(19)中华人民共和国国家知识产权局

(12)发明专利申请

(10)申请公布号 CN 106202200 A
(43)申请公布日 2016.12.07

(21)申请号 201610485392.4

(22)申请日 2016.06.28

(71)申请人 昆明理工大学
地址 650093 云南省昆明市五华区学府路253号

(72)发明人 邵玉斌 王丽霞 刘彩 王晨歌 杜庆治

(51)Int.Cl.
G06F 17/30(2006.01)

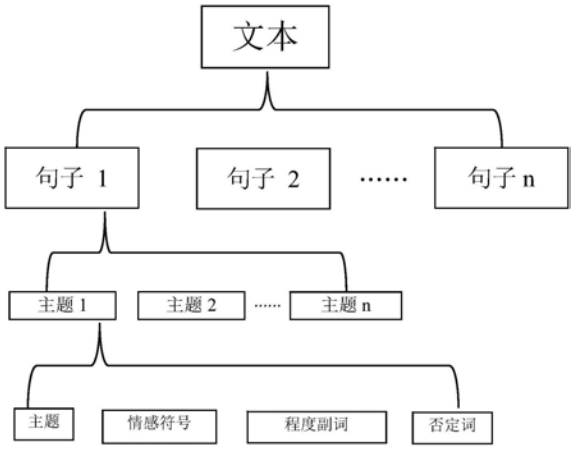
权利要求书2页 说明书5页 附图1页

(54)发明名称

一种基于固定主题的文本情感倾向性分类方法

(57)摘要

本发明公开了一种基于固定主题的文本情感倾向性分类的方法，属于文本情感倾向性分类领域。首先找出句子的主题，根据主题在本句的位置分成两步分别计算此主题前后的情感倾向，最终再计算出此主题的情感倾向。利用特征情感符号和通用情感词典找出句子中的情感符号；在主题词与情感符号之间找否定词和程度副词并计算其对此情感符号的影响；在情感符号之间找连接关系并计算本主题的情感倾向。本发明能帮助用户得到其他用户对某一产品、服务、事件或人物重要属性的倾向程度，并细分出相关用户对此产品、事件或人物各特征方面的情感倾向。



CN 106202200 A

1.一种基于固定主题的文本情感倾向性分类方法,其特征在于,包括以下步骤:

(1)划分中文文本中句子的组成成分

将句子划分为四种成分,包括主题T、情感符号S、修饰情感符号的程度副词W、修饰情感符号的否定词P;

(2)建立特征属性等式,找出中文文本中所有主题T;

(3)找出每个主题T常用的特征情感符号和通用情感词典,按照积极和消极为情感符号赋值D_S;

(4)按句子顺序,找出中文文本中第一个主题T_1,在主题T_1后面从特征情感符号和通用情感词典找出第一个情感符号S_1,其情感倾向值为步骤(3)中对应的情感符号赋值D_S,没有情感符号则情感倾向值为0并继续查找下一主题;

(5)在第一个主题T_1与第一个情感符号S_1之间找出所有的否定词P和程度副词W,并记录其位置P_{ID}和W_{ID},计算第一个情感符号S_1的情感倾向值S_{T11};

①否定词P的个数为偶数时:

$$S_{T11}=\begin{cases} D_S & (D_W=0) \\ D_S \cdot D_W & (D_W \neq 0) \end{cases}$$

②否定词P的个数为奇数时:

$$S_{T11}=\begin{cases} -S & (D_W=0) \\ \dfrac{W_{ID}-P_{ID}}{|W_{ID}-P_{ID}|} \cdot D_S \cdot D_W & (D_W \neq 0) \end{cases}$$

其中D_S为情感符号的赋值,D_w表示程度副词的赋值,W_{ID}为程度副词的位置,P_{ID}为离S最近的否定词的位置;

