

(19)中华人民共和国国家知识产权局



(12)发明专利

(10)授权公告号 CN 103514279 B

(45)授权公告日 2016.10.05

(21)申请号 201310445953.4

(56)对比文件

(22)申请日 2013.09.26

CN 100593783 C, 2010.03.10,

(65)同一申请的已公布的文献号

US 6092035 A, 2000.07.18,

申请公布号 CN 103514279 A

杨超等.“基于情感词典扩展技术的网络舆情倾向性分析”.《小型微型计算机系统》.2010,第31卷(第4期),第691-694页.

(43)申请公布日 2014.01.15

杨频、李涛、赵奎.“一种网络舆情的定量分析方法”.《计算机应用研究》.2009,第26卷(第3期),第1066页至第1078页.

(73)专利权人 苏州大学

审查员 张莹

地址 215123 江苏省苏州市工业园区仁爱路199号

(72)发明人 李寿山 朱珠 周国栋

(74)专利代理机构 北京集佳知识产权代理有限公司 11227

代理人 常亮

(51)Int.Cl.

G06F 17/30(2006.01)

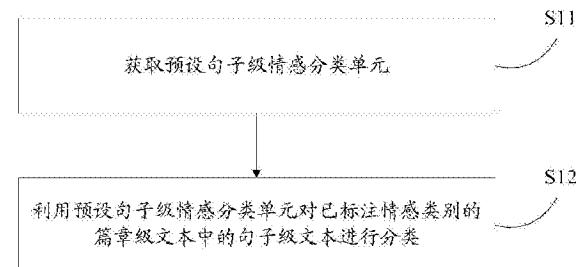
权利要求书5页 说明书17页 附图7页

(54)发明名称

一种句子级情感分类方法及装置

(57)摘要

本申请提供了一种句子级情感分类方法，包括：获取预设句子级情感分类单元；利用预设句子级情感分类单元对已标注情感类别的篇章级文本中的句子级文本进行分类；其中，预设句子级情感分类器的获取过程，包括：分别获取标记为正、负和客观的篇章级文本；对拆分篇章级文本获取到的句子级文本按照正、负和客观类型进行分类标记，得到对应的正、负和客观训练样本；利用正、负和客观训练样本对最大熵分类器进行训练，得到预设句子级情感分类单元。因此，本申请相比于采用人工标注的方式对句子级文本进行分类，提高了工作效率，且由于不需要再使用人工来标注，因此不需要支付人工费，降低了成本。



1. 一种句子级情感分类方法,其特征在于,包括:

获取预设句子级情感分类单元;

利用所述预设句子级情感分类单元对已标注情感类别的篇章级文本中的句子级文本进行分类;

其中,所述预设句子级情感分类单元的获取过程,包括:

分别获取标记为正、负和客观的篇章级文本;

确定所述标记为正的篇章级文本、所述标记为负的篇章级文本、所述标记为正的篇章级文本中的句子文本和所述标记为负的篇章级文本中的句子级文本为二部图的文档向量;

确定所述正的篇章级文本中的词语和所述负的篇章级文本中的词语为二部图的词向量;

计算任意一个文档向量到该文档向量所包含的词向量的词转移概率;

依据所述词转移概率,计算任意一个文档向量到任意一个文档向量的文档转移概率;

依据每个文档转移概率和二部图的标签传播算法,计算每个句子级文本对应的正句子级概率和每个句子级文本对应的负句子级概率;

比较所述正句子级概率和负句子级概率的大小;

在比较结果为所述正句子级概率大于所述负句子级概率的情况下,确定该句子级文本的类别为正;

在比较结果为所述负句子级概率大于所述正句子级概率的情况下,确定该句子级文本的类别为负;

确定类别为正的句子级文本为正训练样本,确定类别为负的句子级文本为负训练样本;

对所述标记为客观的篇章级文本中的句子级文本进行人工标注,确定类别为客观的句子级文本,并确定类别为客观的句子级文本为客观训练样本;

利用所述正、负和客观训练样本对最大熵分类器进行训练,得到预设句子级情感分类单元。

2. 根据权利要求1所述的方法,其特征在于,利用所述正、负和客观训练样本对最大熵分类器进行训练,得到预设句子级情感分类单元的过程,包括:

确定所述正、负和客观训练样本中包含的词语为特征值;

$$\text{依据公式 } P(a,b) = \frac{\exp(\sum_{i=1}^k \lambda_i f_i(a,b))}{\sum_a \exp(\sum_{i=1}^k \lambda_i f_i(a,b))}, \text{ 分别计算每个特征值在 } a \text{ 分别为 } +1, -1 \text{ 和 } 0 \text{ 时, 对}$$

应的正预测条件概率、负预测条件概率和客观预测条件概率,其中,所述b为特征值,P( )为预测条件概率,exp( )为自然数e为底的指数函数,f<sub>i</sub>( )为二值特征函数,λ<sub>i</sub>为特征函数值f<sub>i</sub>(a,b)的权值且相同b对应的不同特征函数值的权值相同,

$\sum_{i=1}^k$  为对每个特征值对应的k个特征函数值进行求和的函数,K为不小于1的整数, $\sum_a$  为对a为不同值时对应的数据进行求和的函数;

利用GIS算法,调整每个特征值对应的正预测条件概率,直至每个特征值各自的正预测条件概率收敛,并将每个特征值各自收敛的正预测条件概率对应的 $\lambda$ 作为每个特征值各自的测试正 $\lambda$ ;

利用GIS算法,调整每个特征值对应的负预测条件概率,直至每个特征值各自的负预测条件概率收敛,并将每个特征值各自收敛的负预测条件概率对应的 $\lambda$ 作为每个特征值各自的测试负 $\lambda$ ;

利用GIS算法,调整每个特征值对应的客观预测条件概率,直至每个特征值各自的客观预测条件概率收敛,并将每个特征值各自收敛的客观预测条件概率对应的 $\lambda$ 作为每个特征值各自的测试客观 $\lambda$ ;

确定所述测试正 $\lambda$ 、测试负 $\lambda$ 和测试客观 $\lambda$ 对应的最大熵分类器为预设句子级情感分类单元。

3. 根据权利要求2所述的方法,其特征在于,所述利用所述预设句子级情感分类单元对已标注情感类别的篇章级文本中的句子级文本进行分类,包括:

确定所述已标注情感类别的篇章级文本中的句子级文本为待分类句子级文本;

确定所述待分类句子级文本所包含的词语为待分类特征值;

预设所述待分类句子级文本的情感类别分别为正、负和客观;

$$\text{依据公式 } P(a,b) = \frac{\exp(\sum_{i=1}^k \lambda_i f_i(a,b))}{\sum_a \exp(\sum_{i=1}^k \lambda_i f_i(a,b))}, \text{ 分别计算每个待分类特征值在 } a \text{ 分别为 } +1, -1 \text{ 和 } 0 \text{ 时, 对应的待分类正预测条件概率、待分类负预测条件概率和待分类客观预测条件概率, 其中, 所述 } b \text{ 为待分类特征值, } P(\quad) \text{ 为待分类预测条件概率, } \sum_{i=1}^k \text{ 为对每个待分类特征值对应的 } k \text{ 个特征函数值进行求和的函数, } K \text{ 为 } 3;}$$

其中,每个待分类特征值对应的各个特征函数值分别对应所述待分类句子级文本的预设情感类型正、负和客观,在计算待分类正预测条件概率时,若待分类特征值对应的词语包含在所述特征值中,则 $\lambda$ 为对应测试正 $\lambda$ ,否则 $\lambda$ 为0,在计算待分类负预测条件概率时,若待分类特征值对应的词语包含在所述特征值中,则 $\lambda$ 为对应测试负 $\lambda$ ,否则 $\lambda$ 为0,在计算待分类客观预测条件概率时,若待分类特征值对应的词语包含在所述特征值中,则 $\lambda$ 为测试客观 $\lambda$ ,否则 $\lambda$ 为0;

将各个待分类特征值对应的待分类正预测条件概率进行乘运算,得到句子级正预测条件概率,将各个待分类特征值对应的待分类负预测条件概率进行乘运算,得到句子级负预测条件概率,将各个待分类特征值对应的待分类客观预测条件概率进行乘运算,得到句子级客观预测条件概率;

比较从所述句子级正预测条件概率、所述句子级负预测条件概率和所述句子级客观预测条件概率的大小;

在比较结果为所述句子级正预测条件概率最大的情况下,确定所述待分类句子级文本的情感类别为正;

在比较结果为所述句子级负预测条件概率最大的情况下,确定所述待分类句子级文本

的情感类别为负；

在比较结果为所述句子级客观预测条件概率最大的情况下，确定所述待分类句子级文本的情感类别为客观。

4. 根据权利要求3所述的方法，其特征在于，在确定所述待分类句子级文本的情感类别之后，还包括：

判断所述待分类句子级文本的情感类别对应的句子级预测条件概率与0.5之差的绝对值是否小于0.05；

若是，将所述待分类句子级文本的情感类别转换为客观；

若否，所述待分类句子级文本的情感类别保持不变。

5. 一种句子级情感分类装置，其特征在于，包括：

第一获取单元，用于获取预设句子级情感分类单元；

预设句子级情感分类单元，用于对已标注情感类别的篇章级文本中的句子级文本进行分类；

其中，第一获取单元，包括：

第二获取单元，用于分别获取标记为正、负和客观的篇章级文本；

第一分类单元，用于对拆分所述篇章级文本获取到的句子级文本按照正、负和客观类型进行分类标记，得到对应的正、负和客观训练样本；

所述第一分类单元包括：

第一确定单元，用于确定所述标记为正的篇章级文本、所述标记为负的篇章级文本、所述标记为正的篇章级文本中的句子文本和所述标记为负的篇章级文本中的句子级文本为二部图的文档向量；

第二确定单元，用于确定所述正的篇章级文本中的词语和所述负的篇章级文本中的词语为二部图的词向量；

第一计算单元，用于计算任意一个文档向量到该文档向量所包含的词向量的词转移概率；

第二计算单元，用于依据所述词转移概率，计算任意一个文档向量到任意一个文档向量的文档转移概率；

第三计算单元，用于依据每个文档转移概率和二部图的标签传播算法，计算每个句子级文本对应的正句子级概率和每个句子级文本对应的负句子级概率；

第一比较单元，用于比较所述正句子级概率和负句子级概率的大小，在比较结果为所述正句子级概率大于所述负句子级概率的情况下，执行第三确定单元，在比较结果为所述负句子级概率大于所述正句子级概率的情况下，执行第四确定单元；

第三确定单元，用于确定该句子级文本的类别为正；

第四确定单元，用于确定该句子级文本的类别为负；

第五确定单元，用于确定类别为正的句子级文本为正训练样本，确定类别为负的句子级文本为负训练样本；

第六确定单元，用于对所述标记为客观的篇章级文本中的句子级文本进行人工标注，确定类别为客观的句子级文本，并确定类别为客观的句子级文本为客观训练样本；

训练单元，用于利用所述正、负和客观训练样本对最大熵分类器进行训练，得到预设句

子级情感分类单元。

6. 根据权利要求5所述的装置,其特征在于,所述训练单元包括:

第七确定单元,用于确定所述正、负和客观训练样本中包含的词语为特征值;

$$\text{第四计算单元,用于依据公式 } P(a,b) = \frac{\exp(\sum_{i=1}^k \lambda_i f_i(a,b))}{\sum_a \exp(\sum_{i=1}^k \lambda_i f_i(a,b))}, \text{ 分别计算每个特征值在a分}$$

别为+1、-1和0时,对应的正预测条件概率、负预测条件概率和客观预测条件概率,其中,所述b为特征值,P( )为预测条件概率,exp( )为自然数e为底的指数函数,f<sub>i</sub>( )为二值特征函数,λ<sub>i</sub>为特征函数值f<sub>i</sub>(a,b)的权值且相同b对应的不同特征函数值的权值相同,Σ<sub>i=1</sub><sup>k</sup>为对

每个特征值对应的k个特征函数值进行求和的函数,K为不小于1的整数,Σ<sub>a</sub>为对a为不同值时对应的数据进行求和的函数;

