

## 基于三维坐标的消费情绪本体库建立及应用

邱云飞<sup>1</sup>, 林明明<sup>1\*</sup>, 邵良杉<sup>2</sup>

(1. 辽宁工程技术大学 软件学院, 辽宁 葫芦岛 125100; 2. 辽宁工程技术大学 系统工程研究所, 辽宁 葫芦岛 125100)

(\* 通信作者电子邮箱 gleylin@163.com)

**摘要:**针对商家好评中存在非真正满意的评价问题,构建一种能够真正反映消费者情绪状态的方法,以减少好评率中非真正满意的评价。针对消费情绪进行了研究,首先从评价中提取出消费情绪词汇,根据消费情绪的特征,将消费情绪划分为7大类,25小类,建立了三维坐标模型;其次,用Protégé来构建消费情绪本体库,根据三维坐标词汇分类算法对消费情绪词汇进行自动划分;然后,根据构建的本体库,用消费情绪判断算法来自动判断消费者的评价。最后,与淘宝的好评率进行比较, $F$ 值达到了95%以上,减少了好评中非真正满意的评价,体现了消费者的真实情绪。

**关键词:**消费情绪; 词汇分类; 三维坐标; 本体库; 评价

**中图分类号:** TP391.1 **文献标志码:** A

### Establishment and application of consumption sentiment ontology library based on three-dimensional coordinate

QIU Yunfei<sup>1</sup>, LIN Mingming<sup>1\*</sup>, SHAO Liangshan<sup>2</sup>

(1. School of Software, Liaoning Technical University, Huludao Liaoning 125100, China;

2. System Engineering Institute, Liaoning Technical University, Huludao Liaoning 125100, China)

**Abstract:** Since the positive comments may have the non-truly satisfied comments, a method which can truly reflect the sentiment of the consumers was constructed in order to decrease the non-truly satisfied comments from the positive comments. The research oriented to the consumption sentiment shows that the consumption sentiment vocabulary should be extracted at first. According to the consumption sentimental features, consumption sentiment got classified into seven classes and twenty-five subclasses, and the three-dimensional coordinate model was established. Afterwards, Protégé was used to build a consumption sentiment ontology library so that the consumption sentiment can be automatically classified by the three-dimensional coordinate vocabulary classification algorithm. Moreover, the consumption sentiment judging algorithm was applied to automatically judge consumer comments based on the completed library. Finally, compared with the positive comment ratio of Taobao, the  $F$ -measure can reach more than 95%. It can eliminate the non-truly satisfied comments from positive comments and reflect the consumer's real emotion.

**Key words:** consumption sentiment; vocabulary classification; three-dimensional coordinate; ontology library; comment

## 0 引言

现在越来越多的人选择在网上购物,然而网购最重要的就是商家的好评率,但是卖家可能会逼迫消费者给予好评,致使商家的好评率并不真实,但是在评价内容中会体现消费者不满意的地方,消费评价就会体现消费者的情绪,所以消费情绪在消费过程中是很重要的。本文就是针对消费情绪来反映消费者的真实态度,从好评率中减少非真正满意的评价,反映了消费者的真实情绪。

目前,国内外对情绪分类的研究还是比较少的,大多研究的是情感分类,国外许多学者研究了情感分类方法,文献[1–2]都是用分类器进行情感倾向分析。文献[3]在分类方法方面进行了探索,提出了一种组合的方法,即通过将多种分类器进行混合叠加来提高情感分类的性能。但是这些文献只是将情感分为积极、消极和中性,并没有细化分类。国内对情感分类的研究刚刚起步,并且应用在消费领域中更少,文献

[4–6]只是对微博情感进行分类,也是分为了积极、消极和中性,并没有细化。

情感分析也用到语义识别的知识,文献[7–9]对其进行了深入的研究。目前国内外主要研究的是情感方面,本文侧重于情绪方面,基于消费评价找出消费者的情绪,在积极和消极的基础上,将其情绪类别细化,更好地反映消费者的情绪。本文要将消费情绪和计算机联系起来,要用计算机的语言来定义消费情绪,用计算机来处理消费情绪,对其进行分类,来判断消费者对商品的满意程度,减少好评率中非真正满意的评价,反映消费者的真实情绪。

## 1 消费情绪词汇

### 1.1 情绪的相关概念

情绪是指人对外界刺激所产生的心理或生理反应。消费情绪是消费者对商品及服务绩效所产生的心理反应和情感回应,消费评价就是这种情绪的体现。根据情绪发生的不同特

**收稿日期:** 2013-03-18; **修回日期:** 2013-05-07。 **基金项目:** 国家自然科学基金资助项目(70971059); 辽宁省高等学校创新团队支持计划项目(2009TD45); 辽宁省高等学校杰出青年学者成长计划项目(LJQ2012027)。

**作者简介:** 邱云飞(1976–),男(蒙古族),辽宁阜新人,教授,博士,CCF会员,主要研究方向:数据挖掘、情感分析; 林明明(1989–),女,辽宁大连人,硕士研究生,主要研究方向:数据挖掘、情感分析; 邵良杉(1961–),男,辽宁凌源人,教授,博士生导师,主要研究方向:数据挖掘、情感分析。

征,将情绪划分为三种表现状态,即:激情、心境和热情<sup>[10]</sup>。

情绪是一种即时的感觉,更能体现消费者的心理状态。情绪并不等同于情感,其主要是指感情过程具有情境性、激动性和暂时性的特点,而且情绪可以包含动词,反映一种表现状态,而情感是指具有稳定性、深刻性和持久性的感情,带有一定的社会因素和外在此因素。基于本文总结出的差别,所以本文研究情绪而不是情感。

