计算窗口中的词对当前句的情感作用

  else if (notList，butList，adverbList. contain(w_i) == true)

    window. put(wi)；

Output：kind，score；

上述算法给出了基于共性的情感计算的方法，是用于分析和实现中西差异算法的基础算法。下面章节中的算法是在上述算法基础上进行修改的，3.3节将给出中西思维差异所导致的情感计算方法的不同。

3.3 中西差异下的量化计算

3.3.1 图形式与直线式

在西方“直线式”思维的影响下，英文表达往往先陈述观点，而后进行论证，在这种情况下，具有情感倾向性的词汇出现在靠近句子的首部的位置上。而汉民族“图形式”的思维方式体现了东方人含蓄、委婉的思维模式，在行文时往往不采用开门见山的行文方式，而是先进行论证，在文章的末尾给出总结性结论，因此在行文时，与“直线式”相反，具有情感倾向性的词汇出现在靠近句子末段的位置上。例如：I am sorry to hear that，得知这个消息我很难过；Happy New Year，新年快乐；He was shocked by what he saw，他被他眼前的一幕震惊了。从例句中可以看出，各个例句的主要情感词：难过(sorry)，快乐(happy)，震惊(shocked)，出现在句子中的位置是不同的。如2.1节所述，目前汉民族的思维模式受到西方民族思维模式的影响，在句子级别的文本中，常采用相对直接的表达方式，但是在段落以及篇章中“图形式”的思维特点则表现得更为明显。通过对本文实验语料词语位置信息的统计，得到如下信息，在中文语料中按照词语的位置与句子长度的商值进行计算，则对句子情感倾向性有贡献的词的商值为68%，而在英文语料中这个商值为41%，这说明在中文中情感词较多出现在句子的靠近句尾的位置上，英文的情感词较多出现在句子前端。在段落和篇章中也体现了“图形式”的特点[16]。因而“图形式”和“直线式”的思维方式体现在词、句子以及段落所在的位置。所以我们采用下面的公式对中文图形式思维模式进行量化。

为实现汉民族“图形式”思维模式，本文采用式(1)作为主要的计算公式。

$$score(A)=\sum_i(1+position(a_i)/count(a|A))\times score(a_i) \tag{1}$$

为实现西方民族“直线式”思维模式，本文采用式(2)作为主要的计算公式。

$$score(A)=\sum_i(2-position(a_i)/count(a|A))\times score(a_i) \tag{2}$$

式中，A表示篇章、段落或句子，a包含于A则a表示段落、句子或词，$score(A)$表示A的情感得分，$position(a_i)$表示情感单元a_i在A中的位置，$count(a|A)$表示A中含有a的总数。从式(1)可以看出，若a_i在A中的位置越靠近末尾，则$score(a_i)$的放大倍数越大，说明a_i更能够代表A表达A的情感倾向。相反地，式(2)体现了西方民族“直线式”的思维特点，$position(a_i)$的值越大，则A的情感得分越高。

3.3.2 抽象性与具体性

在研究中发现，文本的倾向性分析中除了要对词进行分析外，还需要考虑文本的语法结构，其中包括词的词性信息。在研究中西思维的“抽象性”以及“具体性”的差异后，本文发现词性信息也有助于提高判别倾向性的效果，例如形容词在表达情感时具有其他词性的词不具备的优越性。中西民族在表达情感时，首先要考虑使用的词就是形容词，其次从思维模式的抽象性以及“具体性”的特点上来分析。中文的情感表达中，除了优先使用形容词外，更多使用的是具有动词词性的词，这说明了中文具有的动词优势。与此相反，西方民族思维的抽象性特征使得在表达情感时，除主要使用形容词外，优先考虑使用能够准确表达抽象含义的名词。

本文为了满足上述思维以及语言上的不同的特点，将情感词汇本体库中的词进行分级处理，针对不同词性的词以及句子结构采取不同的策略，式(3)给出了对中文文本进行分析时相应的处理办法。

$$score(a)=\sum_{i}^{n}W_i^{adj}+\sum_{j}^{m}W_j^{verb}+\sum_{k}^{l}W_k \tag{3}$$

针对西方民族“抽象性”思维的特点给出下面的公式进行量化：

$$score(a)=\sum_{i}^{n}W_i^{adj}+\sum_{j}^{m}W_j^{noun}+\sum_{k}^{l}W_k \tag{4}$$