(6)在第一个情感符号S_1后面继续查找第二个情感符号S_2,按照步骤(4)和(5)计算第二个情感符号S_2的情感倾向值S_{T12},并按照两者之间的连接关系计算第一个主题T_1后面的情感倾向值,若第一个情感符号S_1后面没有其他情感符号则查找下一主题;

①并列关系:第一个主题T_1后面的情感倾向值为第一个情感符号S_1和第二个情感符号S_2的情感倾向值之和;

②转折关系:第一个主题T_1后面的情感倾向值为第二个情感符号S_2的情感倾向值;

(7)继续顺序查找句子中其他情感符号直至句末或下一个主题词,并按照步骤(6)计算出第一个主题T_1后总的情感倾向值S_{T1A};

(8)查找第一个主题T_1前面的情感符号,并按步骤(4)-(7)计算第一个主题T_1前面的情感倾向值S_{T1B};

(9)计算第一个主题T_1的情感倾向值$S_{T1}=S_{T1A}+S_{T1B}$;

(10)依次查找其他主题并计算情感倾向值,每一句的情感倾向通过本句中所有主题的情感倾向值之和判断。

2.根据权利要求1所述的基于固定主题的文本情感倾向性分类方法,其特征在于:所述步骤(2)中的主题T包括关键词、特征属性和细分特征属性,

$t_1=t_2=t_3=...=t_i=...=t_n$,$t_i$表示与主题T等同的关键词,$i\in[1,n]$;

特征属性等式如下:

$t_i = A_1 + A_2 + A_3 + \ldots + A_p + \ldots + A_m$

其中A_p表示与关键词t_i相关的特征属性，$p \in [1,m]$；

$A_p = a_{p1} + a_{p2} + a_{p3} + \ldots + a_{pq} + \ldots + a_{pk}$

a_{pq}表示特征属性A_p的细分特征属性，$q \in [1,k]$。

3.根据权利要求1所述的基于固定主题的文本情感倾向性分类方法，其特征在于：步骤(3)中针对每个主题T对知网的情感词典、台湾大学的情感词典、大连理工大学的情感词典进行对比找出情感倾向有差别的情感符号，同时按照词频统计方法找出主题T常用的情感符号，将两者的结果叠加作为主题T的特征情感符号，且对所有特征情感符号进行积极和消极的倾向划分；对所有积极的情感符号赋值D_s为1，所有消极的情感符号赋值D_s为-1。

4.根据权利要求1所述的基于固定主题的文本情感倾向性分类方法，其特征在于：所述步骤(5)中将程度副词W的程度倾向按照稍、很、极其进行分类

并赋程度值D_w：稍的程度值为1，很的程度值为2，极其的程度值为3。

一种基于固定主题的文本情感倾向性分类方法

技术领域

[0001] 本发明涉及一种基于固定主题的文本情感倾向性分类方法，属于文本情感倾向性分类领域。

背景技术

[0002] 在网络信息爆炸的时代，如何得到大众对某一事件、产品的观点或看法，即如何从这些评论信息中找出有用的参考数据，是近十几年来国内外相关研究者的重要内容。

[0003] 目前针对情感倾向性分类主要采用的是基于情感词典和基于大规模语料库的机器学习，而不管是基于词典或是机器学习其关键在于情感词典的质量。利用一个情感词典对不同的主题进行分类，情感词典的质量必然达不到专业的要求并且会大大降低情感分类的速度。由于评论信息对象属性特征的多样性，一个主题中往往两个评论方向都存在，即不同的属性拥有不同的情感倾向。这就需要对主题中存在的属性进行逐一分析，使得对此主题的分类更加详细可靠。

发明内容

[0004] 本发明的目的在于提出一种针对某一固定主题的情感倾向性分类方法，使针对主题的情感分类更加详细可靠，具体包括以下过程：

[0005] 首先将句子划分为四种成分，包括主题T、情感符号S、修饰情感符号的程度副词W、修饰情感符号的否定词P；

[0006] 为主题T建立等式，主题T包括关键词、特征属性和细分特征属性，与主题相关的所有关键词t_i可以互相表示：$t_1=t_2=t_3=...=t_i=...=t_n$，$i\in[1,n]$；