## 1.2 消费情绪分类

到目前为止,心理学家和语言学家对情绪和情感做了不同的划分<sup>[11-14]</sup>,因为人类的情感和情绪比较复杂,不容易界定,所以并没有一个统一的标准对其划分。

根据众多学者对情绪、情感的分类和消费的特点,以及根据后文对情绪强度的计算和三维坐标模型建立的需求,本文将消费情绪分为7大类:乐、好、忠、怒、哀、惧、惊。将这7大类归为3种态度。有些词虽然属于同一大类,但是却从不同方向表现这一大类,所以再将这7大类细化为25个子类,即:喜爱、安心、愤怒、憎恶、惊奇等。分类结果如表1所示。

表1 消费情绪分类

最终态度	情绪大类	细分小类
满意	乐	喜爱,安心,快乐,激动
	好	享受,决定,希望,思念,感激,赞扬
	忠	忠诚,相信,崇拜
不满意	怒	愤怒,憎恶,贬责
	哀	后悔,失望,忧愁,怀疑,无奈,焦虑
	惧	恐惧,犹豫
中性	惊	惊奇

## 2 三维坐标模型划分消费情绪

### 2.1 情绪强度定义

本文将消费情绪词汇分为7大类:H(乐)、G(好)、F(忠)、A(怒)、S(哀)、D(惧)、I(惊)。这7大类也就是基本类,其他的词就用这7个基本类组合而成,每个基本类赋上不同程度的值,就能组合成不同的词。

$$\text{Intensity} = \{H, G, F, A, S, D, I\}$$

如果是褒义词,就用乐、好、忠、惊来表示;如果是贬义词,就用怒、哀、惧、惊来表示。在这里要强调指出的是,“惊”是一个特殊的中性类,可以表述积极的,也可以表示消极的,并不是所有的词都需要含有“惊”的意思,当这个词含有“惊”的意思,那么就需要将“惊”这一类赋上相应的值。

本文首先采用人工标注词汇强度,两个实验员同时对词汇进行强度标注,设强度的值域为0~9,取0~9中的奇数,因此强度的值域为 $I = \{0, 1, 3, 5, 7, 9\}$ ,0就是不包含该类的词意,9就是包含最强的该类的词意,选择这样的梯度和范围,可以减少误差,得到准确的结果,用强度向量 $q = (q_x, q_y, q_z)$ 和惊的强度值 $J$ 来表示每个词的强度值。

### 2.2 三维坐标模型构建

根据以上分析,将强度计算用一个三维坐标模型来表示,乐、好、忠分别为X、Y、Z轴的正半轴,怒、哀、惧为X、Y、Z轴的负半轴,正半轴和负半轴类并不需要是一对反义类,因为一个词如果是褒义词,它就不可能含有贬义词的含义,相反一个词如果是贬义词,就不会含有褒义词的含义,所以只需要将正半轴定义为褒义类,负半轴定义为贬义类即可,而“惊”大类不参与计算,只需从“惊”大类的值中,就可以看出含有惊的成分是多少。设褒义词的定义域为积极域 $POS = (0^\circ, 90^\circ)$ ,

贬义词的定义域为消极域 $NEG = (90^\circ, 180^\circ)$ ,中性词的定义域为中性域 $ZH = \{0, 1, 3, 5, 7, 9\}$ 。消费情绪词汇强度的三维坐标模型如图1所示。

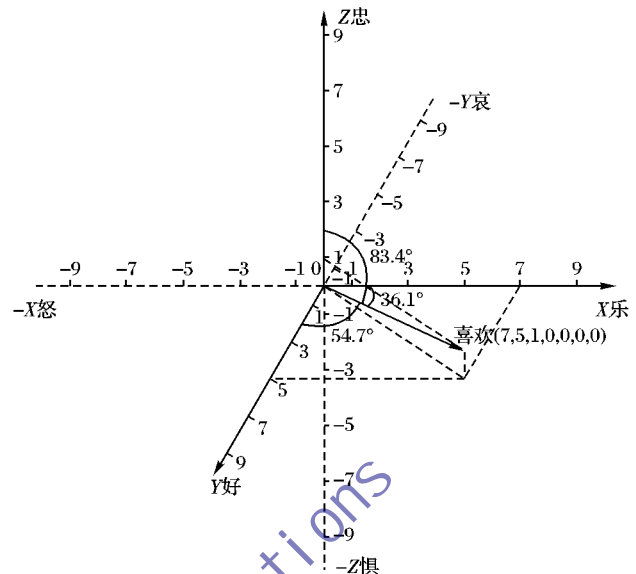


图1 消费情绪词汇强度的三维坐标模型

### 2.3 三维坐标词汇分类算法

设 $W = \{w_1, w_2, \dots, w_n\}$ 为待分类词的集合, $w_i$ 为每一个待分类的词。 $\text{Intensity}(w_i) = (i_1, i_2, \dots, i_7)$ 为每一个词的强度。 $F = (f_1, f_2, \dots, f_n)$ 为分类完的词的集合, $f_i$ 为每一个分类完成的词。设 $P = \{\text{乐, 好, 忠}\}$ 为积极情绪类的集合, $N = \{\text{怒, 哀, 惧}\}$ 为消极情绪类的集合, $Z = \{\text{惊}\}$ 为中性类的集合。