第一调整单元,用于利用GIS算法,调整每个特征值对应的正预测条件概率,直至每个特征值各自的正预测条件概率收敛,并将每个特征值各自收敛的正预测条件概率对应的λ作为每个特征值各自的测试正λ;

第二调整单元,用于利用GIS算法,调整每个特征值对应的负预测条件概率,直至每个特征值各自的负预测条件概率收敛,并将每个特征值各自收敛的负预测条件概率对应的λ作为每个特征值各自的测试负λ;

第三调整单元,用于利用GIS算法,调整每个特征值对应的客观预测条件概率,直至每个特征值各自的客观预测条件概率收敛,并将每个特征值各自收敛的客观预测条件概率对应的λ作为每个特征值各自的测试客观λ;

第八确定单元,用于确定所述测试正λ、测试负λ和测试客观λ对应的最大熵分类器为预设句子级情感分类单元。

7. 根据权利要求6所述的装置,其特征在于,所述预设句子级情感分类单元包括:

第九确定单元,用于确定所述已标注情感类别的篇章级文本中的句子级文本为待分类句子级文本;

第十确定单元,用于确定所述待分类句子级文本所包含的词语为待分类特征值;

预设单元,用于预设所述待分类句子级文本的情感类别分别为正、负和客观;

$$\text{第五计算单元,用于依据公式 } P(a,b) = \frac{\exp(\sum_{i=1}^k \lambda_i f_i(a,b))}{\sum_a \exp(\sum_{i=1}^k \lambda_i f_i(a,b))}, \text{ 分别计算每个待分类特征}$$

值在a分别为+1、-1和0时,对应的待分类正预测条件概率、待分类负预测条件概率和待分类客观预测条件概率,其中,所述b为待分类特征值,P( )为待分类预测条件概率,Σ<sub>i=1</sub><sup>k</sup>为对每个待分类特征值对应的k个特征函数值进行求和的函数,K为3;

其中,每个待分类特征值对应的各个特征函数值分别对应所述待分类句子级文本的预

设情感类型正、负和客观，在计算待分类正预测条件概率时，若待分类特征值对应的词语包含在所述特征值中，则 $\lambda$ 为对应测试正 $\lambda$ ，否则 $\lambda$ 为0，在计算待分类负预测条件概率时，若待分类特征值对应的词语包含在所述特征值中，则 $\lambda$ 为对应测试负 $\lambda$ ，否则 $\lambda$ 为0，在计算待分类客观预测条件概率时，若待分类特征值对应的词语包含在所述特征值中，则 $\lambda$ 为测试客观 $\lambda$ ，否则 $\lambda$ 为0；

第六计算单元，用于将各个待分类特征值对应的待分类正预测条件概率进行乘运算，得到句子级正预测条件概率，将各个待分类特征值对应的待分类负预测条件概率进行乘运算，得到句子级负预测条件概率，将各个待分类特征值对应的待分类客观预测条件概率进行乘运算，得到句子级客观预测条件概率；

第二比较单元，用于比较从所述句子级正预测条件概率、所述句子级负预测条件概率和所述句子级客观预测条件概率的大小，在比较结果为所述句子级正预测条件概率最大的情况下，执行第十一确定单元，在在比较结果为所述句子级负预测条件概率最大的情况下，执行第十二确定单元，在比较结果为所述句子级客观预测条件概率最大的情况下，执行第十三确定单元；

第十一确定单元，用于确定所述待分类句子级文本的情感类别为正；

第十二确定单元，用于确定所述待分类句子级文本的情感类别为负；

第十三确定单元，用于确定所述待分类句子级文本的情感类别为客观。

8. 根据权利要求7所述的装置，其特征在于，还包括：

判断单元，用于判断所述待分类句子级文本的情感类别对应的句子级预测条件概率与0.5之差的绝对值是否小于0.05，若是，执行转换单元，若否，所述待分类句子级文本的情感类别保持不变；

转换单元，用于将所述待分类句子级文本的情感类别转换为客观。

## 一种句子级情感分类方法及装置

### 技术领域

[0001] 本申请涉及自然语言处理及机器学习领域,特别涉及一种句子级情感分类方法及装置。

### 背景技术

[0002] 随着互联网的快速发展,互联网所倡导“以用户为中心,用户参与”的开放式构架理念,使得互联网用户由被动地接受互联网信息向主动创造互联网信息转变。因此,互联网(如博客和论坛)上产生了大量用户参与的、对于诸如人物、事件、产品等有价值的评论信息,这些评论信息表达了用户的各种情感色彩和情感倾向性,如喜、怒、哀、乐和批评、赞扬等。用户可以通过浏览这些主观色彩的评论来了解大众舆论对于某一事件或产品的看法。越来越多的用户乐于在互联网上分享自己的观点或体验,导致评论信息量迅速增加,仅靠人工的方法难以应付网上海量信息的收集和处理,因此文本情感分析技术应运而生。

[0003] 文本情感分析技术利用计算机快速获取和整理相关评价信息,其可以对带有情感色彩的主观性文本进行分析、处理、归纳和推理。

[0004] 情感分类是文本情感分析技术的一项子任务,其利用底层情感信息抽取的结果将情感文本分为若干情感类别,如分为褒贬两类或者其他更细致的情感类别。

[0005] 目前,基于计算机的情感分类,主要对篇章级文本进行分类,对句子级文本进行分类则主要采用人工逐句标注的方式,但是采用人工逐句标注的方式对句子级文本进行分类,效率低,且由于需要支付人工费,因此长期使用人工逐句标注的方式,投资大,成本高。

[0006] 由上可见,采用人工逐句标注的方式对句子级文本进行分类,存在效率低,投资大,成本高的缺点。

### 发明内容

[0007] 为解决上述技术问题,本申请实施例提供一种句子级情感分类方法及装置,以达到提高工作效率,不需要支付人工费,降低了成本的目的,技术方案如下:

[0008] 一种句子级情感分类方法,包括:

[0009] 获取预设句子级情感分类单元;

[0010] 利用所述预设句子级情感分类单元对已标注情感类别的篇章级文本中的句子级文本进行分类;

[0011] 其中,所述预设句子级情感分类单元的获取过程,包括:

[0012] 分别获取标记为正、负和客观的篇章级文本;

[0013] 确定所述标记为正的篇章级文本、所述标记为负的篇章级文本、所述标记为正的篇章级文本中的句子文本和所述标记为负的篇章级文本中的句子级文本为二部图的文档向量;

[0014] 确定所述正的篇章级文本中的词语和所述负的篇章级文本中的词语为二部图的词向量;

- [0015] 计算任意一个文档向量到该文档向量所包含的词向量的词转移概率；
- [0016] 依据所述词转移概率，计算任意一个文档向量到任意一个文档向量的文档转移概率；
- [0017] 依据每个文档转移概率和二部图的标签传播算法，计算每个句子级文本对应的正句子级概率和每个句子级文本对应的负句子级概率；
- [0018] 比较所述正句子级概率和负句子级概率的大小；
- [0019] 在比较结果为所述正句子级概率大于所述负句子级概率的情况下，确定该句子级文本的类别为正；
- [0020] 在比较结果为所述负句子级概率大于所述正句子级概率的情况下，确定该句子级文本的类别为负；
- [0021] 确定类别为正的句子级文本为正训练样本，确定类别为负的句子级文本为负训练样本；
- [0022] 对所述标记为客观的篇章级文本中的句子级文本进行人工标注，确定类别为客观的句子级文本，并确定类别为客观的句子级文本为客观训练样本；
- [0023] 利用所述正、负和客观训练样本对最大熵分类器进行训练，得到预设句子级情感分类单元。
- [0024] 优选的，利用所述正、负和客观训练样本对最大熵分类器进行训练，得到预设句子级情感分类单元的过程，包括：
- [0025] 确定所述正、负和客观训练样本中包含的词语为特征值；

- [0026] 依据公式  $P(a,b) = \frac{\exp(\sum_{i=1}^k \lambda_i f_i(a,b))}{\sum_a \exp(\sum_{i=1}^k \lambda_i f_i(a,b))}$ ，分别计算每个特征值在a分别为+1、-1和0时，对应的正预测条件概率、负预测条件概率和客观预测条件概率，其中，所述b为特征值，P()为预测条件概率， $\exp()$ 为自然数e为底的指数函数， $f_i()$ 为二值特征函数， $\lambda_i$ 为特征函数值 $f_i(a,b)$ 的权值且相同b对应的不同特征函数值的权值相同， $\sum_{i=1}^k$ 为对每个特征值对应的k个特征函数值进行求和的函数，K为不小于1的整数， $\sum_a$ 为对a为不同值时对应的数据进行求和的函数；

- [0027] 利用GIS算法，调整每个特征值对应的正预测条件概率，直至每个特征值各自的正预测条件概率收敛，并将每个特征值各自收敛的正预测条件概率对应的 $\lambda$ 作为每个特征值各自的测试正 $\lambda$ ；
- [0028] 利用GIS算法，调整每个特征值对应的负预测条件概率，直至每个特征值各自的负预测条件概率收敛，并将每个特征值各自收敛的负预测条件概率对应的 $\lambda$ 作为每个特征值各自的测试负 $\lambda$ ；
- [0029] 利用GIS算法，调整每个特征值对应的客观预测条件概率，直至每个特征值各自的客观预测条件概率收敛，并将每个特征值各自收敛的客观预测条件概率对应的 $\lambda$ 作为每个特征值各自的测试客观 $\lambda$ ；

[0030] 确定所述测试正 $\lambda$ 、测试负 $\lambda$ 和测试客观 $\lambda$ 对应的最大熵分类器为预设句子级情感分类单元。

[0031] 优选的,所述利用所述预设句子级情感分类单元对已标注情感类别的篇章级文本中的句子级文本进行分类,包括:

[0032] 确定所述已标注情感类别的篇章级文本中的句子级文本为待分类句子级文本;

[0033] 确定所述待分类句子级文本所包含的词语为待分类特征值;

[0034] 预设所述待分类句子级文本的情感类别分别为正、负和客观;

[0035] 依据公式  $P(a,b) = \frac{\exp(\sum_{i=1}^k \lambda_i f_i(a,b))}{\sum_a \exp(\sum_{i=1}^k \lambda_i f_i(a,b))}$ , 分别计算每个待分类特征值在a分别为+1、-1和0时,对应的待分类正预测条件概率、待分类负预测条件概率和待分类客观预测条件概率,其中,所述b为待分类特征值,P()为待分类预测条件概率, $\sum_{i=1}^k$  为对每个待分类特征值对应的k个特征函数值进行求和的函数,K为3;

[0036] 其中,每个待分类特征值对应的各个特征函数值分别对应所述待分类句子级文本的预设情感类型正、负和客观,在计算待分类正预测条件概率时,若待分类特征值对应的词语包含在所述特征值中,则 $\lambda$ 为对应测试正 $\lambda$ ,否则 $\lambda$ 为0,在计算待分类负预测条件概率时,若待分类特征值对应的词语包含在所述特征值中,则 $\lambda$ 为对应测试负 $\lambda$ ,否则 $\lambda$ 为0,在计算待分类客观预测条件概率时,若待分类特征值对应的词语包含在所述特征值中,则 $\lambda$ 为测试客观 $\lambda$ ,否则 $\lambda$ 为0;

[0037] 将各个待分类特征值对应的待分类正预测条件概率进行乘运算,得到句子级正预测条件概率,将各个待分类特征值对应的待分类负预测条件概率进行乘运算,得到句子级负预测条件概率,将各个待分类特征值对应的待分类客观预测条件概率进行乘运算,得到句子级客观预测条件概率;

[0038] 比较从所述句子级正预测条件概率、所述句子级负预测条件概率和所述句子级客观预测条件概率的大小;

[0039] 在比较结果为所述句子级正预测条件概率最大的情况下,确定所述待分类句子级文本的情感类别为正;

[0040] 在比较结果为所述句子级负预测条件概率最大的情况下,确定所述待分类句子级文本的情感类别为负;

[0041] 在比较结果为所述句子级客观预测条件概率最大的情况下,确定所述待分类句子级文本的情感类别为客观。

[0042] 优选的,在确定所述待分类句子级文本的情感类别之后,还包括:

[0043] 判断所述待分类句子级文本的情感类别对应的句子级预测条件概率与0.5之差的绝对值是否小于0.05;

[0044] 若是,将所述待分类句子级文本的情感类别转换为客观;

[0045] 若否,所述待分类句子级文本的情感类别保持不变。

[0046] 一种句子级情感分类装置,包括:

- [0047] 第一获取单元,用于获取预设句子级情感分类单元;
- [0048] 预设句子级情感分类单元,用于对已标注情感类别的篇章级文本中的句子级文本进行分类;
- [0049] 其中,第一获取单元,包括:
- [0050] 第二获取单元,用于分别获取标记为正、负和客观的篇章级文本;
- [0051] 第一分类单元,用于对拆分所述篇章级文本获取到的句子级文本按照正、负和客观类型进行分类标记,得到对应的正、负和客观训练样本;
- [0052] 所述第一分类单元包括:
- [0053] 第一确定单元,用于确定所述标记为正的篇章级文本、所述标记为负的篇章级文本、所述标记为正的篇章级文本中的句子文本和所述标记为负的篇章级文本中的句子级文本为二部图的文档向量;
- [0054] 第二确定单元,用于确定所述正的篇章级文本中的词语和所述负的篇章级文本中的词语为二部图的词向量;
- [0055] 第一计算单元,用于计算任意一个文档向量到该文档向量所包含的词向量的词转移概率;
- [0056] 第二计算单元,用于依据所述词转移概率,计算任意一个文档向量到任意一个文档向量的文档转移概率;
- [0057] 第三计算单元,用于依据每个文档转移概率和二部图的标签传播算法,计算每个句子级文本对应的正句子级概率和每个句子级文本对应的负句子级概率;
- [0058] 第一比较单元,用于比较所述正句子级概率和负句子级概率的大小,在比较结果为所述正句子级概率大于所述负句子级概率的情况下,执行第三确定单元,在比较结果为所述负句子级概率大于所述正句子级概率的情况下,执行第四确定单元;
- [0059] 第三确定单元,用于确定该句子级文本的类别为正;
- [0060] 第四确定单元,用于确定该句子级文本的类别为负;
- [0061] 第五确定单元,用于确定类别为正的句子级文本为正训练样本,确定类别为负的句子级文本为负训练样本;
- [0062] 第六确定单元,用于对所述标记为客观的篇章级文本中的句子级文本进行人工标注,确定类别为客观的句子级文本,并确定类别为客观的句子级文本为客观训练样本;
- [0063] 训练单元,用于利用所述正、负和客观训练样本对最大熵分类器进行训练,得到预设句子级情感分类单元。
- [0064] 优选的,所述训练单元包括:
- [0065] 第七确定单元,用于确定所述正、负和客观训练样本中包含的词语为特征值;

- [0066] 第四计算单元,用于依据公式  $P(a,b) = \frac{\exp(\sum_{i=1}^k \lambda_i f_i(a,b))}{\sum_a \exp(\sum_{i=1}^k \lambda_i f_i(a,b))}$ , 分别计算每个特征值

在a分别为+1、-1和0时,对应的正预测条件概率、负预测条件概率和客观预测条件概率,其中,所述b为特征值,P()为预测条件概率,exp()为自然数e为底的指数函数,f<sub>i</sub>()为二值特

征函数,  $\lambda_i$  为特征函数值  $f_i(a, b)$  的权值且相同  $b$  对应的不同特征函数值的权值相同,  $\sum_{i=1}^k$  为

对每个特征值对应的  $k$  个特征函数值进行求和的函数,  $K$  为不小于 1 的整数,  $\sum_a$  为对  $a$  为不同值时对应的数据进行求和的函数;

[0067] 第一调整单元, 用于利用 GIS 算法, 调整每个特征值对应的正预测条件概率, 直至每个特征值各自的正预测条件概率收敛, 并将每个特征值各自收敛的正预测条件概率对应的  $\lambda$  作为每个特征值各自的测试正  $\lambda$ ;

[0068] 第二调整单元, 用于利用 GIS 算法, 调整每个特征值对应的负预测条件概率, 直至每个特征值各自的负预测条件概率收敛, 并将每个特征值各自收敛的负预测条件概率对应的  $\lambda$  作为每个特征值各自的测试负  $\lambda$ ;

[0069] 第三调整单元, 用于利用 GIS 算法, 调整每个特征值对应的客观预测条件概率, 直至每个特征值各自的客观预测条件概率收敛, 并将每个特征值各自收敛的客观预测条件概率对应的  $\lambda$  作为每个特征值各自的测试客观  $\lambda$ ;

[0070] 第八确定单元, 用于确定所述测试正  $\lambda$ 、测试负  $\lambda$  和测试客观  $\lambda$  对应的最大熵分类器为预设句子级情感分类单元。

[0071] 优选的, 所述预设句子级情感分类单元包括:

[0072] 第九确定单元, 用于确定所述已标注情感类别的篇章级文本中的句子级文本为待分类句子级文本;

[0073] 第十确定单元, 用于确定所述待分类句子级文本所包含的词语为待分类特征值;

[0074] 预设单元, 用于预设所述待分类句子级文本的情感类别分别为正、负和客观;

[0075] 第五计算单元, 用于依据公式  $P(a, b) = \frac{\exp(\sum_{i=1}^k \lambda_i f_i(a, b))}{\sum_a \exp(\sum_{i=1}^k \lambda_i f_i(a, b))}$ , 分别计算每个待分类特征值在  $a$  分别为 +1、-1 和 0 时, 对应的待分类正预测条件概率、待分类负预测条件概率和待分类客观预测条件概率, 其中, 所述  $b$  为待分类特征值,  $P()$  为待分类预测条件概率,  $\sum_{i=1}^k$  为对每个待分类特征值对应的  $k$  个特征函数值进行求和的函数,  $K$  为 3;

[0076] 其中, 每个待分类特征值对应的各个特征函数值分别对应所述待分类句子级文本的预设情感类型正、负和客观, 在计算待分类正预测条件概率时, 若待分类特征值对应的词语包含在所述特征值中, 则  $\lambda$  为对应测试正  $\lambda$ , 否则  $\lambda$  为 0, 在计算待分类负预测条件概率时, 若待分类特征值对应的词语包含在所述特征值中, 则  $\lambda$  为对应测试负  $\lambda$ , 否则  $\lambda$  为 0, 在计算待分类客观预测条件概率时, 若待分类特征值对应的词语包含在所述特征值中, 则  $\lambda$  为测试客观  $\lambda$ , 否则  $\lambda$  为 0;

[0077] 第六计算单元, 用于将各个待分类特征值对应的待分类正预测条件概率进行乘运算, 得到句子级正预测条件概率, 将各个待分类特征值对应的待分类负预测条件概率进行乘运算, 得到句子级负预测条件概率, 将各个待分类特征值对应的待分类客观预测条件概率进行乘运算, 得到句子级客观预测条件概率;

[0078] 第二比较单元,用于比较从所述句子级正预测条件概率、所述句子级负预测条件概率和所述句子级客观预测条件概率的大小,在比较结果为所述句子级正预测条件概率最大的情况下,执行第十一确定单元,在在比较结果为所述句子级负预测条件概率最大的情况下,执行第十二确定单元,在比较结果为所述句子级客观预测条件概率最大的情况下,执行第十三确定单元;

[0079] 第十一确定单元,用于确定所述待分类句子级文本的情感类别为正;

[0080] 第十二确定单元,用于确定所述待分类句子级文本的情感类别为负;

[0081] 第十三确定单元,用于确定所述待分类句子级文本的情感类别为客观。

[0082] 优选的,还包括:

[0083] 判断单元,用于判断所述待分类句子级文本的情感类别对应的句子级预测条件概率与0.5之差的绝对值是否小于0.05,若是,执行转换单元,若否,所述待分类句子级文本的情感类别保持不变;

[0084] 转换单元,用于将所述待分类句子级文本的情感类别转换为客观。

[0085] 与现有技术相比,本申请的有益效果为:

[0086] 在本申请中,获取预设句子级情感分类单元;利用所述预设句子级情感分类单元对已标注情感类别的篇章级文本中的句子级文本进行分类;其中,所述预设句子级情感分类器的获取过程,包括:分别获取标记为正、负和客观的篇章级文本;对拆分所述篇章级文本获取到的句子级文本按照正、负和客观类型进行分类标记,得到对应的正、负和客观训练样本;利用所述正、负和客观训练样本对最大熵分类器进行训练,得到预设句子级情感分类单元。

[0087] 因此,本申请通过获取到的预设句子级情感分类单元,可以自动对已标注情感类别的篇章级文本中的句子级文本进行分类,相比于采用人工标注的方式对句子级文本进行分类,提高了工作效率,且由于不需要再使用人工来标注,因此不需要支付人工费,降低了成本。

## 附图说明

[0088] 为了更清楚地说明本申请实施例中的技术方案,下面将对实施例描述中所需要使用的附图作简单地介绍,显而易见地,下面描述中的附图仅仅是本申请的一些实施例,对于本领域普通技术人员来讲,在不付出创造性劳动性的前提下,还可以根据这些附图获得其他的附图。

[0089] 图1是本申请提供的一种句子级情感分类方法的一种流程图;

[0090] 图2是本申请提供的一种句子级情感分类方法的一种子流程图;

[0091] 图3是本申请提供的一种句子级情感分类方法的再一种流程图;

[0092] 图4是本申请提供的一种句子级情感分类方法的再一种流程图;

[0093] 图5是本申请提供的一种句子级情感分类方法的再一种子流程图;

[0094] 图6是本申请提供的一种句子级情感分类方法的再一种流程图;

[0095] 图7是本申请提供的一种句子级情感分类装置的一种结构示意图;

[0096] 图8是本申请提供的第一分类单元的一种结构示意图;

[0097] 图9是本申请提供的一种训练单元的一种结构示意图;

[0098] 图10是本申请提供的一种预设句子级情感分类单元的一种结构示意图。

## 具体实施方式

[0099] 下面将结合本申请实施例中的附图,对本申请实施例中的技术方案进行清楚、完整地描述,显然,所描述的实施例仅仅是本申请一部分实施例,而不是全部的实施例。基于本申请中的实施例,本领域普通技术人员在没有做出创造性劳动前提下所获得的所有其他实施例,都属于本申请保护的范围。

[0100] 一个实施例

[0101] 请参见图1,其示出了本申请提供的一种句子级情感分类方法的一种流程图,可以包括以下步骤:

[0102] 步骤S11:获取预设句子级情感分类单元。

[0103] 在本实施例中,预设句子级情感分类单元的获取过程可以参见图2,图2示出了本申请提供的一种句子级情感分类方法的一种子流程图,可以包括以下步骤:

[0104] 步骤S21:分别获取标记为正、负和客观的篇章级文本。

[0105] 其中,标记为正、负的篇章级文本可以为自动标记为正、负的篇章级文本,标记为客观的篇章级文本为人工标记为客观的篇章级文本。获取到的标记为正、负和客观的篇章级文本分别对应的数量可以相同。

[0106] 标记为正、负和客观的篇章级文本可以来自于DVD领域。标记为正、负的篇章级文本可以根据DVD领域已存在的星级自动获取,标记为客观的篇章级文本则通过人工标记的方式从DVD领域中获取。

[0107] 步骤S22:对拆分所述篇章级文本获取到的句子级文本按照正、负和客观类型进行分类标记,得到对应的正、负和客观训练样本。

[0108] 分别对标记为正、负和客观的篇章级文本进行拆分,得到句子级文本,并对拆分得到的句子级文本按照正、负和客观类型进行分类标记,得到对应的正、负和客观训练样本。

[0109] 步骤S23:利用所述正、负和客观训练样本对最大熵分类器进行训练,得到预设句子级情感分类单元。

[0110] 步骤S12:利用所述预设句子级情感分类单元对已标注情感类别的篇章级文本中的句子级文本进行分类。

[0111] 在本实施例中,利用预设句子级情感分类单元可以对已标注情感类别的篇章级文本中的句子级文本进行自动分类。具体的,利用预设句子级情感分类单元可以对已标注情感类别的篇章级文本中的句子级文本自动分类为正、负或客观。

