(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 106156004 B
(45)授权公告日 2019.03.26

(21)申请号 201610519169.7

(22)申请日 2016.07.04

(65)同一申请的已公布的文献号
申请公布号 CN 106156004 A

(43)申请公布日 2016.11.23

(73)专利权人 中国传媒大学
地址 100024 北京市朝阳区定福庄南里7号中国传媒大学

(72)发明人 殷复莲 潘幸艺 刘晓薇 王颜颜

(74)专利代理机构 北京鸿元知识产权代理有限公司 11327
代理人 许向彤 张宁

(51)Int.Cl.
G06F 17/27(2006.01)
G06F 16/35(2019.01)

(56)对比文件
CN 104573046 A,2015.04.29,
CN 105550269 A,2016.05.04,
CN 105701229 A,2016.06.22,
JP 2008176489 A,2008.07.31,
CN 105205124 A,2015.12.30,

审查员 王玮玮

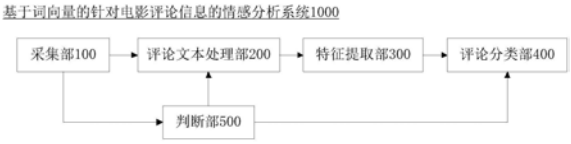
权利要求书6页 说明书12页 附图5页

(54)发明名称

基于词向量的针对电影评论信息的情感分析系统及方法

(57)摘要

本发明提供一种基于词向量的针对电影评论信息的情感分析系统,包括:采集部,采集电影评论,形成评论文本库;评论文本处理部,对评论文本库中的每一条评论进行分词,构建分词后的评论文本库;特征提取部,对分词后的评论文本库中的每一条评论转换为基于词向量的评论向量,完成每一条评论的特征提取,其中,所述词向量为所述评论中每一个词语的词语概率最大化的最优解,所述评论向量为每一条评论中的所有词向量的平均值;评论分类部,存储有分类模型,将所述评论向量输入到所述分类模型中进行训练,得到每一条评论的评论类型。上述情感分析系统不需人工标注,不依赖于情感词典的维修修缮工作。



CN 106156004 B

1.一种基于词向量的针对电影评论信息的情感分析系统，其特征在于，包括：

采集部，采集电影评论，形成评论文本库；

评论文本处理部，对评论文本库中的每一条评论进行分词，构建分词后的评论文本库；

特征提取部，对分词后的评论文本库中的每一条评论转换为基于词向量的评论向量，完成每一条评论的特征提取，其中，所述词向量为所述评论中每一个词语的词语概率最大化的最优解，所述评论向量为每一条评论中的所有词向量的平均值；

评论分类部，存储有分类模型，将所述评论向量输入到所述分类模型中进行训练，得到每一条评论的评论类型，

其中，所述特征提取部包括：

第一设定单元，设定词向量训练窗口的大小、词向量的维度和变化阈值；

映射单元，将分词后的评论文本库中的所有评论中的词去重复后形成词汇表，建立分词后的评论文本库的词语到词汇表中词语的映射；

词向量查找表构建单元，将上述词汇表中的每一个词语的词向量的每一维的数值设定为变量，构建词向量查找表；

第一更新单元，随机生成所述词向量查找表构建单元中各词向量在各维度的数值，设定词向量训练窗口内一个词语为所述词语所在评论的中心词，通过所述中心词的词向量预测所述评论中其他词语的预测概率，通过所述预测概率采用平均对数法和迭代方法不断更新所述其他词语的词向量在每一维度的数值，直到所述数值的变化值小于变化阈值，完成词向量查找表的更新，其中，所述数值的变化值为：

$$B^{a}_{word_{n,I+J},k}=v^{a}_{word_{n,I+J},k}-v^{a-1}_{word_{n,I+J},k}=[\frac{1}{wordc(doc_n)}-O^{a-1}(v_{word_{n,I+J}})]\times v^{a-1}{}_{word_{n,I},k}$$

$$O^{a-1}(word_{n,I+J})=\frac{1}{wordc(doc_n)}\sum_{I=1}^{wordc(doc_n)}\sum_{-win/2\le J\le win/2,J\neq 0}\log p^{a-1}(word_{n,I+J}\mid word_{n,I})$$

$$p^{a-1}(word_{n,I+J}\mid word_{n,I})=\frac{\exp[(v^{a-1}_{word_{n,I+J}})^{T}\bullet v^{a-1}{}_{word_{n,I}}]}{\sum_{X=1}^{m}\exp[(v^{a-1}_{word_X})^{T}\bullet v^{a-1}{}_{word_{n,I}}]}$$

其中，a为迭代次数，为自然数；

$wordc(doc_n)$为第n条评论的词语总数；

m为词汇表中的词语总数；

win为词向量训练窗口的大小；

$word_{n,I}$为第n条评论的第I个词语；

$v^{a-1}{}_{word_{n,I},k}$为第a-1次迭代中，第n条评论中第I个词语的第k维的数值；

$v^{a-1}_{word_{n,I}}$为第a-1次迭代中，第n条评论的第I个词语的词向量；

$v^{a-1}_{word_X}$为第a-1次迭代中，词汇表的第X个词语的词向量；

$p^{a-1}(word_{n,I+J}\mid word_{n,I})$为第a-1次迭代中，通过中心词$word_{n,I}$词向量预测得到词语

$word_{n,I+J}$词向量的预测概率；

$O^{a-1}(word_{n,I+J})$为第a-1次迭代中，第n条评论的除中心词外各词语的预测概率的对数平均值；

$B^{a}_{word_{n,I+J},k}$为词语$word_{n,I+J}$第k维数值在第a-1次迭代和第a次迭代的数值变化；

评论向量构建单元，通过计算每一条评论中的所有词向量的平均值，将所述评论的信息替换为评论向量。

2.根据权利要求1所述的情感分析系统，其特征在于，还包括：

判断部，判断所述采集部采集的评论中是否具有评分，将不具有评分的评论和具有评分的评论分类存储，将所述评分存储到评分数据库。

3.根据权利要求2所述的情感分析系统，其特征在于，还包括：分类模型构建部，用于构建分类模型，包括：

评分训练模型构建单元，构建评分训练模型，其中，所述评分训练模型为设定评分标准，高于标准的评论的评分值设为1，不高于所述标准的评论的评分值设为-1，将每一条评论的评分相对于所述评分标准存储成只包括1和-1的数据集，其中，1表示该条评论为正倾向，-1表示该条评论为负倾向；

分类模型构建单元，构建包括变量的分类模型，其中，所述分类模型为：

$y^{i}=w\cdot rv^{(i)}-b$

$$w=\sum_{n=1}^{G}\alpha_n y^{(n)} rv'(n)$$

$$\arg\min_{w,b}\frac{\|w\|}{2},\text{且}y^{(n)}(w\cdot rv'^{(n)}-b)\geq 1$$

$$E_n=\sum_{k=1}^{G}\alpha_k y^{(k)}\left\langle rv'(k),rv'(n)\right\rangle-y^{(n)}$$

$$\left\langle rv'(k),rv'(n)\right\rangle=\sum_{s=1}^{\dim} rv'_s(k)\cdot rv'_s(n)$$