首先将 $w_i$ 添加到本体库中的“词汇”类中,然后进行情绪强度赋值,再从本体库中读取出词和其强度值 $\text{Intensity}(w_i) = (i_1, i_2, \dots, i_7)$ ,将强度值处理为强度向量 $q = (q_x, q_y, q_z)$ ,以及将 $i_7$ 赋值给惊的强度值 $J$ 。设X、Y、Z轴的单位向量分别为 $x = (1, 0, 0)$ , $y = (0, 1, 0)$ , $z = (0, 0, 1)$ 利用式(1)计算向量 $q$ 与X、Y、Z轴的余弦值。

$$\begin{cases} W_{\cos x}(q) = \cos(\theta_x) = \frac{q \cdot x}{|q| |x|} \\ W_{\cos y}(q) = \cos(\theta_y) = \frac{q \cdot y}{|q| |y|} \\ W_{\cos z}(q) = \cos(\theta_z) = \frac{q \cdot z}{|q| |z|} \end{cases} \quad (1)$$

再将得到的 $W_{\cos x}(q)$ 、 $W_{\cos y}(q)$ 、 $W_{\cos z}(q)$ 用式(2)取反余弦值,得到向量 $q$ 与X、Y、Z轴的夹角 $\theta_x$ 、 $\theta_y$ 、 $\theta_z$ ,根据该词的 $\theta_x$ 、 $\theta_y$ 、 $\theta_z$ 将其归类。

$$\begin{cases} \text{Angle}_x(W_{\cos x}(q)) = \theta_x = \arccos(W_{\cos x}(q)) \\ \text{Angle}_y(W_{\cos y}(q)) = \theta_y = \arccos(W_{\cos y}(q)) \\ \text{Angle}_z(W_{\cos z}(q)) = \theta_z = \arccos(W_{\cos z}(q)) \end{cases} \quad (2)$$

根据计算之后,比较与每个轴的夹角来判断属于哪一类。如果是褒义词,那么所计算出的 $\theta_x$ 、 $\theta_y$ 、 $\theta_z$ 都是小于 $90^\circ$ ,所以只要比较出与哪个轴的夹角最小,该词就属于哪个类,也就是越接近该轴的正半轴;如果是贬义词,那么计算出的 $\theta_x$ 、 $\theta_y$ 、 $\theta_z$ 都是大于 $90^\circ$ ,所以只要比较出与哪个轴的夹角最大,该词就属于哪个类,也就是越接近该轴的负半轴;如果是中性词,不为该词进行前6大类的强度赋值,只为该词赋值“惊”的强度值。由于判断后得到的只是7大类的结果,每个大类还有许多小类,因为它们可能强度相同,但是方向不同,所以还要细

分,之所以没有构建多维坐标,直接判断为25小类,是为了降低算法的复杂度,而且多维坐标的正交性差。

#### 算法1 三维坐标词汇分类算法。

输入  $W$ 。

输出  $F$ 。

```

1) Mark the intensity of word //人工标记情绪词的强度值
2) Iterator  $c_i$  in CLASS =  $\{c_1, c_2, \dots, c_n\}$  //循环本体库中所有类
3) If  $c_i$  = “词汇” //找到词汇类
4) Class. getLocalName(); property. getLocalName() //得到要分类的词汇和强度
5) End iterator
6) If  $c_i \neq \text{null}$  //如果词汇类中不为空
7)  $W_{\cos X}(q); W_{\cos Y}(q); W_{\cos Z}(q)$  //计算词汇的余弦值
8)  $\theta_x = \text{Angle}_x(W_{\cos X}(q));$ 
 $\theta_y = \text{Angle}_y(W_{\cos Y}(q));$ 
 $\theta_z = \text{Angle}_z(W_{\cos Z}(q))$  //计算词汇与各个坐标轴的夹角
9) End if
10) If  $\theta_x < 90 \wedge \theta_y < 90 \wedge \theta_z < 90$  //如果是积极词汇
11)  $\theta_{\text{pos}} = \min\{\theta_x, \theta_y, \theta_z\}$ ; Add  $w_i$  in  $c_{\text{pos}}$  //计算出最小角度的轴,将该词添加到对应的类中
12) Else if  $\theta_x > 90 \wedge \theta_y > 90 \wedge \theta_z > 90$  //如果是消极词汇
13)  $\theta_{\text{neg}} = \max\{\theta_x, \theta_y, \theta_z\}$ ; Add  $w_i$  in  $c_{\text{neg}}$  //计算出最大角度的轴,将该词添加到对应的类中
14) Else //如果是中性类
15) Add  $w_i$  in  $c_{\text{zh}}$  //将该词添加到“惊”大类中
16) End if //结束

```

#### 2.4 具体实例

如:褒义词“喜欢”,它的强度值是:  $\text{Intensity}(\text{喜欢}) = (7, 5, 1, 0, 0, 0, 0)$ 。“喜欢”含有很高的“乐”的含义,含有一部分“好”的含义,同时喜欢某样东西也略带有“忠”的含义,而贬义词类怒、哀、惧肯定不含有,也不含有“惊”的意思,将该强度值处理成强度向量  $q = (7, 5, 1)$ ,  $J=0$ ,根据式(1)~(2)计算得:该词与  $X$  轴夹角  $\theta_x = 36.1^\circ$ ,与  $Y$  轴夹角  $\theta_y = 54.7^\circ$ ,与  $Z$  轴夹角  $\theta_z = 83.4^\circ$ ,经过比较,“喜欢”一词与  $X$  轴的夹角  $\theta_x$  最小,不含有“惊”的意思,所以该词属于“乐”大类。