[0112] 在本实施例中,预设句子级情感分类单元的获取过程不需要循环执行,执行一次即可,在获取预设句子级情感分类单元后,可以使用获取到的预设句子级情感分类单元对每个已标注情感类别的篇章级文本中的每个句子级文本进行自动分类。

[0113] 在本申请中,获取预设句子级情感分类单元;利用所述预设句子级情感分类单元对已标注情感类别的篇章级文本中的句子级文本进行分类;其中,所述预设句子级情感分类器的获取过程,包括:分别获取标记为正、负和客观的篇章级文本;对拆分所述篇章级文本获取到的句子级文本按照正、负和客观类型进行分类标记,得到对应的正、负和客观训练样本;利用所述正、负和客观训练样本对最大熵分类器进行训练,得到预设句子级情感分类

单元。

[0114] 因此,本申请通过获取到的预设句子级情感分类单元,可以自动对已标注情感类别的篇章级文本中的句子级文本进行分类,相比于采用人工标注的方式对句子级文本进行分类,提高了工作效率,且由于不需要再使用人工来标注,因此不需要支付人工费,降低了成本。

[0115] 另一个实施例

[0116] 在本实施例中,示出的是对拆分所述篇章级文本获取到的句子级文本按照正、负和客观类型进行分类标记,得到对应的正、负和客观训练样本的过程,请参见图3,图3示出了本申请提供的一种句子级情感分类方法的再一种流程图,可以包括以下步骤:

[0117] 步骤S31:确定所述标记为正的篇章级文本、所述标记为负的篇章级文本、所述标记为正的篇章级文本中的句子文本和所述标记为负的篇章级文本中的句子级文本为二部图的文档向量。

[0118] 步骤S32:确定所述正篇章级文本中的词语和所述负篇章级文本中的词语为二部图的词向量。

[0119] 步骤S33:计算任意一个文档向量到该文档向量所包含的词向量的词转移概率。

[0120] 步骤S34:依据所述词转移概率,计算任意一个文档向量到任意一个文档向量的文档转移概率。

[0121] 步骤S35:依据每个文档转移概率和二部图的标签传播算法,计算每个句子级文本对应的正句子级概率和每个句子级文本对应的负句子级概率。

[0122] 步骤S36:比较所述正句子级概率和负句子级概率的大小。

[0123] 在比较结果为正句子级概率大于负句子级概率的情况下,确定该句子级文本的类别为正,在比较结果为负句子级概率大于正句子级概率的情况下,确定该句子级文本的类别为负。

[0124] 步骤S37:确定类别为正的句子级文本为正训练样本,确定类别为负的句子级文本为负训练样本。

[0125] 步骤S38:对所述标记为客观的篇章级文本中的句子级文本进行人工标注,确定类别为客观的句子级文本,并确定类别为客观的句子级文本为客观训练样本。

[0126] 当然,在确定正训练样本和负训练样本的执行过程除了步骤S31至步骤S37的实现方式之外,还可以通过人工标注的方式,确定正训练样本和负训练样本。

[0127] 再一个实施例

[0128] 在本实施例中,示出的是利用所述正、负和客观训练样本对最大熵分类器进行训练,得到预设句子级情感分类单元的过程,请参见图4,图4示出的是本申请提供的一种句子级情感分类方法的再一种流程图,可以包括以下步骤:

[0129] 步骤S41:确定所述正、负和客观训练样本中包含的词语为特征值。

[0130] 步骤S42:依据公式  $P(a,b) = \frac{\exp(\sum_{i=1}^k \lambda_i f_i(a,b))}{\sum_a \exp(\sum_{i=1}^k \lambda_i f_i(a,b))}$ , 分别计算每个特征值在a分别为+1、-1和0时,对应的正预测条件概率、负预测条件概率和客观预测条件概率。

[0131] 其中,  $P(a, b) = \frac{\exp(\sum_{i=1}^k \lambda_i f_i(a, b))}{\sum_a \exp(\sum_{i=1}^k \lambda_i f_i(a, b))}$  中的b为特征值,P()为预测条件概率,exp()为自然数e为底的指数函数,f<sub>i</sub>()为二值特征函数,λ<sub>i</sub>为特征函数值f<sub>i</sub>(a,b)的权值且相同b对应的不同特征函数值的权值相同, $\sum_{i=1}^k$ 为对每个特征值对应的k个特征函数值进行求和的函数,K为不小于1的整数, $\sum_a$ 为对a为不同值时对应的数据进行求和的函数。

[0132] 现举例对依据公式,分别计算每个特征值在a分别为+1、-1和0时,对应的正预测条件概率、负预测条件概率和客观预测条件概率的过程进行详细说明。例如,三个训练样本,序号分别为1、2和3,序号为1的训练样本为正训练样本,且正训练样本包括词语x,y,序号为2的训练样本为负训练样本,且负训练样本包括词语c,d,x,序号为3的训练样本为客观训练样本,客观训练样本包括词语e,y。

[0133] 以x为例,对在a分别为+1、-1和0时,对应的正预测条件概率、负预测条件概率和客观预测条件概率的过程进行说明。

[0134] x在正训练样本中和负训练样本中都存在,因此在a为+1时,x对应两个特征函数值,分别为f<sub>1</sub>(1,x)、f<sub>2</sub>(1,x),在a为-1时,对应两个特征函数值,分别为f<sub>1</sub>(-1,x)、f<sub>2</sub>(-1,x),在a为0时,对应两个特征函数值,分别为f<sub>1</sub>(0,x)、f<sub>2</sub>(0,x)。由于相同b对应的不同特征函数值的权值相同,因此x对应的f<sub>1</sub>(1,x)的权值和对应f<sub>2</sub>(1,x)的权值相同,记为λ<sub>1</sub>;x对应的f<sub>1</sub>(-1,x)的权值和对应f<sub>2</sub>(-1,x)的权值相同,记为λ<sub>2</sub>;x对应的f<sub>1</sub>(0,x)的权值和对应f<sub>2</sub>(0,x)的权值相同,记为λ<sub>3</sub>。

[0135] 在a为+1时,依据公式  $P(a, b) = \frac{\exp(\sum_{i=1}^k \lambda_i f_i(a, b))}{\sum_a \exp(\sum_{i=1}^k \lambda_i f_i(a, b))}$ , 可以得到

[0136]

$$P(1, x) = \frac{\exp(\sum_{i=1}^2 \lambda_i f_i(1, x))}{\sum_a \exp(\sum_{i=1}^2 \lambda_i f_i(1, x))} =$$

[0137]

$$P(1, x) = \frac{\exp(\lambda_1 f_1(1, x) + \lambda_2 f_2(1, x))}{\exp(\lambda_1 f_1(1, x) + \lambda_2 f_2(1, x)) + \exp(\lambda_1 f_1(-1, x) + \lambda_2 f_2(-1, x)) + \exp(\lambda_3 f_1(0, x) + \lambda_3 f_2(0, x))}$$

。

[0138]

$$P(1, x) = \frac{\exp(\lambda_1 f_1(1, x) + \lambda_2 f_2(1, x))}{\exp(\lambda_1 f_1(1, x) + \lambda_2 f_2(1, x)) + \exp(\lambda_1 f_1(-1, x) + \lambda_2 f_2(-1, x)) + \exp(\lambda_3 f_1(0, x) + \lambda_3 f_2(0, x))}$$

[0139] 即x的正预测条件概率。

[0140] 在a为-1时,依据公式  $P(a,b) = \frac{\exp(\sum_{i=1}^k \lambda_i f_i(a,b))}{\sum_a \exp(\sum_{i=1}^k \lambda_i f_i(a,b))}$ , 可以得到

[0141]

$$P(-1,x) = \frac{\exp(\sum_{i=1}^2 \lambda_i f_i(-1,x))}{\sum_a \exp(\sum_{i=1}^2 \lambda_i f_i(-1,x))} =$$

[0142]

$$P(-1,x) = \frac{\exp(\lambda_2 f_1(-1,x) + \lambda_2 f_2(-1,x))}{\exp(\lambda_1 f_1(1,x) + \lambda_1 f_2(1,x)) + \exp(\lambda_2 f_1(-1,x) + \lambda_2 f_2(-1,x)) + \exp(\lambda_3 f_1(0,x) + \lambda_3 f_2(0,x))}.$$

[0143]

$$P(-1,x) = \frac{\exp(\lambda_2 f_1(-1,x) + \lambda_2 f_2(-1,x))}{\exp(\lambda_1 f_1(1,x) + \lambda_1 f_2(1,x)) + \exp(\lambda_2 f_1(-1,x) + \lambda_2 f_2(-1,x)) + \exp(\lambda_3 f_1(0,x) + \lambda_3 f_2(0,x))}$$

[0144] 即x的负预测条件概率。

[0145] 在a为0时,依据公式  $P(a,b) = \frac{\exp(\sum_{i=1}^k \lambda_i f_i(a,b))}{\sum_a \exp(\sum_{i=1}^k \lambda_i f_i(a,b))}$ , 可以得到

[0146]

$$P(0,x) = \frac{\exp(\sum_{i=1}^2 \lambda_i f_i(0,x))}{\sum_a \exp(\sum_{i=1}^2 \lambda_i f_i(0,x))} =$$

[0147]

$$P(0,x) = \frac{\exp(\lambda_3 f_1(0,x) + \lambda_3 f_2(0,x))}{\exp(\lambda_1 f_1(1,x) + \lambda_1 f_2(1,x)) + \exp(\lambda_2 f_1(-1,x) + \lambda_2 f_2(-1,x)) + \exp(\lambda_3 f_1(0,x) + \lambda_3 f_2(0,x))}$$

◦

[0148]

$$P(0,x) = \frac{\exp(\lambda_3 f_1(0,x) + \lambda_3 f_2(0,x))}{\exp(\lambda_1 f_1(1,x) + \lambda_1 f_2(1,x)) + \exp(\lambda_2 f_1(-1,x) + \lambda_2 f_2(-1,x)) + \exp(\lambda_3 f_1(0,x) + \lambda_3 f_2(0,x))}$$

[0149] 即x的客观预测条件概率。

[0150] 词语y,c,d,e各自对应的正预测条件概率、负预测条件概率和客观预测条件概率的计算过程如上述x对应的正预测条件概率、负预测条件概率和客观预测条件概率的计算过程,在此不再赘述。

[0151] 每个特征值各自对应的正预测条件概率、负预测条件概率和客观预测条件概率的计算过程也如上述以x为例的计算过程,在此不再赘述。

[0152] 步骤S43:利用GIS算法,调整每个特征值对应的正预测条件概率,直至每个特征值各自的正预测条件概率收敛,并将每个特征值各自收敛的正预测条件概率对应的λ作为每

个特征值各自的测试正 $\lambda$ 。

[0153] 步骤S44:利用GIS算法,调整每个特征值对应的负预测条件概率,直至每个特征值各自的负预测条件概率收敛,并将每个特征值各自收敛的负预测条件概率对应的 $\lambda$ 作为每个特征值各自的测试负 $\lambda$ 。

[0154] 步骤S45:利用GIS算法,调整每个特征值对应的客观预测条件概率,直至每个特征值各自的客观预测条件概率收敛,并将每个特征值各自收敛的客观预测条件概率对应的 $\lambda$ 作为每个特征值各自的测试客观 $\lambda$ 。

[0155] 步骤S46:确定所述测试正 $\lambda$ 、测试负 $\lambda$ 和测试客观 $\lambda$ 对应的最大熵分类器为预设句子级情感分类单元。

[0156] 再一个实施例

[0157] 在本实施例中,示出的是利用预设句子级情感分类单元对已标注情感类别的篇章级文本中的句子级文本进行分类的过程,请参见图5,图5示出的是本申请提供的一种句子级情感分类方法的再一种子流程图,可以包括以下步骤:

[0158] 步骤S51:确定所述已标注情感类别的篇章级文本中的句子级文本为待分类句子级文本。

[0159] 步骤S52:确定所述待分类句子级文本所包含的词语为待分类特征值。

[0160] 步骤S53:预设所述待分类句子级文本的情感类别分别为正、负和客观。

[0161] 步骤S54:依据公式  $P(a,b) = \frac{\exp(\sum_{i=1}^k \lambda_i f_i(a,b))}{\sum_a \exp(\sum_{i=1}^k \lambda_i f_i(a,b))}$ , 分别计算每个待分类特征值在a分别为+1、-1和0时,对应的待分类正预测条件概率、待分类负预测条件概率和待分类客观预测条件概率。

[0162] 其中,所述b为待分类特征值,P()为待分类预测条件概率, $\sum_{i=1}^k$ 为对每个待分类特征值对应的k个特征函数值进行求和的函数,K为3。

[0163] 其中,每个待分类特征值对应的各个特征函数值分别对应所述待分类句子级文本的预设情感类型正、负和客观,在计算待分类正预测条件概率时,若待分类特征值对应的词语包含在所述特征值中,则 $\lambda$ 为对应测试正 $\lambda$ ,否则 $\lambda$ 为0,在计算待分类负预测条件概率时,若待分类特征值对应的词语包含在所述特征值中,则 $\lambda$ 为对应测试负 $\lambda$ ,否则 $\lambda$ 为0,在计算待分类客观预测条件概率时,若待分类特征值对应的词语包含在所述特征值中,则 $\lambda$ 为测试客观 $\lambda$ ,否则 $\lambda$ 为0。

[0164] 现举例对依据公式  $P(a,b) = \frac{\exp(\sum_{i=1}^k \lambda_i f_i(a,b))}{\sum_a \exp(\sum_{i=1}^k \lambda_i f_i(a,b))}$ , 分别计算每个待分类特征值在a分别为+1、-1和0时,对应的待分类正预测条件概率、待分类负预测条件概率和待分类客观预测条件概率的过程进行说明。

[0165] 例如,待分类句子级文本包括词语x,e,h。而特征值为x,y,c,d,e,则在计算h的待

分类正预测条件概率、待分类负预测条件概率和待分类客观预测条件概率时,  $\lambda$  为 0。

[0166] 在计算  $x$  对应的待分类正预测条件概率、待分类负预测条件概率和待分类客观预测条件概率时,  $\lambda$  为  $x$  对应的测试正  $\lambda$ 、测试负  $\lambda$  和测试客观  $\lambda$ 。

[0167] 在计算  $e$  对应的待分类正预测条件概率、待分类负预测条件概率和待分类客观预测条件概率时,  $\lambda$  为  $e$  对应的测试正  $\lambda$ 、测试负  $\lambda$  和测试客观  $\lambda$ 。

[0168] 以  $x$  为例, 对。。。进行说明。令  $x$  对应的测试正  $\lambda$  为  $\lambda'_1$ , 测试负  $\lambda$  为  $\lambda'_2$ , 测试客观  $\lambda$  为  $\lambda'_3$ 。在  $a$  为 +1 时,  $x$  在待分类句子级文本的预设情感类别分别为正、负和客观时对应的特征函数值分别为  $f_1(1, x)$ 、 $f_{-1}(1, x)$  和  $f_0(1, x)$ ; 在  $a$  为 -1 时,  $x$  在待分类句子级文本的预设情感类别分别为正、负和客观时对应的特征函数值分别为  $f_1(-1, x)$ 、 $f_{-1}(-1, x)$  和  $f_0(-1, x)$ ; 在  $a$  为 0 时,  $x$  在待分类句子级文本的预设情感类别分别为正、负和客观时对应的特征函数值分别为  $f_1(0, x)$ 、 $f_{-1}(0, x)$  和  $f_0(0, x)$ 。

[0169] 在  $a$  为 +1 时, 依据公式  $P(a, b) = \frac{\exp(\sum_{i=1}^k \lambda_i f_i(a, b))}{\sum_a \exp(\sum_{i=1}^k \lambda_i f_i(a, b))}$ , 可以得到

[0170]

$$P(1, x) = \frac{\exp(\sum_{i=1}^3 \lambda_i f_i(1, x))}{\sum_a \exp(\sum_{i=1}^3 \lambda_i f_i(1, x))} =$$

[0171]

$$P(1, x) = \frac{\exp(\lambda'_1 f_1(1, x) + \lambda'_1 f_{-1}(1, x) + \lambda'_1 f_0(1, x))}{\exp(\lambda'_1 f_1(1, x) + \lambda'_1 f_{-1}(1, x) + \lambda'_1 f_0(1, x)) + \exp(\lambda'_2 f_1(-1, x) + \lambda'_2 f_{-1}(-1, x) + \lambda'_2 f_0(-1, x)) + \exp(\lambda'_3 f_1(0, x) + \lambda'_3 f_{-1}(0, x) + \lambda'_3 f_0(0, x))}$$

[0172]

$$P(1, x) = \frac{\exp(\lambda'_1 f_1(1, x) + \lambda'_1 f_{-1}(1, x) + \lambda'_1 f_0(1, x))}{\exp(\lambda'_1 f_1(1, x) + \lambda'_1 f_{-1}(1, x) + \lambda'_1 f_0(1, x)) + \exp(\lambda'_2 f_1(-1, x) + \lambda'_2 f_{-1}(-1, x) + \lambda'_2 f_0(-1, x)) + \exp(\lambda'_3 f_1(0, x) + \lambda'_3 f_{-1}(0, x) + \lambda'_3 f_0(0, x))}$$

[0173] 即  $x$  的待分类正预测条件概率。

[0174] 在  $a$  为 -1 时, 依据公式  $P(a, b) = \frac{\exp(\sum_{i=1}^k \lambda_i f_i(a, b))}{\sum_a \exp(\sum_{i=1}^k \lambda_i f_i(a, b))}$ , 可以得到

[0175]

$$P(-1, x) = \frac{\exp(\sum_{i=1}^3 \lambda_i f_i(-1, x))}{\sum_a \exp(\sum_{i=1}^3 \lambda_i f_i(-1, x))} =$$

[0176]

$$P(-1, x) = \frac{\exp(\lambda'_1 f_1(-1, x) + \lambda'_1 f_{-1}(-1, x) + \lambda'_1 f_0(-1, x))}{\exp(\lambda'_1 f_1(-1, x) + \lambda'_1 f_{-1}(-1, x) + \lambda'_1 f_0(-1, x)) + \exp(\lambda'_2 f_1(-1, x) + \lambda'_2 f_{-1}(-1, x) + \lambda'_2 f_0(-1, x)) + \exp(\lambda'_3 f_1(0, x) + \lambda'_3 f_{-1}(0, x) + \lambda'_3 f_0(0, x))}$$

[0177]

$$P(-1, x) = \frac{\exp(\lambda_1 f_1(-1, x) + \lambda_1' f_{-1}(-1, x) + \lambda_1' f_0(-1, x))}{\exp(\lambda_1 f_1(1, x) + \lambda_1' f_{-1}(1, x) + \lambda_1' f_0(1, x)) + \exp(\lambda_2 f_1(-1, x) + \lambda_2' f_{-1}(-1, x) + \lambda_2' f_0(-1, x)) + \exp(\lambda_3 f_1(0, x) + \lambda_3' f_{-1}(0, x))}$$

[0178] 即x的待分类负预测条件概率。

[0179] 在a为0时,依据公式  $P(a, b) = \frac{\exp(\sum_{i=1}^k \lambda_i f_i(a, b))}{\sum_a \exp(\sum_{i=1}^k \lambda_i f_i(a, b))}$ ,可以得到

[0180]

$$P(0, x) = \frac{\exp(\sum_{i=1}^3 \lambda_i f_i(0, x))}{\sum_a \exp(\sum_{i=1}^3 \lambda_i f_i(0, x))} =$$

[0181]

$$P(0, x) = \frac{\exp(\lambda_1 f_1(0, x) + \lambda_1' f_{-1}(0, x) + \lambda_1' f_0(0, x))}{\exp(\lambda_1 f_1(1, x) + \lambda_1' f_{-1}(1, x) + \lambda_1' f_0(1, x)) + \exp(\lambda_2 f_1(-1, x) + \lambda_2' f_{-1}(-1, x) + \lambda_2' f_0(-1, x)) + \exp(\lambda_3 f_1(0, x) + \lambda_3' f_{-1}(0, x))}$$

[0182]

$$P(0, x) = \frac{\exp(\lambda_1 f_1(0, x) + \lambda_1' f_{-1}(0, x) + \lambda_1' f_0(0, x))}{\exp(\lambda_1 f_1(1, x) + \lambda_1' f_{-1}(1, x) + \lambda_1' f_0(1, x)) + \exp(\lambda_2 f_1(-1, x) + \lambda_2' f_{-1}(-1, x) + \lambda_2' f_0(-1, x)) + \exp(\lambda_3 f_1(0, x) + \lambda_3' f_{-1}(0, x))}$$

[0183] 即x的待分类客观预测条件概率。

[0184] 每个待分类特征值各自对应的待分类正预测条件概率、待分类负预测条件概率和待分类客观预测条件概率的计算过程如本实施例中上述以x为例的计算过程,在此不再赘述。

[0185] 步骤S55:将各个待分类特征值对应的待分类正预测条件概率进行乘运算,得到句子级正预测条件概率,将各个待分类特征值对应的待分类负预测条件概率进行乘运算,得到句子级负预测条件概率,将各个待分类特征值对应的待分类客观预测条件概率进行乘运算,得到句子级客观预测条件概率。

[0186] 步骤S56:比较从所述句子级正预测条件概率、所述句子级负预测条件概率和所述句子级客观预测条件概率的大小。

[0187] 在比较结果为句子级正预测条件概率最大的情况下,确定所述待分类句子级文本的情感类别为正;在比较结果为句子级负预测条件概率最大的情况下,确定所述待分类句子级文本的情感类别为负;在比较结果为句子级客观预测条件概率最大的情况下,确定所述待分类句子级文本的情感类别为客观。

[0188] 再一个实施例

[0189] 在本实施例中,在图5示出的利用预设句子级情感分类单元对已标注情感类别的篇章级文本中的句子级文本进行分类的过程的基础上扩展出另一种利用预设句子级情感分类单元对已标注情感类别的篇章级文本中的句子级文本进行分类的过程,请参见图6,图6示出的是本申请提供的一种句子级情感分类方法的再一种流程图,可以包括以下步骤:

[0190] 步骤S61:确定所述已标注情感类别的篇章级文本中的句子级文本为待分类句子级文本。

[0191] 步骤S62:确定所述待分类句子级文本所包含的词语为待分类特征值。

[0192] 步骤S63:预设所述待分类句子级文本的情感类别分别为正、负和客观。

[0193] 步骤S64:依据公式 $P(a,b) = \frac{\exp(\sum_{i=1}^k \lambda_i f_i(a,b))}{\sum_a \exp(\sum_{i=1}^k \lambda_i f_i(a,b))}$ ,分别计算每个待分类特征值在a分别为+1、-1和0时,对应的待分类正预测条件概率、待分类负预测条件概率和待分类客观预测条件概率。

[0194] 步骤S65:将各个待分类特征值对应的待分类正预测条件概率进行乘运算,得到句子级正预测条件概率,将各个待分类特征值对应的待分类负预测条件概率进行乘运算,得到句子级负预测条件概率,将各个待分类特征值对应的待分类客观预测条件概率进行乘运算,得到句子级客观预测条件概率。

[0195] 步骤S66:比较从所述句子级正预测条件概率、所述句子级负预测条件概率和所述句子级客观预测条件概率的大小。

[0196] 在比较结果为句子级正预测条件概率最大的情况下,确定所述待分类句子级文本的情感类别为正;在比较结果为句子级负预测条件概率最大的情况下,确定所述待分类句子级文本的情感类别为负;在比较结果为句子级客观预测条件概率最大的情况下,确定所述待分类句子级文本的情感类别为客观。

[0197] 步骤S61,步骤S62,步骤S63,步骤S64,步骤S65和步骤S66与图5示出的利用预设句子级情感分类单元对已标注情感类别的篇章级文本中的句子级文本进行分类的过程中的步骤S51,步骤S52,步骤S53,步骤S54,步骤S55和步骤S56相同,在此不再赘述。