其中，$rv^{(i)}$为不具有评分的评论向量；

G为具有评分的评论向量的总个数；

w和b为变量，其中，w为垂直于评论向量平面的向量，b为阈值；

α为拉格朗日参数，是D维向量，$\alpha\in R^{D}$

α_k是拉格朗日参数α第k维度的分量；

$y^{(k)}$，$y^{(n)}$是具有评分的第k个和第n个评论向量在数据集中的数值；

$rv'^{(k)}$，$rv'^{(n)}$分别是具有评分的第k个和第n个评论向量；

$rv'_s(k)$，$rv'_s(k)$分别表示第k个和第n个评论向量的第s维分量，$1\leq s\leq dim$；

E_n是被优化的目标函数；

$\langle rv'^{(k)},rv'^{(n)}\rangle$表示对评论向量$rv'^{(k)}$，$rv'^{(n)}$求向量内积；

dim是词向量的维度数；

第一获得单元，通过评论文本处理部和特征提取部对具有评分的评论进行处理，获得所述评论对应的评论向量，将存储所述评论的评分的评分数据库通过评分训练模型转变为只包括1和-1的数据集；

第二获得单元，利用第一获得单元获得的评论向量及其对应的数据集确定所述分类模型的变量。

4.根据权利要求3所述的情感分析系统，其特征在于，所述第二获得单元包括：

第二设定单元，初始化拉格朗日参数α和阈值b及b的待选参数b_1和b_2，设置指定精度ε、容差tol和调和函数C；

计算单元，计算第一获取单元中每一个评论向量对应的E函数值；

第二判断单元，判断上述评论向量的评分相对评分标准的数据与其E函数值的乘积以及其拉格朗日参数是否满足下述条件：$y^{(n)}E_n<-tol$且$\alpha_n<C$，或者$y^{(n)}E_n>tol$且$\alpha_n>0$，如果存在均不满足上述两个条件的评论向量，则发送指令给计算单元，重新计算该评论向量的E值；如果满足上述两个条件之一，发送指令给第二更新单元；

第二更新单元，将满足第二判断单元条件的第一获取单元中的任意两个评论向量配对，更新每一个评论向量的拉格朗日参数，其中，

$$\alpha_n^{(new)}=\begin{cases} H & \alpha_n^{(new,wnc)}>H \\ \alpha_n^{(new,wnc)} & L\le\alpha_n^{(new,wnc)}\le H \\ L & \alpha_n^{(new,wnc)}<L \end{cases}$$

$$\alpha_n^{(new,wnc)}=\alpha_n^{(old)}-\frac{y^{(n)}(E_k-E_n)}{\eta}$$

$\eta=2\langle rv'^{(k)},rv'^{(n)}\rangle-\langle rv'^{(k)},rv'^{(k)}\rangle-\langle rv'^{(n)},rv'^{(k)}\rangle$，且$\eta<0$

$$\begin{cases} L=max(0,\alpha_n^{(old)}-\alpha_k^{(old)}),H=min(C,C+\alpha_n^{(old)}-\alpha_k^{(old)}) & y^{(n)}\neq y^{(k)} \\ L=max(0,\alpha_n^{(old)}+\alpha_k^{(old)}-C),H=min(C,C+\alpha_n^{(old)}-\alpha_k^{(old)}) & y^{(n)}=y^{(k)} \end{cases},\text{且}L\neq H$$

$$\alpha_k^{(new)}=\alpha_k^{(old)}+y^{(k)}y^{(n)}(\alpha_n^{(old)}-\alpha_n^{(new)}),\text{且}|\alpha_n^{(new)}-\alpha_n^{(old)}|\ge\varepsilon$$

其中，$rv'^{(n)}$和$rv'^{(k)}$为满足第二判断单元条件的第一获取单元中的任意两个评论向量；

$\alpha_n^{(old)}$和$\alpha_k^{(old)}$为更新前评论向量$rv'^{(n)}$和$rv'^{(k)}$对应的拉格朗日参数；

$\alpha_n^{(new,wnc)}$为更新过程中评论向量$rv'^{(n)}$待判断的新的拉格朗日参数；

$\alpha_n^{(new)}$和$\alpha_k^{(new)}$是更新后评论向量$rv'^{(n)}$和$rv'^{(k)}$对应的拉格朗日参数；

L和H为$\alpha_n^{(old)}$更新的上限和下限；

η是被优化的目标函数E_n的二阶导数；

E_k是评论向量$rv'^{(k)}$对应的目标函数；

第三更新单元，更新每一个评论相量对应的阈值，其中，

$$b^{(n)}=\begin{cases} b_1^{(new)} & 0<\alpha_k^{(new)}<C \\ b_2^{(new)} & 0<\alpha_n^{(new)}<C \\ \dfrac{b_1^{(new)}+b_2^{(new)}}{2} & \text{其他情况} \end{cases}$$

$b_1^{(new)}=b_1^{(old)}-E_k-y^{(k)}(\alpha_k^{(new)}-\alpha_k^{(old)})\langle rv'^{(k)},rv'^{(k)}\rangle-y^{(n)}(\alpha_n^{(new)}-\alpha_n^{(old)})\langle rv'^{(k)},rv'^{(n)}\rangle$

$b_2^{(new)}=b_2^{(old)}-E_n-y^{(k)}(\alpha_k^{(new)}-\alpha_k^{(old)})\langle rv'^{(k)},rv'^{(n)}\rangle-y^{(n)}(\alpha_n^{(new)}-\alpha_n^{(old)})\langle rv'^{(n)},rv'^{(n)}\rangle$

其中，$b^{(n)}$为更新后评论相量$rv'^{(n)}$对应的阈值b的值；

$b_1^{(old)}$、$b_2^{(old)}$为之前保留的待选参数b_1和b_2；

第二确定单元，根据更新后各评论向量的拉格朗日参数及其对应的阈值确定变量参数w和b，其中，

$$w=\sum_{n=1}^{G}\alpha_n y^{(n)} rv'^{(n)}$$

$\arg\min_{w,b}\frac{\|w\|}{2}$，且$y^{(n)}(w\cdot rv'^{(n)}-b)\geq 1$。

5.根据权利要求1所述的情感分析系统，其特征在于，所述特征提取部还包括：第一判断单元，判断每一条评论的词语总数是否大于词向量训练窗口的大小，其中，当所述评论的词语总数不大于词向量训练窗口的大小时，选择所述评论中的一个词为中心词，对所述词向量查找表进行更新；当所述评论的词语总数大于词向量训练窗口的大小时，所述评论在所述词向量窗口中从左往右或者从右往左显示，依次选择所述词向量训练窗口中的一个词语为中心词，对所述词向量查找表进行更新。