### 3 消费情绪词汇本体库的构建

#### 3.1 本体的概念

本体(Ontology)的概念源于哲学。这里将本体<sup>[15-16]</sup>进行定义:一个本体是一种对共享概念化的形式化的明确规范<sup>[17]</sup>。本文建立的是消费情绪词汇本体库,属于领域本体<sup>[18]</sup>,根据消费情绪词汇的特征,将其明确化、规范化、形式化。人类对消费情绪词汇的定义,计算机并不能理解,要用计算机所能理解的语言对消费情绪词汇进行定义,通过本体的一系列元素对消费情绪词汇进行一个标准的定义,这样可以让计算机来识别和处理消费情绪词汇。

#### 3.2 构建消费情绪本体的基本元素

本文用斯坦福大学开发的软件 Protégé 来构建本体。Protégé 开放源码,用户可以通过点击相应的项目来增加或编辑类、属性、实例等,它已成为国内外众多本体研究机构的首选工具<sup>[19]</sup>。

1)类。类精确地定义了所要描述的知识资源,将其规范化。类有父类和子类,OWL 中所有的类都有一个根类:owl:Thing,用户自定义的类都隐含地继承了这个根类。本文首先在根类 owl:Thing 下建立两个类:消费情绪和词汇。然后依据

第2章对消费情绪的分类,来确定在本体库中类与类之间的关系,首先在消费情绪类下建立三个子类:满意、不满意、中性。再根据表1对情绪的分类,将7个大类归为这三个类中,如:满意的子类有:乐、好、忠。再将这7大类细化成不同的子类,如:乐的子类有:喜爱、安心等;怒的子类有:愤怒、憎恶等。本文还建立了一个词汇类,该类用于新增词汇,存入新增词汇信息,最后再根据推理判断,归入上述建立的消费情绪类中。

2)属性。定义类之后,就要根据情绪词汇的特征定义属性,本文在 Protégé 中定义了数据类型属性。根据上文分析的,描述消费情绪特征的属性有:POS、Meaning、Emotional\_State、Sentence、Polarity、Intensity、X、Y、Z。本体库的属性如表2所示。

表2 本体库属性表

属性名称	comment(注释)	Domain(定义域)	Range(值域)
POS	词性	消费情绪	String
Meaning	词义	消费情绪	String
Emotional_State	情绪状态	消费情绪	String
Sentence	例句	消费情绪	String
Polarity	极性	消费情绪	String
Intensity	强度	消费情绪	String
X	与X轴夹角	消费情绪	Float
Y	与Y轴夹角	消费情绪	Float
Z	与Z轴夹角	消费情绪	Float

3)实例。根据上述建立好的类、属性以及之间的关系,消费情绪本体库的框架已经建立好,接下来需要向建立好的框架中添加内容,也就是需要添加实例。首先向“词汇”类中添加实例,根据推理判断后将实例归入各个情绪类中。

4)规则。本文根据三维坐标模型,用语义网规则语言(Semantic Web Rule Language, SWRL)进行编写规则,然后转换为 Jess 进行推理判断,将消费情绪词汇自动分类。如:属于“乐”类的规则如下:

词汇( $?x$ )  $\wedge$  X( $?x, ?a$ )  $\wedge$  Y( $?x, ?b$ )  $\wedge$  Z( $?x, ?c$ )  $\wedge$  swrlb:lessThan( $?a, 90$ )  $\wedge$  swrlb:lessThan( $?b, 90$ )  $\wedge$  swrlb:lessThan( $?c, 90$ )  $\wedge$  swrlb:lessThan( $?a, ?b$ )  $\wedge$  swrlb:lessThan( $?a, ?c$ )  $\rightarrow$  乐( $?x$ )

#### 3.3 构建本体库的步骤

本文的知识来源于《同义词词林》<sup>[20]</sup>和淘宝中的评价。首先在 owl:Thing 下建立了两个类:“消费情绪”和“词汇”。在“消费情绪”类下建立了三个最终的态度类,再依次建立子类。建立词汇类是为了新增新词,当要纳入新词的时候首先将词汇的基本信息填入,如:词性、极性、词义等。接着从本体库中读出纳入的新词汇,然后进行人工标注强度,经过计算与各个坐标轴的夹角,将夹角自动录入新词汇的属性中,再进行推理,将新加入的词汇重新归类,得到最终的情绪类。

### 4 消费情绪判断

淘宝中的评价有许多买家是习惯性好评,但是在评价中他们会写到不满意的地方,这些评价实质是不满意的评价,然而在淘宝中会显示为好评,这样就会使淘宝中的好评率并不是真实的。本文基于三维坐标模型建立本体库,将情绪词汇进行划分,利用本体库判断评价中买家是否真正满意,提高好评率的真实度。

#### 4.1 评价内容预处理

本文借助的是中国科学院 ICTCLAS50 的分词系统,从句



子集  $L = \{l_1, l_2, \dots, l_n\}$  ( $n$  为要判断的句子数) 中, 依次取出每个句子, 对句子进行分词。有些词前面有否定意义的词修饰, 本文处理的方法就是导入用户词典, 然后进行判断, 这样提高了判断的准确率。