[0198] 步骤S67:在确定所述待分类句子级文本的情感类别之后,判断所述待分类句子级文本的情感类别对应的句子级预测条件概率与0.5之差的绝对值是否小于0.05。

[0199] 在判断结果为待分类句子级文本的情感类别对应的句子级预测条件概率与0.5之差的绝对值小于0.05时,执行步骤S68,在判断结果为待分类句子级文本的情感类别对应的句子级预测条件概率与0.5之差的绝对值不小于0.05,执行步骤S69。

[0200] 步骤S68:将所述待分类句子级文本的情感类别转换为客观。

[0201] 步骤S69:所述待分类句子级文本的情感类别保持不变。

[0202] 与上述方法实施例相对应,本申请提供了一种句子级情感分类装置的一种结构示意图,请参见图7,句子级情感分类装置包括:第一获取单元71和预设句子级情感分类单元72。

[0203] 第一获取单元71,用于获取预设句子级情感分类单元。

[0204] 其中,第一获取单元71包括:第二获取单元、第一分类单元和训练单元。

[0205] 第二获取单元,用于分别获取标记为正、负和客观的篇章级文本。

[0206] 第一分类单元,用于对拆分所述篇章级文本获取到的句子级文本按照正、负和客观类型进行分类标记,得到对应的正、负和客观训练样本。

[0207] 训练单元,用于利用所述正、负和客观训练样本对最大熵分类器进行训练,得到预设句子级情感分类单元。

[0208] 预设句子级情感分类单元72,用于对已标注情感类别的篇章级文本中的句子级文

本进行分类。

[0209] 另一个实施例

[0210] 本实施例中,示出的是第一分类单元的具体构成,请参见图8,图8示出的是第一分类单元的一种结构示意图。第一分类单元包括:第一确定单元81、第二确定单元82、第一计算单元83、第二计算单元84、第三计算单元85、第一比较单元86、第三确定单元87、第四确定单元88、第五确定单元89和第六确定单元810。

[0211] 第一确定单元81,用于确定所述标记为正的篇章级文本、所述标记为负的篇章级文本、所述标记为正的篇章级文本中的句子文本和所述标记为负的篇章级文本中的句子级文本为二部图的文档向量。

[0212] 第二确定单元82,用于确定所述正篇章级文本中的词语和所述负篇章级文本中的词语为二部图的词向量。

[0213] 第一计算单元83,用于计算任意一个文档向量到该文档向量所包含的词向量的词转移概率。

[0214] 第二计算单元84,用于依据所述词转移概率,计算任意一个文档向量到任意一个文档向量的文档转移概率。

[0215] 第三计算单元85,用于依据每个文档转移概率和二部图的标签传播算法,计算每个句子级文本对应的正句子级概率和每个句子级文本对应的负句子级概率。

[0216] 第一比较单元86,用于比较所述正句子级概率和负句子级概率的大小,在比较结果为所述正句子级概率大于所述负句子级概率的情况下,执行第三确定单元87,在比较结果为所述负句子级概率大于所述正句子级概率的情况下,执行第四确定单元88。

[0217] 第三确定单元87,用于确定该句子级文本的类别为正。

[0218] 第四确定单元88,用于确定该句子级文本的类别为负。

[0219] 第五确定单元89,用于确定类别为正的句子级文本为正训练样本,确定类别为负的句子级文本为负训练样本。

[0220] 第六确定单元810,用于对所述标记为客观的篇章级文本中的句子级文本进行人工标注,确定类别为客观的句子级文本,并确定类别为客观的句子级文本为客观训练样本。

[0221] 再一个实施例

[0222] 在本实施例中,示出的是训练单元的具体结构,请参见图9,图9示出的是本申请提供的一种训练单元的一种结构示意图,训练单元包括:

[0223] 第七确定单元91、第四计算单元92、第一调整单元93、第二调整单元94、第三调整单元95和第八确定单元96。

[0224] 第七确定单元91,用于确定所述正、负和客观训练样本中包含的词语为特征值。

[0225] 第四计算单元92,用于依据公式  $P(a,b) = \frac{\exp(\sum_{i=1}^k \lambda_i f_i(a,b))}{\sum_a \exp(\sum_{i=1}^k \lambda_i f_i(a,b))}$ , 分别计算每个特征值在a分别为+1、-1和0时,对应的正预测条件概率、负预测条件概率和客观预测条件概率,其中,所述b为特征值,P()为预测条件概率,exp()为自然数e为底的指数函数,f<sub>i</sub>()为二值

特征函数,  $\lambda_i$  为特征函数值  $f_i(a, b)$  的权值且相同  $b$  对应的不同特征函数值的权值相同,  $\sum_{i=1}^k$

为对每个特征值对应的  $k$  个特征函数值进行求和的函数,  $K$  为不小于 1 的整数,  $\sum_a$  为对  $a$  为不同值时对应的数据进行求和的函数。

[0226] 第一调整单元 93, 用于利用 GIS 算法, 调整每个特征值对应的正预测条件概率, 直至每个特征值各自的正预测条件概率收敛, 并将每个特征值各自收敛的正预测条件概率对应的  $\lambda$  作为每个特征值各自的测试正  $\lambda$ 。

[0227] 第二调整单元 94, 用于利用 GIS 算法, 调整每个特征值对应的负预测条件概率, 直至每个特征值各自的负预测条件概率收敛, 并将每个特征值各自收敛的负预测条件概率对应的  $\lambda$  作为每个特征值各自的测试负  $\lambda$ 。

[0228] 第三调整单元 95, 用于利用 GIS 算法, 调整每个特征值对应的客观预测条件概率, 直至每个特征值各自的客观预测条件概率收敛, 并将每个特征值各自收敛的客观预测条件概率对应的  $\lambda$  作为每个特征值各自的测试客观  $\lambda$ 。

[0229] 第八确定单元 96, 用于确定所述测试正  $\lambda$ 、测试负  $\lambda$  和测试客观  $\lambda$  对应的最大熵分类器为预设句子级情感分类单元。

[0230] 再一个实施例

[0231] 在本实施例中, 示出的是预设句子级情感分类单元的具体结构, 请参见图 10, 图 10 示出的是本申请提供的一种预设句子级情感分类单元的一种结构示意图, 预设句子级情感分类单元包括:

[0232] 第九确定单元 101、第十确定单元 102、预设单元 103、第五计算单元 104、第六计算单元 105、第二比较单元 106、第十一确定单元 107、第十二确定单元 108 和第十三确定单元 109。

[0233] 第九确定单元 101, 用于确定所述已标注情感类别的篇章级文本中的句子级文本为待分类句子级文本。

[0234] 第十确定单元 102, 用于确定所述待分类句子级文本所包含的词语为待分类特征值。

[0235] 预设单元 103, 用于预设所述待分类句子级文本的情感类别分别为正、负和客观。

[0236] 第五计算单元 104, 用于依据公式  $P(a, b) = \frac{\exp(\sum_{i=1}^k \lambda_i f_i(a, b))}{\sum_a \exp(\sum_{i=1}^k \lambda_i f_i(a, b))}$ , 分别计算每个待分类特征值在  $a$  分别为 +1、-1 和 0 时, 对应的待分类正预测条件概率、待分类负预测条件概率和待分类客观预测条件概率, 其中, 所述  $b$  为待分类特征值,  $P()$  为待分类预测条件概率,  $\sum_{i=1}^k$

为对每个待分类特征值对应的  $k$  个特征函数值进行求和的函数,  $K$  为 3。

[0237] 其中, 每个待分类特征值对应的各个特征函数值分别对应所述待分类句子级文本的预设情感类型正、负和客观, 在计算待分类正预测条件概率时, 若待分类特征值对应的词语包含在所述特征值中, 则  $\lambda$  为对应测试正  $\lambda$ , 否则  $\lambda$  为 0, 在计算待分类负预测条件概率时,

若待分类特征值对应的词语包含在所述特征值中,则 $\lambda$ 为对应测试负 $\lambda$ ,否则 $\lambda$ 为0,在计算待分类客观预测条件概率时,若待分类特征值对应的词语包含在所述特征值中,则 $\lambda$ 为测试客观 $\lambda$ ,否则 $\lambda$ 为0。

[0238] 第六计算单元105,用于将各个待分类特征值对应的待分类正预测条件概率进行乘运算,得到句子级正预测条件概率,将各个待分类特征值对应的待分类负预测条件概率进行乘运算,得到句子级负预测条件概率,将各个待分类特征值对应的待分类客观预测条件概率进行乘运算,得到句子级客观预测条件概率。

[0239] 第二比较单元106,用于比较从所述句子级正预测条件概率、所述句子级负预测条件概率和所述句子级客观预测条件概率的大小,在比较结果为所述句子级正预测条件概率最大的情况下,执行第十一确定单元107,在在比较结果为所述句子级负预测条件概率最大的情况下,执行第十二确定单元108,在比较结果为所述句子级客观预测条件概率最大的情况下,执行第十三确定单元109。

[0240] 第十一确定单元107,用于确定所述待分类句子级文本的情感类别为正。

[0241] 第十二确定单元108,用于确定所述待分类句子级文本的情感类别为负。

[0242] 第十三确定单元109,用于确定所述待分类句子级文本的情感类别为客观。

[0243] 在上述装置实施例中,句子级情感分类装置还可以包括:判断单元和转换单元。

[0244] 判断单元,用于判断所述待分类句子级文本的情感类别对应的句子级预测条件概率与0.5之差的绝对值是否小于0.05,若是,执行转换单元,若否,所述待分类句子级文本的情感类别保持不变。

[0245] 转换单元,用于将所述待分类句子级文本的情感类别转换为客观。

[0246] 需要说明的是,本说明书中的各个实施例均采用递进的方式描述,每个实施例重点说明的都是与其他实施例的不同之处,各个实施例之间相同相似的部分互相参见即可。对于装置类实施例而言,由于其与方法实施例基本相似,所以描述的比较简单,相关之处参见方法实施例的部分说明即可。

[0247] 最后,还需要说明的是,在本文中,诸如第一和第二等之类的关系术语仅仅用来将一个实体或者操作与另一个实体或操作区分开来,而不一定要求或者暗示这些实体或操作之间存在任何这种实际的关系或者顺序。而且,术语“包括”、“包含”或者其任何其他变体意在涵盖非排他性的包含,从而使得包括一系列要素的过程、方法、物品或者设备不仅包括那些要素,而且还包括没有明确列出的其他要素,或者是还包括为这种过程、方法、物品或者设备所固有的要素。在没有更多限制的情况下,由语句“包括一个……”限定的要素,并不排除在包括所述要素的过程、方法、物品或者设备中还存在另外的相同要素。

[0248] 以上对本申请所提供的一种句子级情感分类方法及装置进行了详细介绍,本文中应用了具体个例对本申请的原理及实施方式进行了阐述,以上实施例的说明只是用于帮助理解本申请的方法及其核心思想;同时,对于本领域的一般技术人员,依据本申请的思想,在具体实施方式及应用范围上均会有改变之处,综上所述,本说明书内容不应理解为对本申请的限制。