6.根据权利要求1所述的情感分析系统，其特征在于，所述评论文本处理部包括：

第一分词单元，对每一条电影评论遍历，根据句尾的标点符号以及空格符，将每一条评论分割为一个或多个短句；

第二分词单元，基于Trie树结构对评论文本库进行词图扫描，生成每一条中汉字所有可能成词情况所构成的有向无环图，获得所述有向无环图的多种切割方案；

第一确定单元，采用了动态规划查找所述有向无环图基于词频的最大概率路径，确定切割方案。

7.一种利用权利要求1所述情感分析系统进行情感分析的方法，其特征在于，包括：

采集电影评论，形成评论文本库；

对评论文本库中的每一条评论进行分词，构建分词后的评论文本库；

将分词后的评论文本库中的所有评论中的词去重复后形成词汇表，建立分词后的评论文本库的词语到词汇表中的词语的映射；

设定词向量维度，将上述词汇表中的每一个词的词向量的每一维的数值设定为变量，

构建词向量查找表；

随机生成所述词向量查找表的各词向量在各维度的数值；

设定词向量训练窗口的大小，以所述词向量训练窗口内一个词语为所述词语所在评论的中心词，通过所述中心词的词向量预测所述评论中其他词语的预测概率，通过所述预测概率采用平均对数法和迭代方法不断更新所述其他词语的词向量在每一维度的数值，直到所述数值的变化值小于变化阈值，完成词向量查找表的更新；

将每一条评论中的词向量映射到所述更新后的词向量查找表的数值进行平均计算，从而将每一条评论的文本信息替换为评论向量；

将每一条评论的评论向量代入到分类模型中进行训练，得到每一条评论的评论类型。

8.根据权利要求7所述的情感分析方法，其特征在于，还包括：

判断每一条评论的词语总数是否大于词向量训练窗口的大小；

当所述评论的词语总数不大于词向量训练窗口的大小时，选择所述评论中的一个词为中心词，对所述词向量查找表进行更新；

当所述评论的词语总数大于词向量训练窗口的大小时，所述评论在所述词向量窗口中从左往右或者从右往左显示，依次选择所述词向量训练窗口中的一个词语为中心词，对所述词向量查找表进行更新。

9.根据权利要求7所述的情感分析方法，其特征在于，还包括：

判断所述采集部采集的每一条评论中是否具有评分；

如果所述评论中具有评分，设定评分标准，高于标准的评论的评分值设为1，不高于所述标准的评论的评分值设为-1，得到所述评论的评论类型，其中，1表示该条评论为正倾向，-1表示该条评论为负倾向；

如果所述评论中不具有评分，通过评论文本处理部、特征提取部和评论分类部获得所述评论的评论类型。

10.根据权利要求7所述的情感分析方法，其特征在于，还包括构建分类模型的步骤，所述步骤包括：

构建评分训练模型，其中，所述评分训练模型包括设定评分标准，高于标准的评论的评分设为1，不高于所述标准的评论的评分设为-1，将每一条评论的评分相对于所述评分标准存储成只包括1和-1的数据集，其中，1表示该条评论为正倾向，-1表示该条评论为负倾向；

构建包括变量的分类模型；

通过采集部采集具有评分的评论，形成评论文本库和评分数据库；

将所述评分数据库中各评论的评分在评分训练模型中进行训练，得到各评论的所述数据集；

通过评论文本处理部和特征提取部获得所述评论文本库中各评论的评论向量；

利用上述评论向量及其对应的数据集获得分类模型的变量，完成分类模型的构建。

11.根据权利要求10所述的情感分析方法，其特征在于，所述利用评论向量及其对应的数据集获得分类模型的变量的方法，包括：

初始化具有评分的各评论向量的拉格朗日参数和阈值，设置指定精度和容差；

计算上述评论向量对应的E函数值；

筛选出满足条件的评论向量，其中，所述条件为$y^{(n)}E_n<-tol$且$\alpha_n<C$，或者$y^{(n)}E_n>tol$

且$\alpha_n>0$，$y^{(n)}$是具有评分的第n个评论向量在数据集中的数值，E_n是被优化的目标函数，tol是容差，C是调和函数，α_n是拉格朗日参数α第n维度的分量；

将满足上述条件的评论向量中任意两个评论向量进行配对，更新每一个评论向量的拉格朗日参数；

更新上述每一个评论向量对应的阈值；

根据更新后各评论向量的拉格朗日参数及其对应的阈值确定变量参数。

12.根据权利要求7所述的情感分析方法，其特征在于，所述对评论文本库中的每一条评论进行分词的步骤包括：

根据句尾的标点符号以及空格符，将评论文本库中的每一条评论分割为一个或多个短句；

基于Trie树结构对评论文本库进行词图扫描，得到所述短句中汉字所有可能成词情况，构成有向无环图，得到每一条评论的所述短句的多个分割方案；

记录每一条评论的所述多个分割方案形成的所有词语，以及该词语在所述评论文本库中出现的次数，得到每一个词语出现的频率，其中，所述频率$p(w_n)$为：

$$p(w_n)=\frac{freq(w_n)}{\sum_{i=1}^{t}freq(w_i)}$$

其中，$p(w_n)$是词语w_n出现的频率；

$freq(w_n)$是词语w_n出现的频数；

t为所有有向无环图中所有可能成词情况构成的词语的总数；

基于所述频率，采用查找最大概率路径的方法确定每一条评论的切割方案。

基于词向量的针对电影评论信息的情感分析系统及方法

技术领域

[0001] 本发明涉及数据挖掘技术领域，更为具体地，涉及一种基于词向量的针对电影评论信息的情感分析系统及方法。

背景技术

[0002] 随着互联网的迅速发展，网络上的信息爆炸式增长，海量信息成为人们日常中重要的信息来源。随着使用互联网的在线用户数增长，越来越多的用户倾向于在博客、论坛、微博、在线视频中发表针对电影的观感和评论。如何处理激增的文本、从中获取关键信息，成为当前十分重要的信息处理技术问题。在线影评网站中的影视评论文本，博客、论坛、微博中具有多种讨论视频作品的文章。对电影的评估而言，如何对从电影大众评论中抽取主观性观点，量化计算大众的正面倾向或负面倾向，是自然语言处理在实际问题中的重要应用。

[0003] 传统的自然语言处理方法是基于词语计数的统计模型，以词频为重要的文本特征，这一方法在多项自然语言处理的任务中已有丰富的研究。根据其需求特性，情感分析可采用机器学习中的分类方法实现，包括有监督学习与无监督学习。有监督学习由评论文本及评分组合的训练样本训练得到分类模型，其中采用词袋模型，分类模型的训练方法包括贝叶斯分类、最大熵模型和支持向量机模型等。无监督学习方法是基于情感词典的方法，修建与维护一个大型的情感词典受到成本与规模的限制，在此基础上，已有基于种子词与词语关系自动构建词典的方法。基于传统的情感分析方法，或依赖于修建并维护完善的领域针对性强的情感词典，或依赖于大量的人工文本标注工作，这通常需要消耗大量人工精力。在信息改变迅速的在线电影评论应用中，如何减少人工标注和对情感词典的维护修缮工作，是一个亟待解决的问题。