#### 4.2 情绪词的隶属度函数

对于句子  $l_i$ , 与本体库匹配成功的关键词可能有很多, 而且这些词的情绪倾向也可能不同, 即使同一情绪倾向, 也很难严格将句子归类, 所以本文采用隶属度函数。而在一个句子中程度副词也起着很重要的决定, 如: “喜欢” “很喜欢” “有点喜欢”, 这三个词的喜欢程度是不同的, 本文根据韩容洙<sup>[21]</sup> 给出的程度副词等级和其他学者的总结<sup>[22]</sup> 以及评价的实际情况, 给出程度副词及其权重系数如表 3 所示。

表 3 程度副词及其权重系数

程度副词	权重
绝、绝顶、极其、极为、极端、极度、极大、及	9
异常、万分、特别、特	7
格外、分外、十分、太	5
多么、多、何等、何其、非常、好、好不、甚为、甚、很	3
蛮、颇、颇为、较、较为、比较、有些	2
稍微、稍稍、微微、多少、有点儿	1

对于一个句子来说, 当前面表达很多观点时, 可能在最后会有一个总结性的分句来陈述总体的情绪倾向, 这个总结性的分句在整个句子中会起着决定性的作用, 因此也要找到这样的句子, 找总结性的句子关键是要找总结词汇, 找到总结词汇, 将总结性句子中的关键情绪词的角度乘以权重系数, 这样就会增加这个词的情感倾向比重。根据语言学以及淘宝评价, 本文给出总结性的词汇如表 4 所示。

表 4 总结性词汇

总结性词汇	权重
总体来说、总的来说、总而言之、总之、综上所述、综上所述	1.5

设  $w_j$  为句子  $l_i$  中的情绪词汇, 其中  $1 \leq j \leq n$ ,  $n$  为  $l_i$  中情绪词的数量。对于词  $w_j$ , 与各个轴的夹角不同, 因此所占的积极域的比率也不同, 对于句子  $l_i$  中的关键词  $W = \{w_1, w_2, \dots, w_n\}$ , 根据式(1) ~ (2) 计算每个关键词  $w_j$  与  $X$ 、 $Y$ 、 $Z$  轴的夹角为  $\theta_{xj}$ 、 $\theta_{yj}$ 、 $\theta_{zj}$ , 以及所占“惊”大类的含量  $J_j$ , 每一个关键词可以表示为  $w_j = (\theta_{xj}, \theta_{yj}, \theta_{zj}, J_j)$ 。对于词  $w_j$ , 不能说其完全属于哪一大类, 也不能说其完全不属于哪一大类, 所以用式(3) ~ (5) 计算每个关键词在各个类中的比例, 用隶属度函数  $A_{POS}(\theta)$ 、 $A_{NEG}(\theta)$  和  $A_J(J)$  表示。 $A_{POS}(\theta)$  表示积极词汇所占积极域的比率,  $A_{NEG}(\theta)$  表示消极词汇所占消极域的比率,  $A_J(J)$  表示中性词汇所占中性域的比率。

$$A_{POS}(\theta) = \frac{90 - \theta \times (1/\sigma) \times (1/\delta)}{90} \quad (3)$$

$$A_{NEG}(\theta) = \frac{180 - (180 - \theta) \times (1/\sigma) \times (1/\delta)}{180} \quad (4)$$

$$A_J(J) = \frac{9 - (9 - J) \times (1/\sigma) \times (1/\delta)}{9} \quad (5)$$

其中:  $\sigma \in B = \{b \mid b \text{ 为程度副词所占的权重}\}$ ,  $\delta \in C = \{c \mid c \text{ 为总结性词汇所占的权重}\}$ 。

#### 4.3 消费情绪判断算法

计算每个词汇的隶属度函数后, 就用每个词的隶属度的值来计算  $l_i$  的积极和消极的总比率,  $LP_{POSX}$  表示  $l_i$  中所有属于“乐”大类的词的总比率, 同理,  $LP_{POSY}$ 、 $LP_{POSZ}$ 、 $LP_{NEGX}$ 、 $LP_{NEGY}$ 、

$LP_{NEGZ}$  表示各个类的总比率, 用式(6) 和式(7) 分别计算积极和消极总比率, 各个类的总比例值可能大于 1, 这并不影响计算的结果。

$$\begin{cases} LP_{POSX} = \sum_{j=1}^n A_{POS}(\theta_{xj}) \\ LP_{POSY} = \sum_{j=1}^n A_{POS}(\theta_{yj}) \\ LP_{POSZ} = \sum_{j=1}^n A_{POS}(\theta_{zj}) \end{cases} \quad (6)$$

$$\begin{cases} LP_{NEGX} = \sum_{j=1}^n A_{NEG}(\theta_{xj}) \\ LP_{NEGY} = \sum_{j=1}^n A_{NEG}(\theta_{yj}) \\ LP_{NEGZ} = \sum_{j=1}^n A_{NEG}(\theta_{zj}) \end{cases} \quad (7)$$