[0007] 为关键词t_i建立特征属性等式：

[0008] $t_i=A_1+A_2+A_3+...+A_p+...+A_m$，$p\in[1,m]$；

[0009] 为A_p建立一个细分特征属性等式：$A_p=a_{p1}+a_{p2}+a_{p3}+...+a_{pq}+...+a_{pk}$

[0010] a_{pq}表示特征属性A_p的细分特征属性，$q\in[1,k]$。

[0011] 然后为每一个主题T找出其常用特征情感符号：通过对知网的情感词典、台湾大学的情感词典、大连理工大学的情感词典进行对比找出情感倾向有差别的情感符号，同时按照词频统计方法找出主题T中常用的情感符号，将两者结果叠加作为主题T的特征情感符号，且对所有特征情感符号进行积极和消极的倾向划分。将得到特征情感符号之后的情感词典合并得到通用情感词典。最后对所有积极的情感符号赋值D_s为1，所有消极的情感符号赋值D_s为-1。

[0012] 再次考虑到不同的程度副词对情感符号的影响不同，需要对程度副词W赋值处理，具体操作为：将程度副词W的程度倾向按照“稍、很、极其”进行分类并赋程度值D_w：“稍”的程度值为1，“很”的程度值为2，“极其”的程度值为3。

[0013] 按照下面步骤计算并得到句子的情感倾向：

[0014] (1)按句子顺序，找出中文文本中第一个主题T_1，在主题T_1后面从特征情感符号和

通用情感词典找出第一个情感符号S_1,其情感倾向值为对应的情感符号赋值D_s,没有情感符号则情感倾向值为0并继续查找下一主题;

[0015] (2)在第一个主题T_1与第一个情感符号S_1之间找出所有的否定词P和程度副词W,并记录其位置P_{ID}和W_{ID},计算第一个情感符号S_1的情感倾向值S_{T11};

[0016] ①否定词P的个数为偶数时:

[0017] $$S_{T11}=\begin{cases} D_S & (D_W=0) \\ D_S \cdot D_W & (D_W \neq 0) \end{cases}$$

[0018] ②否定词P的个数为奇数时:

[0019] $$S_{T11}=\begin{cases} -S & (D_W=0) \\ \dfrac{W_{ID}-P_{ID}}{|W_{ID}-P_{ID}|} \cdot D_S \cdot D_W & (D_W \neq 0) \end{cases}$$

[0020] 其中D_S为情感符号的赋值,D_w表示程度副词的赋值,W_{ID}为程度副词的位置,P_{ID}为离S最近的否定词的位置;

[0021] (3)在第一个情感符号S_1后面继续查找第二个情感符号S_2,按照步骤(4)和(5)计算第二个情感符号S_2的情感倾向值S_{T12},并按照两者之间的连接关系计算第一个主题T_1后面的情感倾向值,若第一个情感符号S_1后面没有其他情感符号则查找下一主题;

[0022] ①并列关系:第一个主题T_1后面的情感倾向值为第一个情感符号S_1和第二个情感符号S_2的情感倾向值之和;

[0023] ②转折关系:第一个主题T_1后面的情感倾向值为第二个情感符号S_2的情感倾向值;

[0024] (4)继续顺序查找句子中其他情感符号直至句末或下一个主题词,并按照上述步骤计算出第一个主题T_1后总的情感倾向值S_{T1A};

[0025] (5)查找第一个主题T_1前面的情感符号,并按步骤(4)-(7)计算第一个主题T_1前面的情感倾向值S_{T1B};

[0026] (6)计算第一个主题T_1的情感倾向值$S_{T1}=S_{T1A}+S_{T1B}$;