发明内容

[0004] 鉴于上述问题，本发明的目的是提供一种不需人工标注，不依赖于情感词典的维修修缮工作的基于词向量的针对电影评论信息的情感分析系统及方法。

[0005] 根据本发明的一个方面，提供一种基于词向量的针对电影评论信息的情感分析系统，包括：采集部，采集电影评论，形成评论文本库；评论文本处理部，对评论文本库中的每一条评论进行分词，构建分词后的评论文本库；特征提取部，对分词后的评论文本库中的每一条评论转换为基于词向量的评论向量，完成每一条评论的特征提取，其中，所述词向量为所述评论中每一个词语的词语概率最大化的最优解，所述评论向量为每一条评论中的所有词向量的平均值；评论分类部，存储有分类模型，将所述评论向量输入到所述分类模型中进行训练，得到每一条评论的评论类型，其中，所述特征提取部包括：第一设定单元，设定词向量训练窗口的大小、词向量的维度和变化阈值；映射单元，将分词后的评论文本库中的所有评论中的词去重复后形成词汇表，建立分词后的评论文本库的词语到词汇表中词语的映射；词向量查找表构建单元，将上述词汇表中的每一个词语的词向量的每一维的数值设定

为变量，构建词向量查找表；第一更新单元，随机生成所述词向量查找表构建单元中各词向量在各维度的数值，设定词向量训练窗口内一个词语为所述词语所在评论的中心词，通过所述中心词的词向量预测所述评论中其他词语的预测概率，通过所述预测概率采用平均对数法和迭代方法不断更新所述其他词语的词向量在每一维度的数值，直到所述数值的变化值小于变化阈值，完成词向量查找表的更新，其中，所述数值的变化值为：

[0006]

$$B^{a}_{word_{n,I+J},k}=v^{a}_{word_{n,I+J},k}-v^{a-1}_{word_{n,I+J},k}=[\frac{1}{wordc(doc_n)}-O^{a-1}(v_{word_{n,I+J}})]\times v^{a-1}{}_{word_{n,I},k}$$

[0007]

$$O^{a-1}(word_{n,I+J})=\frac{1}{wordc(doc_n)}\sum_{I=1}^{wordc(doc_n)}\sum_{-win/2\le J\le win/2,j\neq 0}\log p^{a-1}(word_{n,I+J}\mid word_{n,I})$$

[0008]

$$p^{a-1}(word_{n,I+J}\mid word_{n,I})=\frac{\exp[(v^{a-1}_{word_{n,I+J}})^{T}\bullet v^{a-1}{}_{word_{n,I}}]}{\sum_{X=1}^{m}\exp[(v^{a-1}_{word_X})^{T}\bullet v^{a-1}{}_{word_{n,I}}]}$$

[0009] 其中，a为迭代次数，为自然数；

[0010] $wordc(doc_n)$为第n条评论的词语总数；

[0011] m为词汇表中的词语总数；

[0012] win为词向量训练窗口的大小；

[0013] $word_{n,I}$为第n条评论的第I个词语；

[0014] $v^{a-1}{}_{word_{n,I},k}$为第a-1次迭代中，第n条评论中第I个词语的第k维的数值；

[0015] $v^{a-1}_{word_{n,I}}$为第a-1次迭代中，第n条评论的第I个词语的词向量；

[0016] $v^{a-1}_{word_X}$为第a-1次迭代中，词汇表的第X个词语的词向量；

[0017] $p^{a-1}(word_{n,I+J}\mid word_{n,I})$为第a-1次迭代中，通过中心词$word_{n,I}$词向量预测得到词语$word_{n,I+J}$词向量的预测概率；

[0018] $O^{a-1}(word_{n,I+J})$为第a-1次迭代中，第n条评论的除中心词外各词语的预测概率的对数平均值；

[0019] $B^{a}_{word_{n,I+J},k}$为词语$word_{n,I+J}$第k维数值在第a-1次迭代和第a次迭代的数值变化；评论向量构建单元，通过计算每一条评论中的所有词向量的平均值，将所述评论的信息替换为评论向量。

[0020] 根据本发明的另一个方面，提供一种利用上述情感分析系统进行情感分析的方法，包括：采集电影评论，形成评论文本库；对评论文本库中的每一条评论进行分词，构建分词后的评论文本库；将分词后的评论文本库中的所有评论中的词去重复后形成词汇表，建立分词后的评论文本库的词语到词汇表中的词语的映射；设定词向量维度，将上述词汇表中的每一个词的词向量的每一维的数值设定为变量，构建词向量查找表；随机生成所述词向量查找表的各词向量在各维度的数值；设定词向量训练窗口的大小，以所述词向量训练

窗口内一个词语为所述词语所在评论的中心词，通过所述中心词的词向量预测所述评论中其他词语的预测概率，通过所述预测概率采用平均对数法和迭代方法不断更新所述其他词语的词向量在每一维度的数值，直到所述数值的变化值小于变化阈值，完成词向量查找表的更新；将每一条评论中的词向量映射到所述更新后的词向量查找表的数值进行平均计算，从而将每一条评论的文本信息替换为评论向量；将每一条评论的评论向量代入到分类模型中进行训练，得到每一条评论的评论类型。

[0021] 本发明所述基于词向量的针对电影评论信息的情感分析系统及方法，采用将评论文本转换成基于词向量的评论向量，词向量和评论向量的训练是无监督学习，能够克服维护情感词典和手工标注文本的巨大工作量问题，另外，评论向量是对词向量的简单的向量求平均运算，计算过程的消耗小，因此方法的实现过程十分简单，而且有效。

附图说明

[0022] 通过参考以下结合附图的说明及权利要求书的内容，并且随着对本发明的更全面理解，本发明的其它目的及结果将更加明白及易于理解。在附图中：

[0023] 图1是本发明基于词向量的针对电影评论信息的情感分析系统的一个实施例的构成框图；

[0024] 图2是本发明所述情感分析系统的特征提取部的构成框图；

[0025] 图3是本发明所述情感分析系统的评论文本处理部的构成框图；

[0026] 图4是本发明所述有向无环图的示意图；

[0027] 图5是本发明基于词向量的针对电影评论信息的情感分析方法的一个实施例的流程图；

[0028] 图6是本发明所述对评论文本库中的每一条评论进行分词的方法的流程图；

[0029] 图7是本发明基于词向量的针对电影评论信息的情感分析系统的另一个实施例的构成框图；

[0030] 图8是本发明基于词向量的针对电影评论信息的情感分析方法的另一个实施例的流程图；

[0031] 图9是本发明分类模型构建部的构成框图。

[0032] 在所有附图中相同的标号指示相似或相应的特征或功能。

具体实施方式

[0033] 在下面的描述中，出于说明的目的，为了提供对一个或多个实施例的全面理解，阐述了许多具体细节。然而，很明显，也可以在没有这些具体细节的情况下实现这些实施例。以下将结合附图对本发明的具体实施例进行详细描述。