式(3)中，a 代表句子，$score(a)$ 表示 a 的情感得分，W_i^{adj} 表示在句子中的第 i 个形容词，W_j^{verb} 表示在句子中的第 j 个动词，W_k 表示在句子中非形容词以及非动词的第 k 个词。式(3)不是简单的加和运算，计算过程为：首先为每个句子设置一个词语队列，用来收集句子中用于计算句子情感的词汇，当收集到形容词时，该形容词将置于词语队列中所有非形容词词语前、已收集的形容词后；当收集的词为动词时，则将该动词置于所有已收集的非形容词以及非动词词语前，其他词性的词语按照先进先出的原则进入队列。

在式(4)中，a 表示句子，$score(a)$ 表示 a 的情感得分，W_i^{adj} 表示在句子中的第 i 个形容词，W_j^{noun} 表示在句子中的第 j 个名词，W_k 表示在句子中非形容词以及非动词的第 k 个词。式(4)给出了西方民族“抽象性”量化方法，与汉民族“具体性”不同的是，英文具有名词优势，所以 W_j^{noun} 取代了式(3)的 W_j^{verb} 体现英文的名词优势。

3.3.3 散点视与聚点视

“散点视”体现在汉语表达中为使用多个表达同样含义的词汇，甚至是多个名词、形容词或数词充当谓语，而西方民族的“聚点视”体现在使用单一的词汇来表达一句话的含义。为了实现汉民族“散点视”思维特点和西方民族“聚点视”思维特点，采用视窗，即在特定的位置选择不同视窗大小来分别体现汉民族“散点视”思维特点和西方民族“聚点视”思维特点。基础算法如下：

```
Input:Sentence(s),s={w1,w2,w3,…,wn}
Initialize:viewWindow (n);
  //初始化一个大小为 n 的视窗
While (i<n)
  if (position(wi) is important)
  //如果 w1 处在一个重要的位置上
    viewWindow. put(wi);
      comp (viewWindow);
  //计算视窗中的词汇对句子的情感贡献
Output:kind,score;
```

上述算法给出了视窗的作用，即为句子的情感词挑选较重要的词进行情感计算。在算法中 w_i 所在的重要位置上，通过实验表明，在实现汉民族“散点视”的视窗的大小为 6 时，识别的效果是最好的，如果视窗过小则不能够收集到有情感贡献的词，过大则会获取到其他句子中的情感贡献词。同样地，在英文语料集上采用“聚点视”的视窗为 2 时，识别的效果较好。

综上所述，针对 3 种思维特点给出了具体的量化方法，但是在实际的语言环境中，作者在表达情感时不会考虑使用了哪种情感；同时一个句子或者篇章不仅仅体现一个思维及语言特性，往往是多个思维特点的结合，因此在对中文文本进行情感倾向性分析时，要同时考虑“图形式”、“具体性”以及“散点视”的思维特点。同样，对英文文本进行情感倾向性分析时，要考虑“直线式”、“抽象性”以及“聚点视”的思维特点。在这里我们给出了基于汉民族和西方民族思维共性的情感倾向性计算方法，同时针对两种思维特点的不同，给出了不同的处理方法。

4 实验

4.1 实验语料集

为了说明本文方法的有效性，在中英文两个语料集上进行了实验。本文采用的中文语料集为谭松波[19]评价语料集，该语料集由 3 部分组成，包括了关于酒店、电子产品以及股市的评论性文章，其中酒店评价文本 40000 篇，电子产品评价文本 1608 篇，股票评价文本 1047 篇，共 6655 篇褒贬评论性文本。语料的内容少到一句话，多到一篇文章，能够涵盖句子级、段落级以及篇章级，消除了在文章结构上的特殊性。同时语料评论主体涉及了酒店、电子产品以及股市 3 个方面的评论性文本，保证了语料的多样性。英文语料采用的是 Bo Pang[20]关于电影的评论文本，共 2000 篇篇章级文本，语料的平均长度为 778 个单词，其中褒义文本 1000 篇，贬义文本 1000 篇。

4.2 实验设计

本文设计了 6 组实验，即中文语料集上的 3 组实验以及英文语料集上的 3 组实验。在中文语料集的 3 组实验使用的对比对象为 Turney 的基于词典的方法，在 Turney 方法中依次加入汉民族“图形式”、“具体性”以及“聚点视”思维模式，并给出相应的实验结果。在英文语料集上，同样使用 Turney 方法为 Baseline 依次加入西方民族“直线式”、“抽象性”以及“聚点视”思维模式，同时给出相应的实验结果。