[0027] (7)依次查找其他主题并计算情感倾向值,每一句的情感倾向通过本句中所有主题的情感倾向值之和判断。

[0028] 本发明的有益效果:与现有情感分类的技术相比,本发明是在确定研究的主题之后再对此主题进行情感分类之前的分析。分析之后得到的情感大词典包括特征情感符号和通用情感词典,整个大词典质量更加可靠,最终的情感分类效率更高,且本发明针对主题的多个属性进行了单独分析,使得分类结果更加详细可靠。

[0029] 本发明能帮助用户得到其他用户对某一产品、服务、事件或人物重要属性的倾向程度,并细分出相关用户对此产品、事件或人物各特征方面的情感倾向。

附图说明

[0030] 图1是文本结构图;

[0031] 图2是句子中主题的情感倾向分析流程图。

具体实施方案

[0032] 为了更加清楚、方便地描述本发明，下面结合附图及具体实施例对本发明进一步说明。

[0033] 以一则评论华为荣耀7的短文为例：

[0034] 华为荣耀7是国产手机中的战斗机。祝愿华为品牌举国产手机大旗，做大、做强民族品牌。荣耀7一到，拆开包装一看，还真是惊艳，并且还真不是一般地惊喜，系统流畅，电池容量大，想不到还带有指纹锁。

[0035] 分析以上文本，文本中包含下面内容：

[0036] 句子1：“华为荣耀7是国产手机中的战斗机。”

[0037] 句子2：“祝愿华为品牌举国产手机大旗，做大、做强民族品牌。”

[0038] 句子3：“荣耀7一到，拆开包装一看，还真是惊艳，并且还真不是一般地惊喜，系统流畅，电池容量大，想不到还带有指纹锁。”

[0039] 首先确定关键词：手机＝华为荣耀7＝荣耀7＝华为荣耀7手机 （式1）

[0040] 手机＝运行+屏幕+摄像头+通话+连网+电池+外观+价格+附赠品 （式2）

[0041] 运行＝内存+CPU+系统 （式3）

[0042] 屏幕＝尺寸+分辨率 （式4）

[0043] 通过对知网的情感词典、台湾大学的情感词典、大连理工大学的情感词典进行对比找出情感倾向有差别的情感符号和按照词频统计方法找出各主题的特殊情感符号，根据式2得到特征属性表1：

[0044] 表1特征属性表------华为荣耀7

[0045]

主题名	特征积极情感符号	特征消极情感符号
手机	性价比高、战斗机、神机、功能强大、质量好、好品牌、有蓝牙功能、有热点	太重、太沉、垃圾货、杂牌、无蓝牙功能、无热点
运行	运行流畅、系统流畅	机子卡
屏幕	大	小
摄像头	像素高、图片清楚	像素低、图片模糊
通话情况	说话清楚	杂音
连网情况	信号强	信号弱
电池	容量大、待机时间长、通话时间长、充电快	容量小、待机时间短、通话时间短、充电慢
外观	手感好、颜色好看、做工精、材质好	手感不好、颜色不好看、做工粗糙
价格	实惠、便宜	不便宜
附赠品	送耳机、送手机壳、有赠品	不送耳机、不送手机壳、无附赠品

[0046] 按照式3、式4，分别得细分特征属性表2、表3：

[0047] 表2细分特征属性表------运行

[0048]

主题名	特征积极情感符号	特征消极情感符号

内存	内存大	内存小
CPU	主频高、快	慢
系统	流畅、快	反应慢、不好

[0049] 表3细分特征属性表------屏幕

[0050]

主题名	特征积极情感符号	特征消极情感符号
尺寸	屏大	屏小
分辨率	高	低

[0051] 可按照所需的实际情况来分类主题的属性和特征情感符号的极性，如用户对属性中的附赠品不需要关注，可以不将此属性归入表中；如需要更全的手机属性可以按照上式继续添加。

[0052] 第一句中的第一个主题T_1＝华为荣耀7，在主题T_1后面从特征情感符号和通用情感词典找出第一个情感符号S_1＝战斗机，其情感符号赋值为1，则其情感倾向值为1。