[0034] 以下将结合附图对本发明的具体实施例进行详细描述。

[0035] 图1是本发明基于词向量的针对电影评论信息的情感分析系统，如图1所示，所述情感分析系统，包括：

[0036] 采集部100，采集电影评论，形成评论文本库corpus，其中，

[0037] $$\text{corpus}=\begin{pmatrix} doc_1 \\ \vdots \\ doc_D \end{pmatrix}$$

[0038] 其中：doc_D表示第D条电影评论文本，例如，利用正则表示法从广播电视公司已有的节目数据库或者利用网站API接口从网站上或者利用网络爬虫从视频网站上或者上述三种方式任意组合采集电影评论文本及电影评分数据；

[0039] 评论文本处理部200，对评论文本库corpus中的每一条评论进行分词，构建分词后的评论文本库$\text{corpus}_{segment}$，其中，

[0040] $$\text{corpus}_{segment}=\begin{pmatrix} docseg_1 \\ \vdots \\ docseg_D \end{pmatrix}=\begin{pmatrix} word_{1,1},word_{1,2},\cdots word_{1,wordc(doc_1)} \\ \vdots \\ word_{D,1},word_{D,2},\cdots word_{D,wordc(doc_D)} \end{pmatrix}$$

[0041] 其中，$docseg_D$是第D条分词后的电影评论，$word_{D,1}$是第D条电影评论中第1个词，$wordc(doc_D)$是第D条电影评论的词语总数；

[0042] 特征提取部300，对分词后的评论文本库$\text{corpus}_{segment}$中的每一条评论转换为基于词向量的评论向量，完成每一条评论的特征提取，其中，所述词向量为所述评论中每一个词语的词语概率最大化的最优解，所述评论向量为每一条评论中的所有词向量的平均值，详细地，将在图2中进行描述；

[0043] 评论分类部400，存储有分类模型，将所述评论向量输入到所述分类模型中进行训练，得到每一条评论的评论类型。

[0044] 优选地，还包括：判断部500，判断所述采集部100采集的评论中是否具有评分，将具有评分的评论和不具有评分的评论分类储存，且将所述评分存储到评分数据库。

[0045] 另外，优选地，所述评论分类部400还设定评分标准，高于标准的评论的评分设为1，不高于所述标准的评论的评分设为-1，从而输出具有评分的电影评论的评论类型，其中，1表示该条评论为正倾向，-1表示该条评论为负倾向，所述评分标准可以根据具有评分的评论从分类模型中得到的评论类型设定，优选地，所述评分标准为满分值的一半。

[0046] 图2是本发明所述情感分析系统的特征提取部的构成框图，如图2所示，所述特征提取部300包括：

[0047] 第一设定单元310，设定词向量训练窗口的大小、词向量的维度和变化阈值，例如，分词后的评论文本库$\text{corpus}_{segment}$包括两条评论，即

[0048] $$\text{corpus}_{segment}=\begin{pmatrix} docseg_1 \\ docseg_2 \end{pmatrix}=\begin{pmatrix} 我,很,喜欢,这,部,电影 \\ 演技,太,差 \end{pmatrix}$$

[0049] 在第一设定单元310可以设定词向量训练窗口的大小win＝6，词向量的维度数dim＝10，词向量变化阈值为0.0001；

[0050] 映射单元320，将分词后的评论文本库中的所有评论中的词语去重复后形成词汇表，建立分词后的评论文本库的词语到词汇表中的词语的映射，例如，分词后的评论文本库中的词语去重后形成的词汇表为V＝{w_1,w_2,・・・w_m}，则建立所述评论文本库中的词$word_{i,j}$到与其相同的词汇表中的词语w_k的映射，其中，m为词汇表V的总词汇数，w_k是词汇表中第k个词；1≤k≤m，又如，词汇表V＝{w_1,w_2,・・・w_9}＝{“我”,“很”,“喜欢”,“这”,“部”,

“电影”,“演技”,“太”,“差”},建立诸如$word_{1,1}=w_1$的多条映射;

[0051] 词向量查找表构建单元330,将上述词汇表中的每一个词语的词向量的每一维的数值设定为变量,构建词向量查找表,例如,词向量查找表LT

[0052] $$LT=\left\langle v_{w_1} \quad v_{w_2} \quad \cdots \quad v_{w_m} \right\rangle$$

[0053] $$v_{w_i}=\left\langle v_{wd_{i,1}} \quad v_{wd_{i,2}} \quad \cdots \quad v_{wd_{i,\dim}} \right\rangle^T$$

[0054] 其中,v_{w_i}是词汇表中第i个词的词向量;$v_{wd_{i,\dim}}$是词汇表中第i个词的词向量中第dim维的数值,例如,$LT=\left\langle v_{w_1} \quad v_{w_2} \quad v_{w_3} \quad v_{w_4} \quad v_{w_5} \quad v_{w_6} \quad v_{w_7} \quad v_{w_8} \quad v_{w_9} \right\rangle$;

[0055] 第一更新单元340,随机生成所述词向量查找表构建单元中各词向量在各维度的数值,设定词向量训练窗口内一个词语为所述词语所在评论的中心词,通过所述中心词的词向量预测所述评论中其他词语的预测概率,通过所述预测概率采用平均对数法和迭代方法不断更新所述其他词语的词向量在每一维度的数值,直到所述数值的变化值小于变化阈值(例如,0.0001),完成词向量查找表的更新,其中,所述数值的变化值为:

[0056]

$$B^a_{word_{n,I+J},k}=v^a_{word_{n,I+J},k}-v^{a-1}_{word_{n,I+J},k}=[\frac{1}{wordc(doc_n)}-O^{a-1}(v_{word_{n,I+J}})]\times v^{a-1}{}_{word_{n,I},k}$$

[0057]

$$O^{a-1}(word_{n,I+J})=\frac{1}{wordc(doc_n)}\sum_{I=1}^{wordc(doc_n)}\sum_{-win/2\le J\le win/2,j\neq 0}\log p^{a-1}(word_{n,I+J}\mid word_{n,I})$$

[0058]

$$p^{a-1}(word_{n,I+J}\mid word_{n,I})=\frac{\exp[(v^{a-1}_{word_{n,I+J}})^T\bullet v^{a-1}{}_{word_{n,I}}]}{\sum_{X=1}^{m}\exp[(v^{a-1}_{word_X})^T\bullet v^{a-1}{}_{word_{n,I}}]}$$

[0059] 其中,a为迭代次数,为自然数;

[0060] wordc(docn)为第n条评论的词语总数;

[0061] m为词汇表中的词语总数;