最后再根据式(8) 计算句子的总倾向性  $LP$ :

$$LP = LP_{POSX} + LP_{POSY} + LP_{POSZ} - LP_{NEGX} - LP_{NEGY} - LP_{NEGZ} \quad (8)$$

如果  $LP > 0$ , 那么该句子为满意的评价; 如果  $LP < 0$ , 那么该句子为不满意的评价。如果句子没有褒义词和贬义词, 只有“惊”大类的词, 那么该句子只能判断出属于“惊”大类。当判断出为最终的情绪倾向后, 再判断出最终属于哪一大类。如果是积极类, 用式(9) 来找出所占积极类比重最大的类即为该句子所属类别; 如果是消极类, 用式(10) 来找出所占消极类比重最大的类即为该句子所属类别。这样可以判断消费者的最终情绪。

$$LP_{POS} = \max\{LP_{POSX}, LP_{POSY}, LP_{POSZ}\} \quad (9)$$

$$LP_{NEG} = \max\{LP_{NEGX}, LP_{NEGY}, LP_{NEGZ}\} \quad (10)$$

算法 2 消费情绪判断算法。

输入  $L$ 。

输出  $H = \{h_1, h_2, \dots, h_n\}$ 。

```

1) For each  $i$  in  $L = \{l_1, l_2, \dots, l_n\}$  //对要判断的句子进行循环处理
2) Divide  $l_i$  into  $W = \{w_1, w_2, \dots, w_n\}$  //对每个句子进行分词
3) For each  $j$  in  $W = \{w_1, w_2, \dots, w_n\}$  //对每个句子中的词进行循环
4) Get  $w_j = (\theta_{xj}, \theta_{yj}, \theta_{zj}, J_j)$  //计算每个词与坐标轴的夹角以及“惊”的程度
5)  $A_{POS}(\theta_j)$ ;  $A_{NEG}(\theta_j)$ ;  $A_J(J_j)$  //计算每个词的隶属度函数
6) End for; //一句话中关键词计算结束
7)  $LP_{POSX}$ ;  $LP_{POSY}$ ;  $LP_{POSZ}$ ;  $LP_{NEGX}$ ;  $LP_{NEGY}$ ;  $LP_{NEGZ}$ 
   //分别计算每一大类所占比率
8) Get  $LP$  //得到句子总的情绪倾向性
9) If  $LP > 0$  //如果句子的总倾向性为积极性
10) Get  $LP_{POS}$ 
   //找出积极性中所占比率最大的一类即为该句子最终所属类
11) Else if  $LP < 0$  //如果句子的总倾向性为消极性
12) Get  $LP_{NEG}$ 
   //找出消极性中所占比率最大的一类即为该句子最终所属类
13) Else //句子的总倾向性为“惊”大类
14) End if;
15) End for; //所有句子循环结束

```

## 5 实验评价

本文的数据来自淘宝评价, 本文从淘宝的每个类中各获

取了500条评价,共7500条评价,经过两名工作人员的人工标注,其中选取7126条作为实验数据。首先对7126条评价使用中国科学院分词系统ICTCLAS5.0对句子进行分词,找出每个句子的关键词,然后根据消费情绪判断算法对句子进行判断分类。

为了验证本文方法的正确性、准确性和实用性进行了实验,将得到的数据与淘宝自动的好评率作比较。本文方法对消费者的评价做了正确的判断,有些消费者习惯性好评,在淘宝评价系统中就会认为是好评,这样淘宝的好评率其实并不是完全真实的,本文采用消费情绪判断算法可以把评价正确分类,得到一个相对真实的好评率。实验数据如表5所示。

表5 各个商品类的好评数量比较

名称	评价总数	真正好评数	淘宝显示的好评数	本文方法测出的好评数	淘宝中真正的好评数	本文方法测出真正的好评数
虚拟	481	298	330	315	290	294
服装类	463	272	316	298	265	267
鞋包配饰	472	267	298	279	260	262
运动户外	481	306	339	316	300	301
珠宝手表	482	304	334	320	298	302
数码	479	305	341	318	299	300
家电	482	298	328	309	290	295
美容护发	470	274	307	296	268	269
母婴用品	475	284	315	299	279	281
家居建材	476	280	314	298	272	275
美食特产	478	274	304	291	269	271
日用百货	471	267	288	272	258	260
汽车用品	479	282	307	293	276	278
文化玩乐	468	287	304	291	279	281
本地生活	469	270	309	285	259	262
总数	7126	4268	4734	4480	4162	4198

根据数据得到折线图,如图2所示。从图2中可以看出淘宝好评数量要高于真实好评数量,这其中有許多非真实的好评,这样就给消费者造成误导,本文降低了好评数量中非真实的好评数量,从评价本身入手,判断出消费者真实消费情绪,这样使好评率更真实,才能为消费者提供一个正确的引导。为了验证本文方法的全面性和准确性,评价标准采用召回率 $R$ 、准确率 $P$ 和 $F$ 值,用来评价整体性能。根据算法表示特点,定义召回率 $R$ 、准确率 $P$ 和 $F$ 值的计算公式如下:

$$R = \frac{\text{系统判断出的正确好评数量}}{\text{真正的好评数量}}$$

$$P = \frac{\text{系统判断出的正确好评数量}}{\text{系统判断的好评数量}}$$

$$F = \frac{2 \times R \times P}{R + P}$$

实验数据的 $F$ 值比较如图3所示。

对淘宝产品不同分类的准确率和召回率以及 $F$ 值进行计算,得到结果如表6。从表6中可以看出本文方法的准确率要高于淘宝准确率,而召回率并没有明显提高,因为淘宝自身的召回率就已经很高了,这是因为淘宝中的好评率基本上全部包含了消费者真实的好评,而很少有消费者给的是差评,而评价内容是好评,所以召回率就会很高。最后从图3可以明显看出淘宝和本文方法的 $F$ 值的差异,淘宝评价的 $F$ 值为92%,而本文方法的 $F$ 值为96%,降低了好评中的非真实好评数量。