[0053] T_1与S_1之间无程度副词则D_{W1}＝0，无否定词，则主题T_1的情感倾向值S_{T1}＝1。

[0054] 主题T_1所在句子中没有其他情感符号，则第一句的情感倾向值为1，表明此句情感倾向为积极。

[0055] 第二句没有包含任何主题，所以不相关，在此不做分析。

[0056] 第三句中的主题T_2＝荣耀7，T_3＝系统，T_4＝电池。

[0057] 主题T_2后的第一个情感符号S_2＝惊艳，其情感符号赋值D_{S2}为1，在主题T_2与情感符号S_2之间找程度副词W＝真是，其中D_{W2}＝2，无否定词则情感符号S_2的情感倾向值$S_{T21}=D_{S2}\cdot D_{W2}=1\cdot 2=2$。

[0058] 在情感符号S_2位置之后找到情感符号S_3＝惊喜，D_{S3}＝1，情感符号S_2与情感符号S_3之间修饰情感符号S_3的程度副词W＝一般，D_{W3}＝1，否定词P＝不是，否定词P个数为1，W_{ID}＝310，P_{ID}＝309，位置标号第一位为句子序号，后两位为词在本句中的序号，对句子进行分词处理，其中对分词之后的词汇进行标注，序号为00、01、02、03....，则情感符号S_3的情感倾向值

[0059] $$S_{T22}=\frac{W_{ID}-P_{ID}}{|W_{ID}-P_{ID}|}\cdot D_{S3}\cdot D_{W3}$$

[0060] 情感符号S_3与情感符号S_2为并列关系，且主题T_2前后没有其他情感符号，则主题T_2的情感倾向值$S_{T2}=S_{T21}+S_{T22}=2+1=3$，表明主题$T_2$的情感倾向是积极的。

[0061] 主题T_3后的第一个情感符号S_4＝流畅，D_{S4}＝1，S_4前无程度副词和否定词，则其情感倾向值$S_{T31}=1$，主题T_3前面也没有其他情感符号，因此主题T_3的情感倾向值$S_{T3}=S_{T31}=1$，表明主题T_3的情感倾向为积极。

[0062] 主题T_4后的第一个情感符号S_5＝容量大，D_{S5}＝1，S_5前无程度副词和否定词，则其情感倾向值$S_{T41}=1$，主题T_4前面也没有其他情感符号，因此主题T_4的情感倾向值$S_{T4}=S_{T41}=1$，表明主题T_4的情感倾向为积极。

[0063] 上面结合附图对本发明的具体实施例作了详细说明，但是本发明并不限于上述实施例，在本领域普通技术人员所具备的知识范围内，还可以在不脱离本发明宗旨的前提下