[0062] win为词向量训练窗口的大小;

[0063] $word_{n,I}$为第n条评论的第I个词语;

[0064] $v^{a-1}word_{n,I},k$为第a-1次迭代中,第n条评论中第I个词语的第k维的数值;

[0065] $v^{a-1}_{word_{n,I+J},k}$为第a-1次迭代中,第n条评论中第I+J个词语的第k维的数值;

[0066] $v^{a}_{word_{n,I+J},k}$为第a次迭代中,第n条评论中第I+J个词语的第k维的数值;

[0067] $v^{a-1}_{word_{n,I}}$为第a-1次迭代中,第n条评论的第I个词语的词向量;

[0068] $v^{a-1}_{word_X}$为第a-1次迭代中,词汇表的第X个词语的词向量;

[0069] $p^{a-1}(word_{n,I+J}|word_{n,I})$为第a-1次迭代中,通过中心词$word_{n,I}$词向量预测得到词语

$word_{n,I+J}$词向量的预测概率；

[0070] $O^{a-1}(word_{n,I+J})$为第a-1次迭代中，第n条评论的除中心词外各词语的预测概率的对数平均值；

[0071] $B^{a}_{word_{n,I+J},k}$为词语$word_{n,I+J}$第k维数值在第a-1次迭代和第a次迭代的数值变化；

[0072] 评论向量构建单元340，通过计算每一条评论中的所有词向量的平均值，将所述评论的信息替换为评论向量：

[0073] $$rv^{(D)}=\frac{1}{wordc(doc_D)}\sum_{t=1}^{wordc(doc_D)} v_{word_{D,t}}$$

[0074] $RV=(rv^{(1)}\ldots rv^{(n)}\cdots rv^{(D)})$

[0075] 其中，$rv(D)$是电影评论docsegD的电影评论向量；RV表示评论向量矩阵。

[0076] 优选地，所述第一更新单元随机生成所述词向量查找表中所述变量的初始值不小于0且不大于1，例如，$v_{w_i}=[0.63851984,\ 0.24690361,\ 0.64098248,0.49932827,\ 0.37650779,0.96173185,\ 0.31899279,\ 0.00182628,\ 0.65638327,0.54103694]^{T}$；

[0077] 此外，优选地，所述特征提取部300还包括：第一判断单元350，判断每一条评论的词语总数是否大于词向量训练窗口的大小，其中，当所述评论的词语总数不大于词向量训练窗口的大小时，选择所述评论中的一个词为中心词，对所述词向量查找表进行更新；当所述评论的词语总数大于词向量训练窗口的大小时，所述评论在所述词向量窗口中从左往右或者从右往左显示，依次选择所述词向量训练窗口中的一个词语为中心词，对所述词向量查找表进行更新，例如，词向量训练窗口的大小win＝3，评论$docseg_1$[我，很，喜欢，这，部，电影]中词语数为$wordc(doc_1)$＝6，首先以窗口[“我”，“很”，“喜欢”]中的“很”为的中心词，对词向量查找表进行更新，然后以窗口[“很”，“喜欢”，“这”]中的“喜欢”为的中心词，对词向量查找表进行更新。

[0078] 当评论文本库中的评论很多时，优选地，所述第一更新单元340，随机筛选r个评论，更新满足阈值条件的所述评论中词语的词向量，重复进行上述筛选，直到词汇表中所有词于的词向量更新完成，例如，r取m/100或m/10。

[0079] 图3是本发明所述情感分析系统的评论文本处理部的构成框图，如图3所示，所述评论文本处理部200包括：

[0080] 第一分词单元210，对每一条电影评论遍历，根据句尾的标点符号以及空格符，将每一条评论分割为一个或多个短句，例如，

[0081] $$corpus_{sentences}=\begin{pmatrix} sent_{1,1},sent_{1,2},\cdots sent_{1,sentc(doc_1)} \\ \vdots \\ sent_{D,1},sent_{D,2},\cdots sent_{D,sentc(doc_D)} \end{pmatrix}$$

[0082] 其中，$corpus_{sentences}$是电影评论语料按标点符号切割后的短句语料，$1\leqslant n\leqslant D$.$sentc(doc_D)$是第D条电影评论的总短句数，$sent_{i,j}$是第i条电影评论中的第j条短句，$1\leqslant j\leqslant sentc(doc_i)$。

[0083] 第二分词单元220，基于Trie树结构对评论文本库进行词图扫描，生成每一条评论

中汉字所有可能成词情况所构成的有向无环图，所述有向无环图由多个结点和连结节点的边组成，如图4所示，有向图是指图中的每条边具有一个方向的图，有向无环图是指，无法从任意顶点出发经过若干条边回到该点的有向图，例如，

[0084] $sent_{i,j}=(chara_1,chara_2,\cdots,chara_l)$

[0085] 其中，每一个$chara_l$(字符l)是$sent_{i,j}$中的第l个字符；l是$sent_{i,j}$的总字符数。

[0086] 考虑每个字符左边和右边的位置，则有l+1个点对应，点的编号从0到l，把候选词看成边，可以根据词典生成一个有向无环图，如图4所示，有向无环图是一个有向正权重的图，有向无环图中的边都是词典中的词语，边的起点和终点分别是词的开始和结束位置。对字符数为l的$sent_{i,j}$，假设$chara_1chara_2$(字符1字符2)、$chara_2chara_3$(字符2字符3)和$chara_{l-1}chara_l$(字符l-1字符l)在词典中，其他字符组合均不在词典中，则生成有向无环图如下切割方案有两个选择：路径1：0-1-3-4-5-……-(l-1)-(l+1)；路径2：0-2-3-4-5-……-(l-1)-(l+1)。

[0087] 第一确定单元230，采用了动态规划查找有向无环图基于词频的最大概率路径，找出基于词频的最大切分路径，确定切割方案。

[0088] 图5是本发明基于词向量的针对电影评论信息的情感分析方法的流程图，如图5所示，所述情感分析方法包括：

[0089] 首先，在步骤S510中，采集电影评论，形成评论文本库；

[0090] 在步骤S520中，对评论文本库中的每一条评论进行分词，构建分词后的评论文本库；

[0091] 在步骤S530中，将分词后的评论文本库中的所有评论中的词语去重复后形成词汇表，建立分词后的评论文本库的词语到词汇表中的词语的映射；

[0092] 在步骤S540中，设定词向量的维度，将上述词汇表中的每一个词语的词向量的每一维的数值设定为变量，构建词向量查找表；

[0093] 在步骤S550中，随机生成所述词向量查找表的各词向量在各维度的数值；

[0094] 在步骤S560中，设定词向量训练窗口的大小，以所述词向量训练窗口内一个词语为所述词语所在评论的中心词，通过所述中心词的词向量预测所述评论中其他词语的预测概率，通过所述预测概率采用平均对数法和迭代方法不断更新所述其他词语的词向量在每一维度的数值，直到所述数值的变化值小于变化阈值，完成词向量查找表的更新；