实验 1 设计为在 Baseline 中加入中文图形式思维特征与在 Baseline 中加入英文直线式思维的对比实验。图 1 给出了实验的结果。

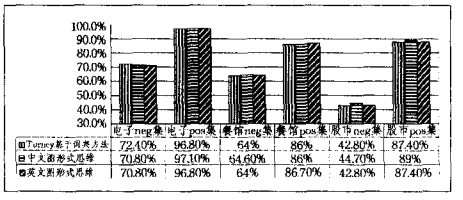

图 1 实验 1 结果

在实验 1 的基础上，实验 2 给出了分别加入中文动词词性特征与英文名词特征后的对比实验，用以体现汉民族的具体性和西方民族抽象性的思维特点，实验结果如图 2 所示。

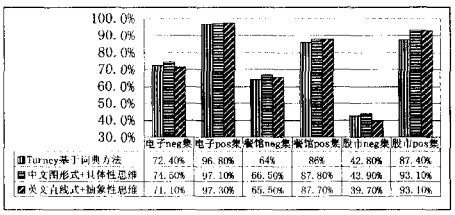

图 2　实验 2 结果

实验 3 在上述算法的基础上对中西方的思维方式分别加入汉民族散点视算法，以及西方民族聚点视算法，实验结果如图 3 所示。

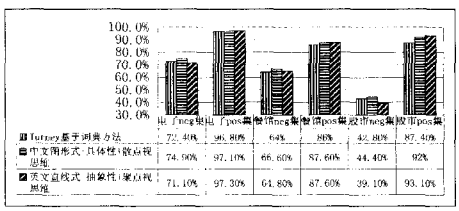

图 3　实验 3 结果

上述实验给出了总体的效果显示，下面针对不同语料的特点进行实验分析。图 4 以及图 5 为在电子产品语料中实验 1—实验 3 的实验结果。

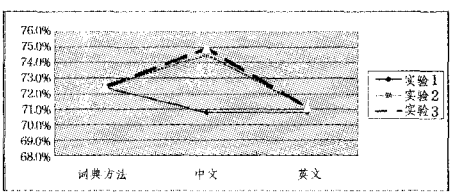

图 4　电子 neg 集实验结果

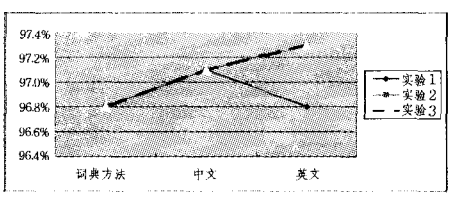

图 5　电子 pos 集实验结果

从图 4 和图 5 的曲线上可以看出，当加入新的中文的思维特征后，实验的准确率有所提高。因电子 neg 集的文本篇幅较长，且内容中多使用动词形式，故在实验中，每增加一个特征，实验效果都有所增加。从电子 pos 集中的实验可以看到，实验 2 的英文方法结果好于中文的方法，原因在于在文本中多使用名词，影响了结果效果。实验 2 较实验 1 效果有明显的提高，但是加入了新的特征后效果并没有提高，原因是语料中使用的情感词不是很丰富，更多的文本采用单一的情感词来表述自己的情感。

图 6 和图 7 为在餐馆评价集合上的实验结果。从该语料的两个语料集上的结果上看，当在语料中加入“具体性”因素后，实验效果有明显的提高。说明了“具体性”在中文中表现十分明显，但是视窗大小要根据具体的语料来设定，以达到最好的实验效果。