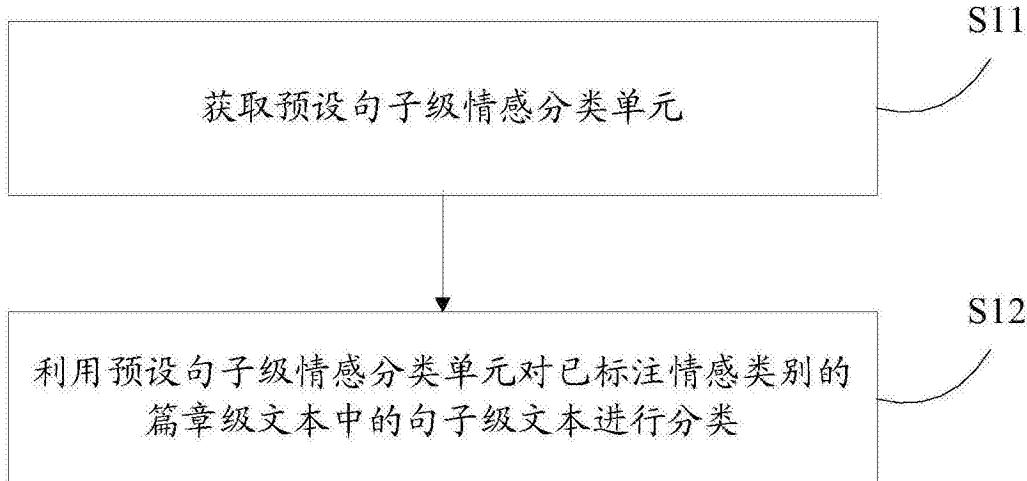


图1

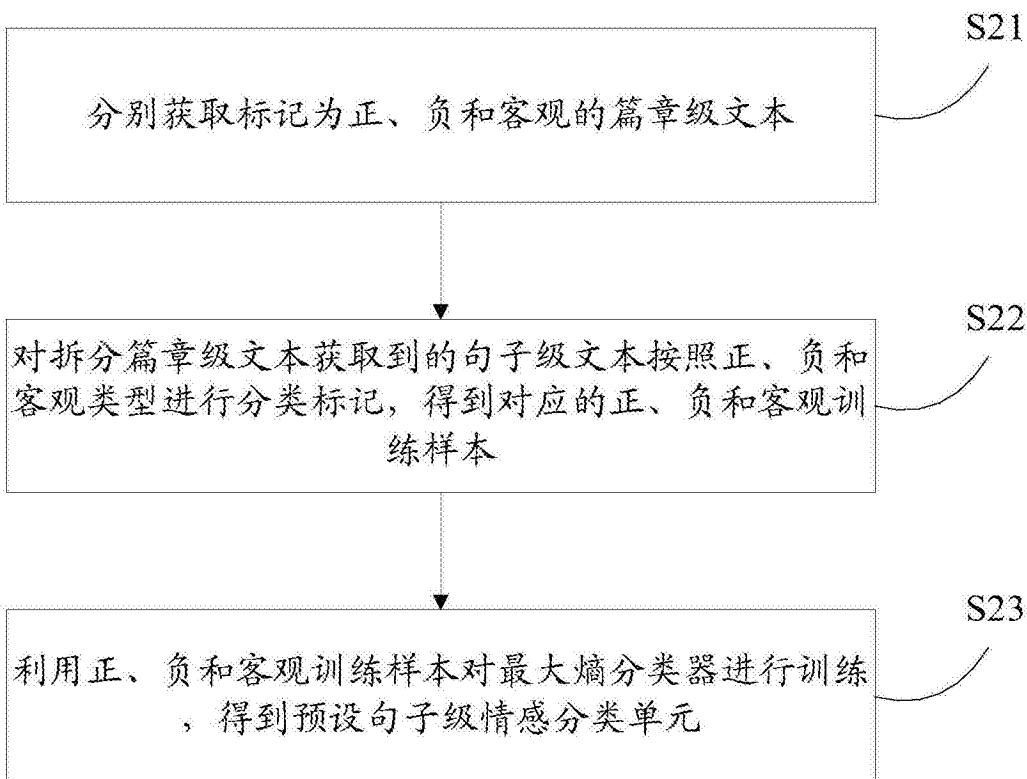


图2

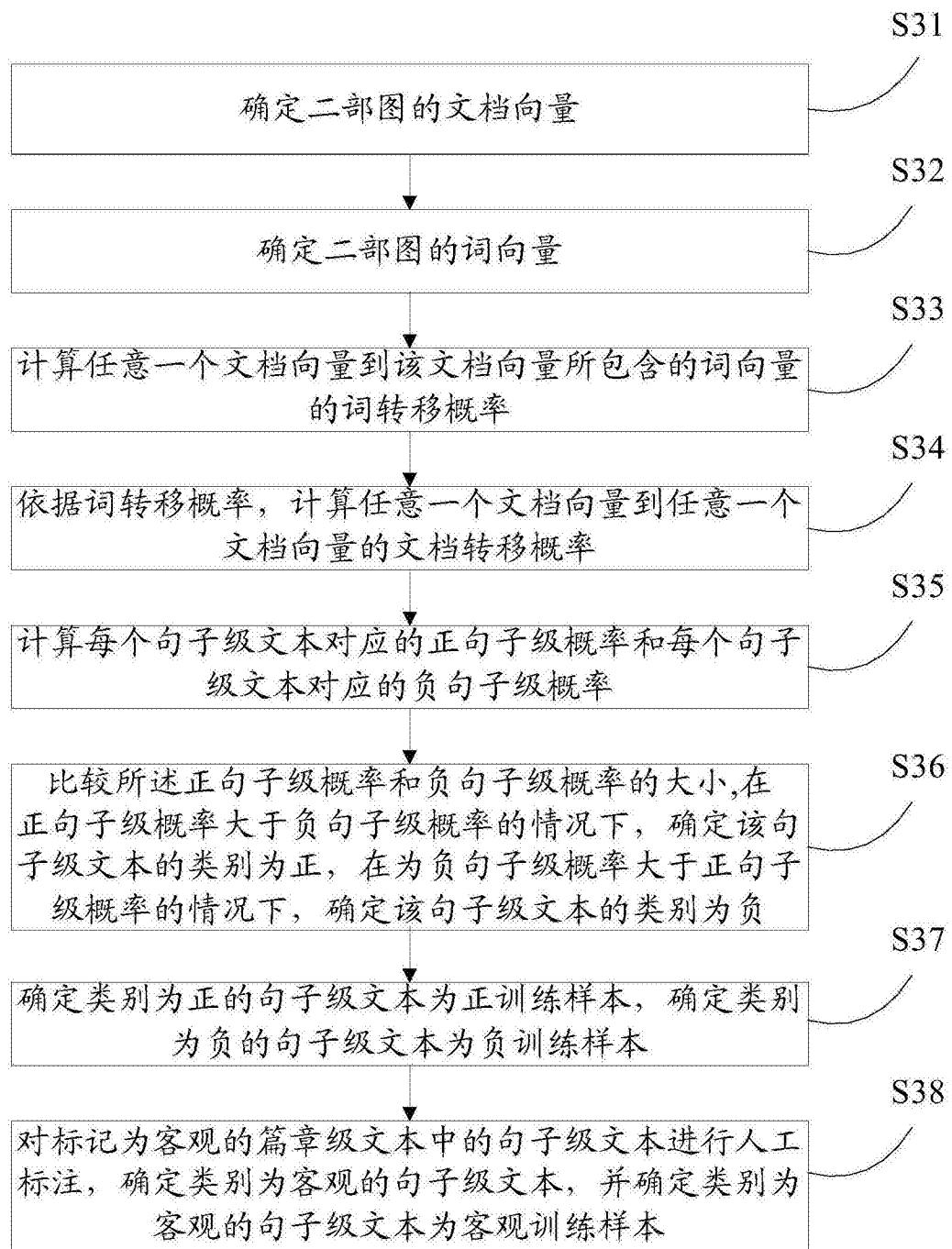


图3

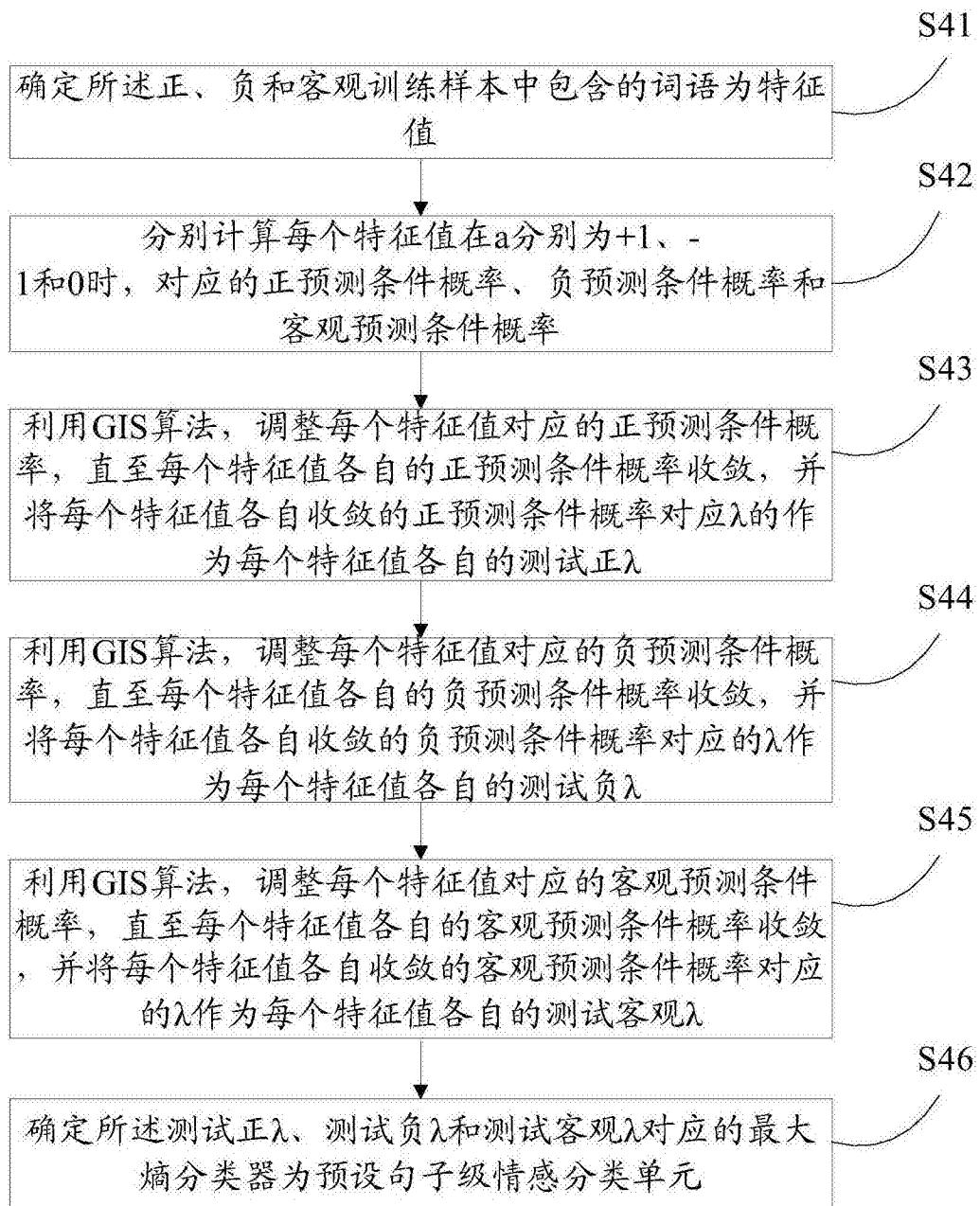


图4

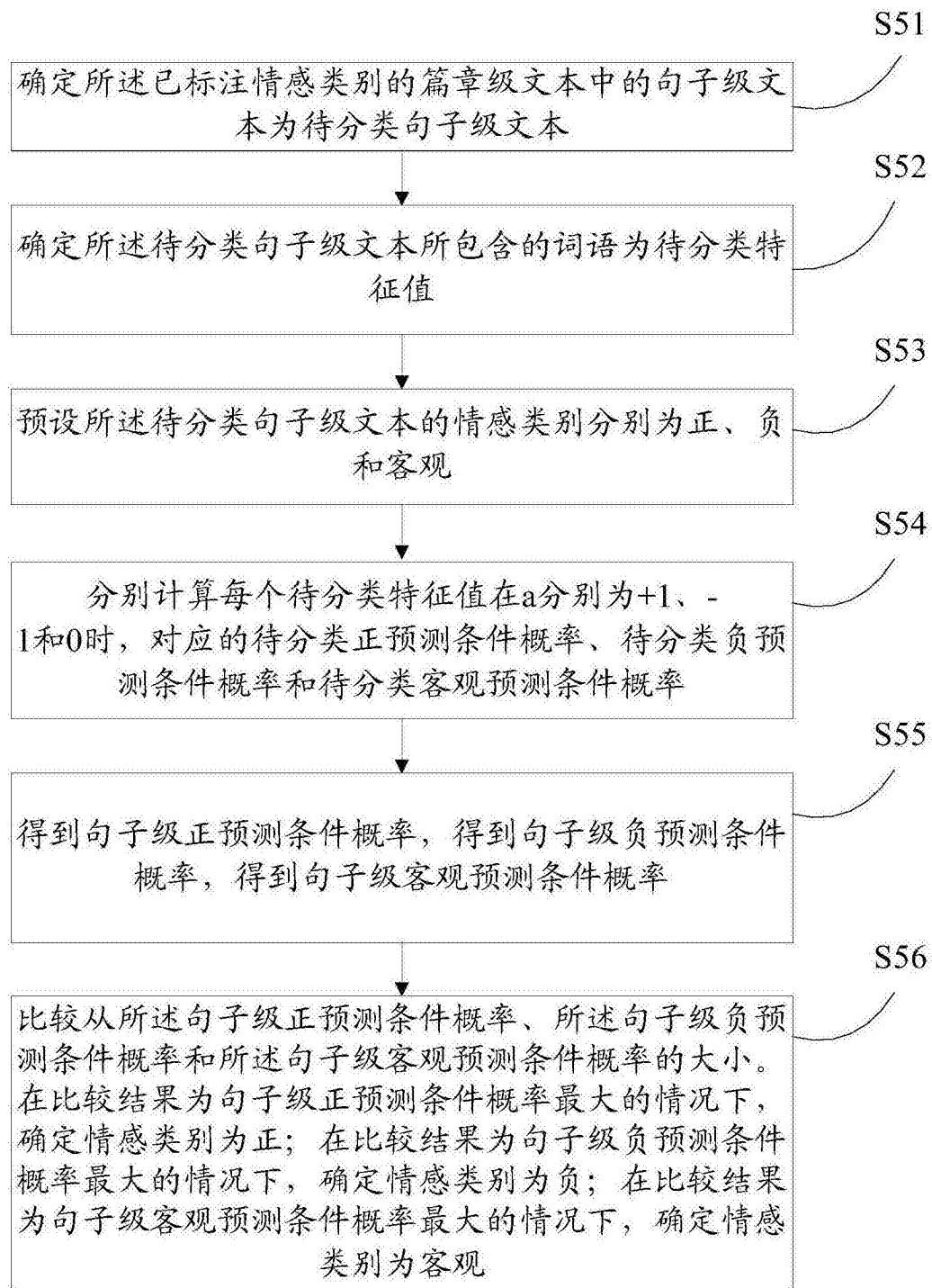


图5