表6 淘宝与本文方法的准确率、召回率和 $F$ 值比较 %

名称	淘宝			本文方法		
	准确率	召回率	$F$ 值	准确率	召回率	$F$ 值
虚拟	87.88	97.32	92.36	93.33	98.66	95.92
服装类	83.86	97.43	90.14	89.60	98.16	93.68
鞋包配饰	87.25	97.38	92.04	93.91	98.13	95.97
运动户外	88.50	98.04	93.02	95.25	98.37	96.78
珠宝手表	89.22	98.03	93.42	94.38	99.34	96.79
数码	87.68	98.03	92.57	94.34	98.36	96.31
家电	88.41	97.32	92.65	95.47	98.99	97.20
美容护发	87.30	97.81	92.25	90.88	98.18	94.39
母婴用品	88.57	98.24	93.16	93.98	98.94	96.40
家居建材	86.62	97.14	91.58	92.28	98.21	95.16
美食特产	88.49	98.18	93.08	93.13	98.91	95.93
日用百货	89.58	96.63	92.97	95.59	97.38	96.47
汽车用品	89.90	97.87	93.72	94.88	98.58	96.70
文化玩乐	91.78	97.21	94.42	96.56	97.91	97.23
本地生活	83.82	95.93	89.46	91.93	97.04	94.41
总数	87.92	97.52	92.47	93.71	98.36	95.98

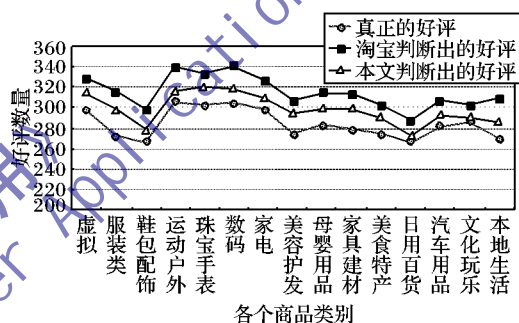


图2 好评数量比较统计

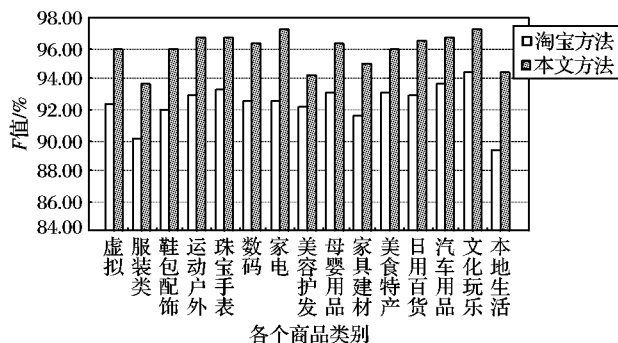


图3 淘宝与本文方法的 $F$ 值比较

## 6 结语

随着网络购物的兴起,人们越来越多在网上购物,购物后的评价是很重要的,体现着消费者的消费情绪,卖家的逼迫或习惯性好评使好评率并不真实,但是在评价内容本身会体现出消费者的不满意。本文基于三维坐标模型,用三维坐标词汇分类算法将消费情绪词汇分类,构建本体库,用计算机的语言来定义消费情绪词汇,再用消费情绪判断算法得出消费者的真实情绪,减少了好评中的非真实好评数,使评价内容更真实,体现了消费者的真实情绪。下一步研究的关键就是,根据评价的不同方向,来进一步对评价进行分析,比如:许多消费者评价快递的好坏,这对于商品本身并没有作用,所以要针对不同方向进行分析研究,提高准确度。