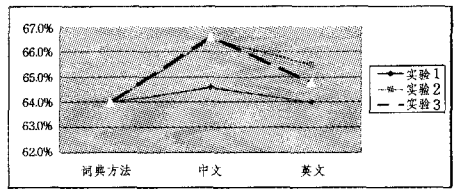

图 6　餐馆 neg 集实验结果

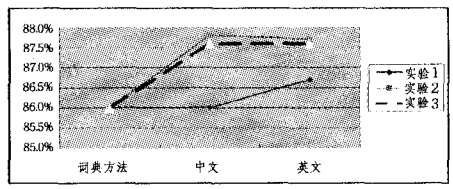

图 7　餐馆 pos 集实验结果

图 8 和图 9 为在股市相关评价中的实验结果。

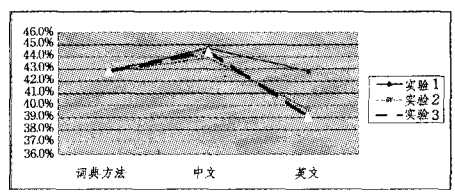

图 8　股市 neg 集实验结果

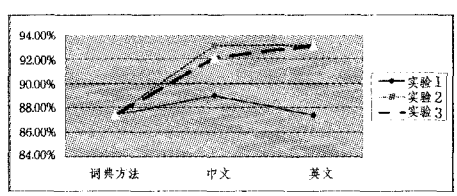

图 9　股市 pos 集实验结果

在图 8 和图 9 中，由于股票的专业词汇过多，部分词汇不在情感词汇本体库中出现，例如在股市 neg 集中使用“把指数拉了下来”、“预期从紧”等词汇，导致了识别的效果不理想，这正是基于词典的方法局限性所在。虽然整体的准确率不高，但是从图 8 和图 9 可以看出，按照汉民族思维模式分析后的文本的情感倾向性分析的效果要高于使用西方民族思维模式的效果，原因在于股市的语料集中，大部分的文本属于篇章级别的文本。正是因为篇章级别的文本篇幅较长，所以在组织语言和构思过程中其更能体现汉民族思维的特点；正是因为文本的思维特点表现明显，所以提出的方法就更能体现出优势。

为了同中文语料上的实验进行对比，本文同样在英文语料上进行了实验，同样将 Turney 的基于词典的方法作为 Baseline。实验 4 设计为在 Baseline 中加入中文图形式思维特征与 Baseline 中加入英文直线式思维特征的对比实验。实验 5 设计为在实验 4 的基础上分别加入“具体性”以及“抽象性”思维特点后的对比实验。实验 6 设计为在实验 5 的基础上分别加入“散点视”以及“聚点视”思维特点后的对比实验，表 1 给出了实验的结果。

从图 9 的结果上看，英文文本应用西方民族直线式思维特征后，倾向性识别的准确率比基于词典的方法以及应用汉民族思维特征后的效果要高。该语料集篇幅较长，再加之西方文明受其他文明的影响不是很深刻，所以其保持了本民族原有的特点，符合西方民族思维特点，故实验效果有所提高。下面针对两个不同的语料集进行分析。

表 1　英文语料集上的实验结果

	neg 集	pos 集
Turney	64.6%	73.6%
中文图形式	65.0%	72.0%
英文直线式	70.3%	83.6%
中文图形式＋动词优势	62.3%	78.7%
英文直线式＋名词优势	78.3%	88.3%
中文图形式＋动词优势＋散点视	58.0%	75.8%
英文直线式＋名词优势＋聚点视	79.0%	89.1%

从图 10 和图 11 中可以看到，同样在英文语料集上的实验中考虑英文的“抽象性”中的名词优势时，实验效果有明显的提升。中文由于不符合西方民族的思维方式，因此效果较差。实验结果说明了本文方法的有效性。

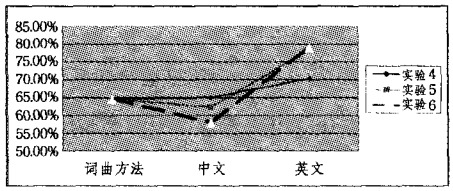

图 10　英文 neg 集实验结果

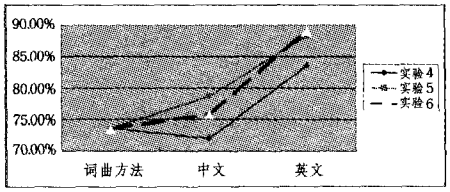

图 11　英文 pos 集实验结果

结束语　本文从思维决定语言的角度出发，通过对汉民族以及西方民族的思维特点的研究，总结出汉民族“图形式”“具体性”、“散点视”的思维特点以及西方民族“直线式”、“抽象性”、“聚点视”的思维特点，根据各自思维特点的表现，本文给出了相应量化思维特点的方法。通过对关于思维特点理论的分析以及实验结果的验证，得出以下结论：目前汉民族的思维方式受到西方民族思维方式的影响，在使用篇幅较短的文本或语言进行表达时，常体现出西方民族思维方式的特点，但是在篇幅较长的文本中，汉民族思维的特点则表现得很明显；而英美文化并没有受到其他文化的影响，所以按照西方民族的思维特点分析英文文本时，情感倾向性分析的效果有所提高。