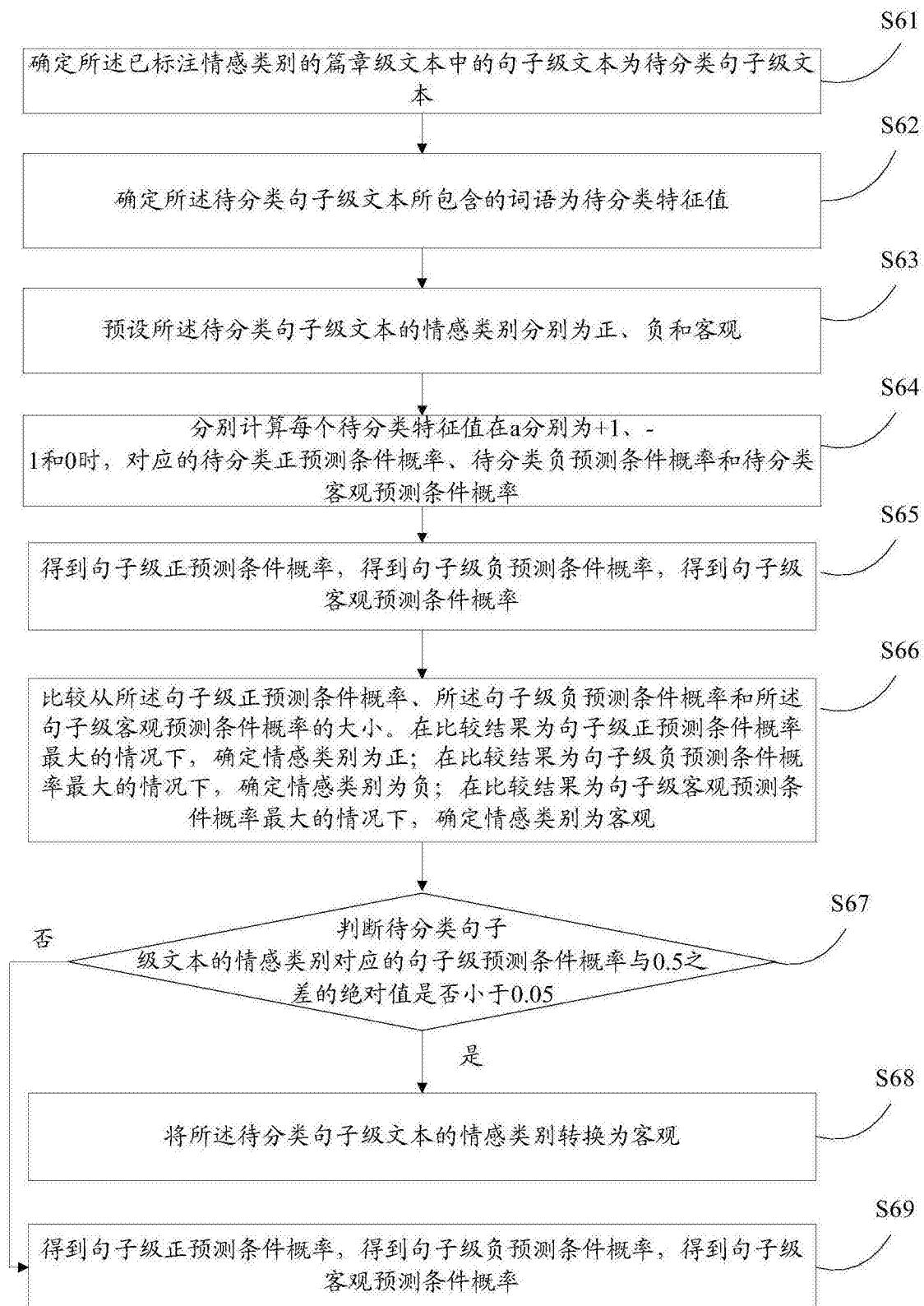


图6

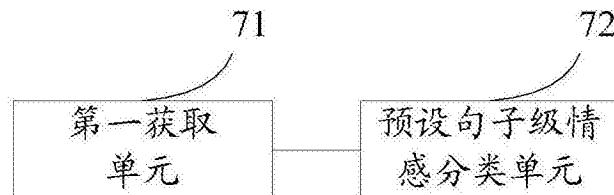


图7

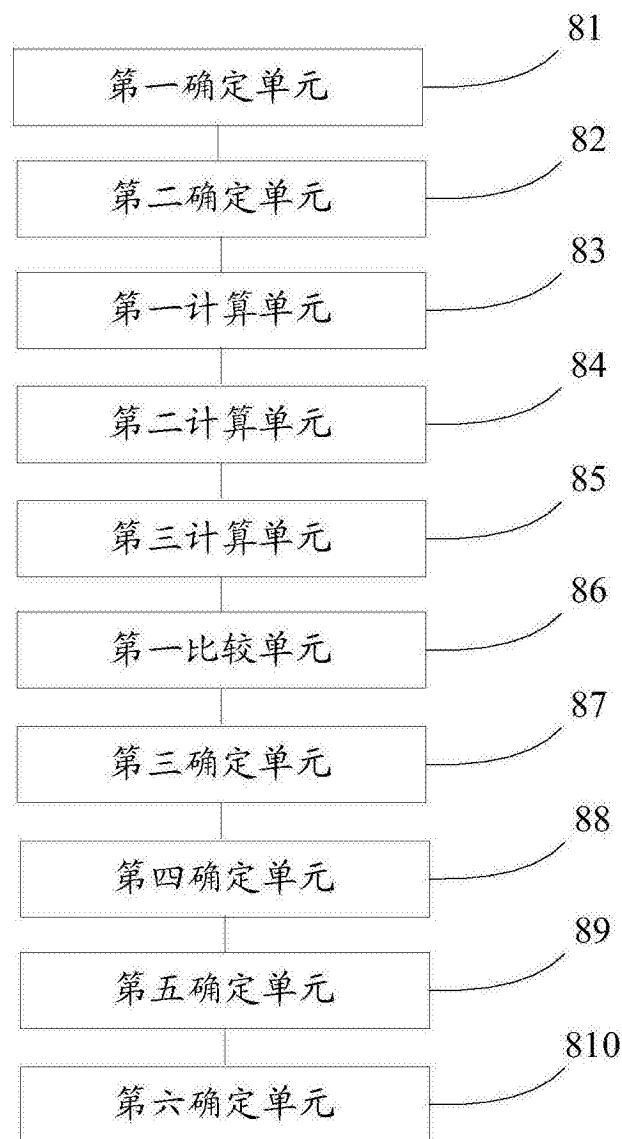


图8

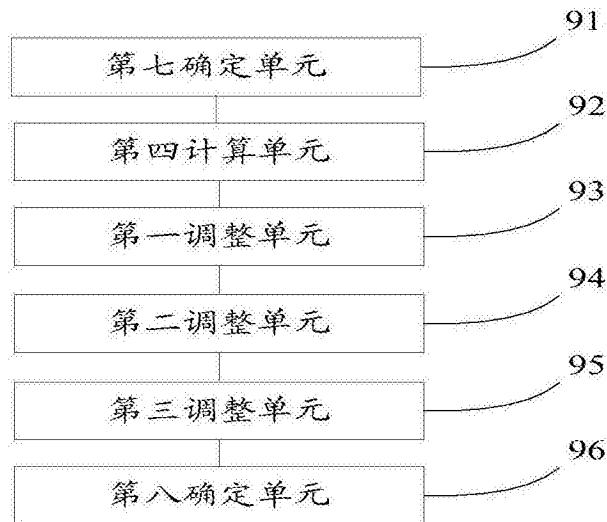


图9

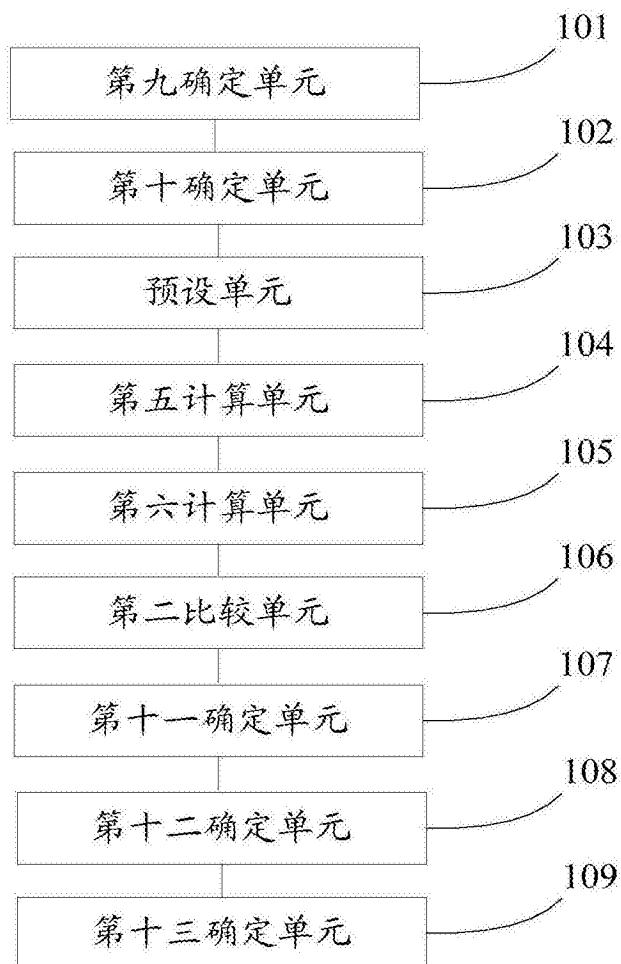


图10