### 参考文献:

- [1] PAK A, PAROUBEK P. Twitter as a corpus for sentiment analysis

- and opinion mining [C]// Proceedings of the International Conference on Language Resources and Evaluation Conference. Malta: LREC, 2010: 1320–1326.
- [2] MELVILLE P, GRYC W, LAWRENCE R D. Sentiment analysis of blogs by combining lexical knowledge with text classification [C]// Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. New York: ACM Press, 2009: 1275–1284.
- [3] GO A, BHAYANI R, HUANG L. Twitter sentiment classification using distant supervision [EB/OL]. [2012-10-10]. <http://cs.stanford.edu/people/alecmgo/papers/TwitterDistantSupervision09.pdf>.
- [4] 庞磊, 李寿山, 周国栋. 基于情绪知识的中文微博情感分类方法 [J]. 计算机工程, 2012, 38(13): 156–158.
- [5] 刘鲁, 刘志明. 基于机器学习的中文微博情感分类实证研究 [J]. 计算机工程与应用, 2012, 48(1): 1–4.
- [6] 张珊, 于留宝, 胡长军. 基于表情图片与情感词的中文微博情感分析 [J]. 计算机科学, 2012, 39(23): 146–148.
- [7] 徐琳宏, 林鸿飞, 赵晶. 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): 96–101.
- [8] 徐琳宏, 林鸿飞, 潘宇. 情感词汇本体的构造 [J]. 情报学报, 2008, 27(2): 180–185.
- [9] 崔大志, 孙立伟. 在线评论情感词汇模糊本体库构建 [J]. 辽宁工程技术大学学报, 2010, 12(4): 395–398.
- [10] 邹忠民. 作家的情绪心理 [J]. 昭通师范高等专科学校学报, 1992, 14(2): 64–71.
- [11] 林传鼎. 社会主义心理学中的情绪问题——在中国社会心理学研究会成立大会上的报告 (摘要) [J]. 社会心理科学, 2006, 21(1): 37–37.
- [12] 许小颖, 陶建华. 汉语情感系统中情感划分的研究 [C]// 第一届中国情感计算及智能交互学术会议论文集. 北京: 中国中文信息学会, 2003: 199–205.
- [13] EKMAN P. Facial expression and emotion [J]. American Psychologist, 1993, 48(4): 384–392.
- [14] ZHANG Y, LI Z M, REN F J. *et al.* Semi-automatic emotion recognition from textual input based on the constructed emotion thesaurus [C]// Proceedings of 2005 IEEE International Conference on Natural Language Processing and Knowledge Engineering. Piscataway, NJ: IEEE Press, 2005: 571–576.
- [15] NECHES R, FIKES R E, GRUBER T R, *et al.* Enabling technology for knowledge sharing [J]. AI Magazine, 1991, 12(3): 36–56.
- [16] 宋炜, 张铭. 语义网简明教程 [M]. 北京: 高等教育出版社, 2004.
- [17] GRUBER T R. Toward principles for the design of ontologies used for knowledge sharing [J]. International Journal of Human Computer Studies, 1995, 43(5): 907–928.
- [18] 杨卓. 本体理论研究初探 [J]. 中小企业管理与科技, 2011(21): 142–143.
- [19] 滕悦明. 基于本体的远程教学辅助系统的设计与实现 [D]. 北京: 北京邮电大学, 2007.
- [20] 梅家驹, 竺一鸣, 高蕴琦, 等. 同义词词林 [M]. 上海: 上海辞书出版社, 1996.
- [21] 韩容珠. 现代汉语的程度副词 [J]. 汉语学习, 2000(2): 12–15.
- [22] 王海, 冯向前, 钱钢. 网页在线评论情感倾向的直觉模糊分类 [J]. 计算机工程与应用, 2013, 49(1): 148–152.

(上接第 2539 页)

限  $t$ , 来实现对系统参数的控制和对成员的进一步约束, 二者可相互配合和相互约束, 在安全性和效率两方面达到最佳平衡点, 但该方案也有系统参数较多、成员数量大时开销会增加的不足, 这是以后需完善的方向。

#### 参考文献:

- [1] FATEMI M, EGHLIDOS T, AREF M. An efficient multistage secret sharing scheme using linear one-way functions and bilinear maps [EB/OL]. [2012-03-02]. <http://eprint.iacr.org/2012/121>.
- [2] CARLES R, LEONOR Y, YANG J. Finding lower bounds on the complexity of secret sharing schemes by linear programming [EB/OL]. [2012-03-02]. <http://eprint.iacr.org/2012/464>.
- [3] TANG C, GAO S, ZHANG C. The optimal linear secret sharing scheme for any given access structure [EB/OL]. [2012-03-02]. <http://eprint.iacr.org/2011/147>.
- [4] CRAMER R, DAMGARD I, MAURER U. General secure multi-party computation from any linear secret-sharing scheme [C]// EUROCRYPT 2000: Proceedings of the 19th International Conference on Theory and Application of Cryptographic Techniques. New York: ACM Press, 2000: 316–334.
- [5] NIKOY V, NIKOVA S, PRENEEL B. Multi-party computation from any linear secret sharing scheme secure against adaptive adversary: the zero-error case [EB/OL]. [2012-03-02]. <http://eprint.iacr.org/2003/006>.
- [6] YUEN K, CHEONG S W. A secret sharing scheme of prime numbers based on hardness of factorization [EB/OL]. [2012-03-02]. <http://eprint.iacr.org/2012/222>.
- [7] KAYA K, SELCUK A. Secret sharing extensions based on the chinese reminder theorem [EB/OL]. [2012-03-02]. <http://eprint.iacr.org/2010/096>.
- [8] FATEMI M, EGHLIDOS T, AREF M. A multi-stage secret sharing scheme using all-or-nothing transform approach [C]// Proceedings of ICICS 2009. New York: ACM Press, 2009: 449–458.
- [9] WANG S J, TSAI Y R, SHEN J. Dynamic threshold multi-secret sharing scheme using elliptic curve and bilinear maps [C]// FGNCN 2008: Proceedings of the Second International Conference on Future Generation Communication and Networking. Piscataway, NJ: IEEE Press, 2008, 2: 405–410.
- [10] WONG T M, WANG C X, WING J M. Verifiable secret redistribution for threshold sharing schemes [EB/OL]. [2012-03-02]. <http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.160.4811>.
- [11] PAUL M B, ANYI V. Algorithmic chaos [EB/OL]. [2010-07-05]. <http://arxiv.org/abs/nlin/0303016v1>.
- [12] LANGLOIS D, VERNIZZI F. Nonlinear perturbations of cosmological scalar fields [EB/OL]. [2010-07-05]. <http://arxiv.org/pdf/astro-ph/0610064.pdf>.
- [13] BEIMEL A, ISHAI Y. On the power of nonlinear secret-sharing [EB/OL]. [2012-01-12]. <http://eprint.iacr.org/2001/030>.
- [14] 张利远, 张恩. 基于中国剩余定理的可验证理性秘密共享方案 [J]. 计算机应用, 2012, 32(11): 3143–3146.
- [15] 罗黎霞, 张峻. 基于双线性映射的动态门限签名方案 [J]. 计算机应用, 2010, 30(3): 677–679.