[0095] 在步骤S570中，将每一条评论中的词向量映射到所述更新后的词向量查找表的数值进行平均计算，从而将每一条评论的文本信息替换为评论向量；

[0096] 在步骤S580中，将每一条评论的评论向量代入到分类模型中进行训练，得到每一条评论的评论类型。

[0097] 在步骤S520中，所述对评论文本库中的每一条评论进行分词的方法，如图6所示，包括：

[0098] 首先，在步骤S521中，根据句尾的标点符号以及空格符，将评论文本库中的每一条评论分割为一个或多个短句；

[0099] 在步骤S522中，基于Trie树结构对评论文本库进行词图扫描，得到所述短句中汉字所有可能成词情况，构成有向无环图，得到每一条评论的所述短语的多个分割方案；

[0100] 在步骤S523中，记录每一条评论的所述多个分割方案形成的所有词语，以及该词

语在所述评论文本库中出现的次数，得到每一个词语出现的频率，其中，所述频率$p(w_n)$为：

[0101] $$p(w_n)=\frac{freq(w_n)}{\sum_{i=1}^{t}freq(w_i)}$$

[0102] 其中，$p(w_n)$是词语w_n出现的频率；

[0103] $freq(w_n)$是词语w_n出现的频数；

[0104] t为所有有向无环图中所有可能成词情况构成的词语的总数；

[0105] $\sum_{i=1}^{t}freq(w_i)$是词汇表中所有词语出现的总频数；

[0106] 在步骤S524中，将每一种切割方案中不存在与词典中的词语的频率用词典中最小频率替代，基于所述频率，采用查找最大概率路径的方法确定每一条评论的切割方案，优选地，采用从右往左查找最大概率路径的方法确定每一条评论的切割方案，例如，

[0107] $p(Node_{l+1})=1.0$

[0108] $p(Node_s)=p(Node_{s+1})\times max(p(w_{s,last}))$，$1\leqslant s\leqslant l$

[0109] 其中，$Node_{l+1}$是评论doc_i从左往右第l+1个节点；

[0110] $Node_s$是评论doc_n中的第j条短句$sent_{i,j}$从左往右第s个节点；

[0111] $p(Node_{s+1})$是评论doc_i从左往右第s+1个节点的概率，即最后一个字符的右边节点；

[0112] $p(Node_s)$是评论doc_i从左往右第s个节点的概率；

[0113] $w_{s,last}$是到$Node_s$为止的从左往右最后的候选词语；

[0114] $p(w_{s,last})$表示$w_{s,last}$出现的频率；

[0115] $max(w_{s,last})$表示到$Node_s$为止的最后的候选词语的最大出现概率。

[0116] 通过上式，得到每一个短句中不同节点设置的不同概率，找到每一个短句最大概率的节点设置，即获得该短句的最大概率路径，确定了该短句的切割方案。

[0117] 在本发明的另一个实施例中，如图7所示，所述情感分析系统1000除上述采集部100、评论文本处理部200、特征提取部300、判断部500还包括：

[0118] 分类模型构建部600，用于构建分类模型，其中，

[0119] 所述采集部100采集电影评论；

[0120] 所述判断部500，判断所述采集部100采集的评论中是否具有评分，将具有评分的评论和不具有评分的评论分类储存，且将所述评分存储到评分数据库；

[0121] 所述分类模型构建部600包括：

[0122] 评分训练模型构建单元610，构建评分训练模型，其中，所述评分训练模型包括设定评分标准，高于标准的评论的评分值设为1，不高于所述标准的评论的评分值设为-1，将每一条评论的评分相对于所述评分标准存储成只包括1和-1的数据集，其中，1表示该条评论为正倾向，-1表示该条评论为负倾向；

[0123] 分类模型构建单元620，构建包括变量的分类模型；

[0124] 第一获得单元630，通过评论文本处理部和特征提取部对具有评分的评论进行处理，获得所述评论对应的评论向量，将存储所述评论的评分的评分数据库通过评分训练模型转变为只包括1和-1的数据集；

[0125] 第二获得单元640，利用所述评论向量及其对应的数据集获得分类模型的变量，详

细地，将在图9中进行描述。

[0126] 优选地，上述情感分析系统还包括评论分类部400，通过存储的分类模型得到具有评分的评论的评论向量的评论类型，对分类模型构建部600的变量起到修正作用。

[0127] 采用上述情感分析系统对电影评论进行情感分析的方法，如图8所示，包括：

[0128] 在步骤S810中，采集电影评论；

[0129] 在步骤S820中，判断所述电影评论中是否具有评分，将具有评分的评论和不具有评分的评论分类储存，且将所述评分存储到评分数据库，例如，将具有评分的评论存储到评论文本库$\mathrm{corpus}'=\begin{pmatrix} doc_1' \\ \vdots \\ doc_G' \end{pmatrix}$，其中，G为具有评分的电影评论的总条数；$\mathrm{doc}_G'$表示第G条电影评论文本；将所述评分存储到评分数据库$\mathrm{scores}=\begin{pmatrix} score^{(1)} \\ \vdots \\ score^{(G)} \end{pmatrix}$，其中，$\mathrm{score}^{(G)}$表示第G条电影评论对应评分，$0\leqslant \mathrm{score}^{(G)} \leqslant \mathrm{score}_{max}$；$\mathrm{score}_{max}$为满分值，通常$\mathrm{score}_{max}\in(5,10)$；

[0130] 在步骤S830中，构建评分训练模型，其中，所述评分训练模型包括设定评分标准，高于标准的评论的评分设为1，不高于所述标准的评论的评分设为-1，将每一条评论的评分相对于所述评分标准存储成只包括1和-1的数据集，其中，1表示该条评论为正倾向，-1表示该条评论为负倾向，例如，以满分值得一半为标准，将评分数据库scores的评分数据转为待遇测变量数据集Y，具体地：

[0131] $$\mathrm{Y}=\left(y^{(1)} \quad \cdots \quad y^{(n)} \quad \cdots \quad y^{(G)}\right)=\begin{pmatrix} f(score^{(1)}) \\ \vdots \\ f(score^{(n)}) \\ \vdots \\ f(score^{(G)}) \end{pmatrix}^T$$

[0132] $$y^{(n)}=f(score^{(n)})=\begin{cases} 1 & score^{(n)} > score_{max}/2 \\ -1 & score^{(n)} \leq score_{max}/2 \end{cases}$$

[0133] 其中，$\mathrm{y}^{(1)}$，$\mathrm{y}^{(n)}$，$\mathrm{y}^{(G)}$是第1、n、G条评论相对评分标准的数据；

[0134] Y是具有评分的评论相对于评分标准的数据构成的数据集；

[0135] 在步骤S840中，构建包括变量的分类模型，其中，所述包括变量的分类模型为：

[0136] $\mathrm{y}^{i}=\mathrm{w}\cdot \mathrm{rv}^{(i)}-\mathrm{b}$

[0137] $$w=\sum_{n=1}^{G}\alpha_n y^{(n)} rv'^{(n)}$$

[0138] $\arg\min_{w,b}\frac{\|w\|}{2}$,且$y^{(n)}(w \cdot rv'^{(n)}-b)\geqslant 1$