作出各种变化。

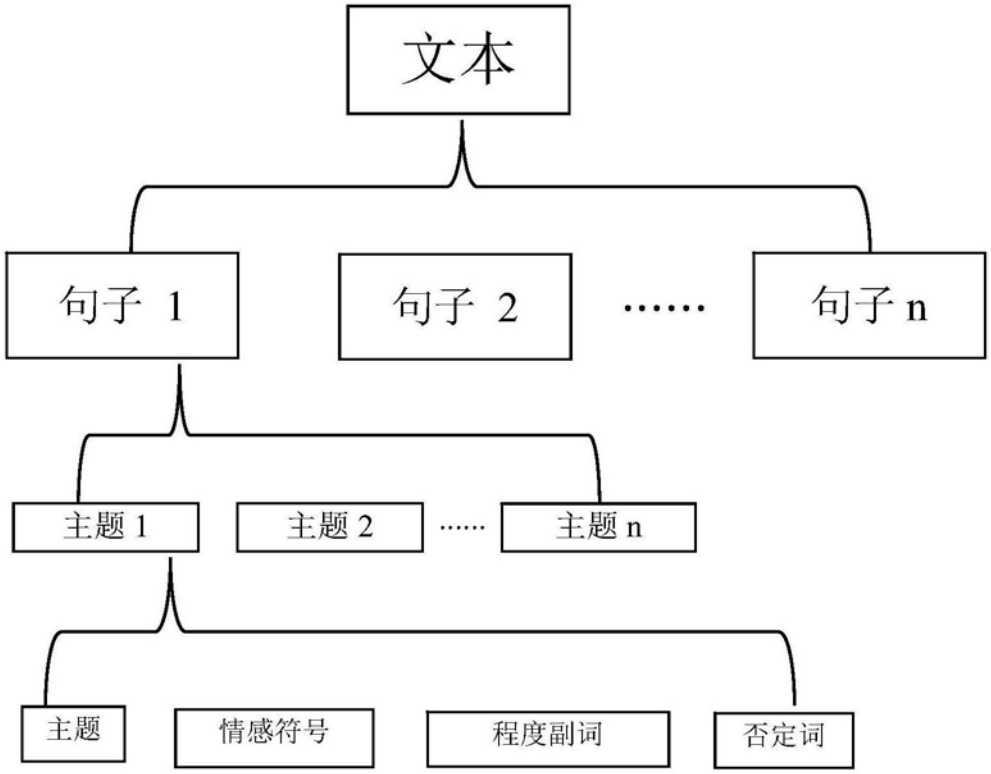

文本
句子 1
句子 2
……
句子 n
主题 1
主题 2
……
主题 n
主题
情感符号
程度副词
否定词

图1

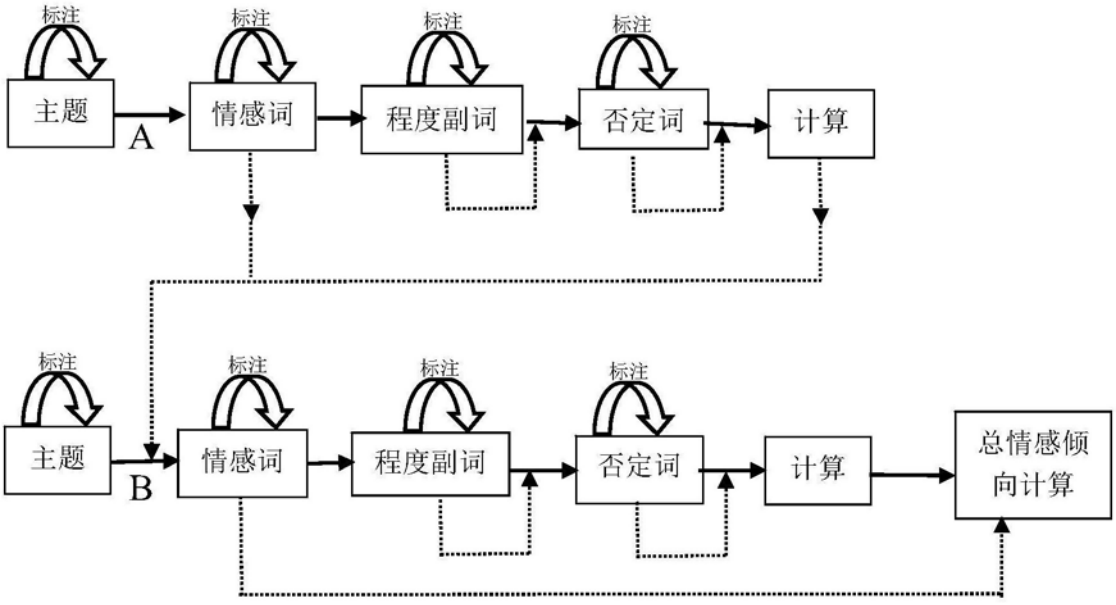

标注
主题
A
情感词
程度副词
否定词
计算
主题
B
总情感倾向计算

图2