在研究中发现“图形式”、“直线式”、“散点视”、“聚点视”思维特点由于涉及到篇章布局相关思维，因此主要在篇章级别的文本里体现；而“具体性”和“抽象性”的思维特点在不同篇幅的文本中均有体现，因此在进行情感分析时，根据文字下表现不同的思维特点，应用相应的量化方法，将会提高情感倾向性分析的效果。从实验结果上看，本文的方法还存在不足之处，例如“图形式”、“直线式”、“散点视”、“聚点视”相关的实验方法对情感倾向性分析计算的效果提升并不明显，有时因为语料的特殊性，会使得分析效果下降。故采用这些思维特点的相关应用时，需要对语料有较多的了解，这些思维特点的量化方法有待提高，根据思维特点给出更优的量化方法为今后的工作重点。

参 考 文 献

[1] Riloff E, Patwardhan S, Wiebe J. Feature subsumption for opinion analysis[C]// Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP). Sydney, Australia, 2006: 1035-1045

[2] Jakob, Niklas, Gurevych I. Extracting Opinion Targets in a Single- and Cross-Domain Setting with Conditional Random Fields [C]// Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP). Massachusetts, USA, 2010: 1035-1045

[3] Qiu Guang, Liu Bing, Bu Jia-jun, et al. Opinion Word Expansion and Target Extraction through Double Propagation [J]. Computational Linguistics, 2011, 37(1): 9-27

[4] Zhang Lei, Liu Bing. Extracting Resource Terms for Sentiment Analysis[C]// Proceedings of the 5th International Joint Conference on Natural Language Processing (ICNLP). Chiang Mai, Thailand, 2011: 1171-1179

[5] Morinaga S, Yamanishi K, Tateishi K, et al. Mining product reputations on the Web[C]// Proceedings of the ACM SIGKDD Conference on Knowledge Discovery and Data Mining (SIGKDD). Edmonton, Alberta, Canada, 2002: 341-349

[6] Tong R M. An operational system for detecting and tracking opinions in on-line discussion[C]// Proceedings of the Workshop on Operational Text Classification. New Orleans, Louisiana, USA, 2001

[7] Yi J, Nasukawa T, Bunescu R, et al. Sentiment analyzer: Extracting sentiments about a given topic using natural language processing techniques[C]// Proceedings of the IEEE International Conference on Data Mining (ICDM). Melbourne, Florida, USA, 2003: 427-434

[8] Kim S-M, Hovy E. Determining the sentiment of opinions[C]// Proceedings of the International Conference on Computational Linguistics (COLING). Geneva, Switzerland, 2004

[9] 赵军，许洪波，黄萱菁. 中文倾向性分析评测技术报告[R]. 第一届中文倾向性分析评测会议，2008

[10] 许洪波，姚天昉，黄萱菁. 第二届中文倾向性分析评测(COAE2009)[C]// 第五届全国信息检索学术会议第二届中文倾向性分析评测会议. 上海，中国，2009

[11] Du Wei-fu, Tan Song-bo. Building domain-oriented sentiment lexicon by improved information bottleneck[C]// Proceedings of the 18th ACM Conference on Information and Knowledge Management (CIKM). HongKong, China, 2009: 1749-1752

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

[13] Ding Xiao-wen, Liu Bing, Yu P S. A Holistic Lexicon-Based Approach to Opinion Mining[C]// Proceedings of First ACM International Conference on Web Search and Data Mining (WSDM). California, USA, 2008: 231-240

[14] Du Wei-fu, Tan Song-bo. Building domain-oriented sentiment lexicon by improved information bottleneck[C]// Proceedings of the 18th ACM Conference on Information and Knowledge Management (CIKM). HongKong, China, 2009: 1749-1752

[15] 徐琳宏，林鸿飞. 认知视角下的文本情感计算[J]. 计算机科学，2010, 37(12): 182-185

[16] 连淑能. 论中西思维方式[J]. 外语与外语教学，2002, 155(2): 40-48

[17] 王福祥，吴汉樱. 文化与语言[M]. 北京：外语教学与研究出版社，1994

[18] 徐琳宏，林鸿飞，潘宇，等. 情感词汇本体的构造[J]. 情报学报，2008, 27(2): 180-185

[19] Wu Qiong, Tan Song-bo, et al. SentiRank: Cross-Domain Graph Ranking for Sentiment Classification[C]// IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology. Milano, Italy, 2009: 309-314

[20] Pang Bo, Lee L. A Sentimental Education: Sentiment Analysis Using Subjectivity Summarization Based on Minimum Cuts[C]// Proceedings of the Association for Computational Linguistics (ACL). Barcelona, Spain, 2004