[0139] $$E_n=\sum_{k=1}^{G}\alpha_k y^{(k)}\left\langle rv'^{(k)},rv'^{(n)}\right\rangle-y^{(n)}$$

[0140] $$\left\langle rv'^{(k)},rv'^{(n)}\right\rangle=\sum_{s=1}^{\dim}rv_s'^{(k)}\cdot rv_s'^{(n)}$$

[0141] 其中，$rv^{(i)}$为不具有评分的评论向量；

[0142] G为具有评分的评论向量的总个数；

[0143] E_n是被优化的目标函数；

[0144] w和b为变量，其中，w为垂直于评论向量平面的向量，b为阈值；

[0145] α为拉格朗日参数，是D维向量，$\alpha\in R^D$

[0146] α_k是拉格朗日参数α第k维度的分量；

[0147] $y^{(k)}$，$y^{(n)}$是具有评分的第k个和第n个评论向量在数据集中的数值；

[0148] $rv'^{(k)}$，$rv'^{(n)}$分别是具有评分的第k个和第n个评论向量；

[0149] $rv_s'^{(k)}$，$rv_s'^{(k)}$分别表示第k个和第n个评论向量的第s维分量，$1\leqslant s\leqslant dim$；

[0150] $\langle rv'^{(k)},rv'^{(n)}\rangle$表示对评论向量$rv'^{(k)}$，$rv'^{(n)}$求向量内积；

[0151] 在步骤S850中，将所述评分数据库中各评论的评分在评分训练模型中进行训练，得到各评论的所述数据集；

[0152] 在步骤S860中，通过评论文本处理部和特征提取部对存储具有评分的评论进行处理，得到所述评论的评论向量，例如，$RV'=(rv'^{(1)}\cdots rv'^{(n)}\cdots rv'^{(G)})$，其中，RV'表示评论向量矩阵；$rv'^{(G)}$是电影评论$doc_G'$的电影评论向量；

[0153] 在步骤S870中，利用具有评分的评论的评论向量及其对应的数据集确定分类模型的变量，完成分类模型的构建；

[0154] 在步骤S880中，通过评论文本处理部和特征提取部对存储不具有评分的评论进行处理，得到所述评论的评论向量；

[0155] 在步骤S890中，将上述评论向量输入上述分类模型，得到不具有评分的评论的评论类型。

[0156] 图9示出了所述第二获得单元的构成框图，如图9所示，所述第二获得单元640包括：

[0157] 第二设定单元641，初始化拉格朗日参数$\alpha=(\alpha_1,\alpha_2,\cdots,\alpha_D)$、阈值b及b的待选参数$b_1$和$b_2$，$\alpha_1=\alpha_2=\cdots=\alpha_D=0$，$b=b_1=b_2=0$；设置指定精度ε(例如$\varepsilon=10^{-5}$)；设置容差tol和调和函数C；

[0158] 计算单元642，遍历第一获取单元630中的评论向量，计算每一个评论向量对应的E函数值，例如，评论向量$rv'^{(n)}$对应的E函数值E_n；

[0159] 第二判断单元643，判断评论向量的评分相对评分标准的数据与其E函数值的乘积以及其拉格朗日参数是否满足下述条件：$y^{(n)}E_n<-tol$且$\alpha_n<C$，或者$y^{(n)}E_n>tol$且$\alpha_n>0$，如果存在均不满足上述两个条件的评论向量，则发送指令给计算单元642，重新计算该评论向

量的E值;如果满足上述两个条件之一,发送指令给第二更新单元644;

[0160] 第二更新单元644,将满足第二判断单元643条件的第一获取单元630中的任意两个评论向量配对,更新每一个评论向量的拉格朗日参数,其中,

[0161] $$\alpha_n^{(new)}=\begin{cases} H & \alpha_n^{(new,wnc)}>H \\ \alpha_n^{(new,wnc)} & L\le\alpha_n^{(new,wnc)}\le H \\ L & \alpha_n^{(new,wnc)}<L \end{cases}$$

[0162] $$\alpha_n^{(new,wnc)}=\alpha_n^{(old)}-\frac{y^{(n)}(E_k-E_n)}{\eta}$$

[0163] $\eta=2<rv'^{(k)},rv'^{(n)}>-<rv'^{(k)},rv'^{(k)}>-<rv'^{(n)},rv'^{(k)}>$,且$\eta<0$

[0164]

$$\begin{cases} L=max(0,\alpha_n^{(old)}-\alpha_k^{(old)}),H=min(C,C+\alpha_n^{(old)}-\alpha_k^{(old)}) & y^{(n)}\neq y^{(k)} \\ L=max(0,\alpha_n^{(old)}+\alpha_k^{(old)}-C),H=min(C,C+\alpha_n^{(old)}-\alpha_k^{(old)}) & y^{(n)}=y^{(k)} \end{cases},\text{且}L\neq H$$

$$\alpha_k^{(new)}=\alpha_k^{(old)}+y^{(k)}y^{(n)}(\alpha_n^{(old)}-\alpha_n^{(new)}),\text{且}|\alpha_n^{(new)}-\alpha_n^{(old)}|\geqslant\varepsilon$$

[0165] 其中,$rv'^{(n)}$和$rv'^{(k)}$为满足第二判断单元643条件的第一获取单元630中的任意两个评论向量;

[0166] $\alpha_n^{(old)}$和$\alpha_k^{(old)}$为更新前评论向量$rv'^{(n)}$和$rv'^{(k)}$对应的拉格朗日参数;

[0167] $\alpha_n^{(new,wnc)}$为更新过程中评论向量$rv'^{(n)}$待判断的新的拉格朗日参数;

[0168] $\alpha_n^{(new)}$和$\alpha_k^{(new)}$是更新后评论向量$rv'^{(n)}$和$rv'^{(k)}$对应的拉格朗日参数;

[0169] L和H为$\alpha_n^{(old)}$更新的上限和下限;

[0170] η是被优化的目标函数E_n的二阶导数;

[0171] 第三更新单元645,更新每一个评论相量对应的阈值,具体地,包括:

[0172] $$b^{(n)}=\begin{cases} b_1^{(new)} & 0<\alpha_k^{(new)}<C \\ b_2^{(new)} & 0<\alpha_n^{(new)}<C \\ \dfrac{b_1^{(new)}+b_2^{(new)}}{2} & \text{其他情况} \end{cases}$$

[0173] $b_1^{(new)}=b_1^{(old)}-E_k-y^{(k)}(\alpha_k^{(new)}-\alpha_k^{(old)})<rv'^{(k)},rv'^{(k)}>-y^{(n)}(\alpha_n^{(new)}-\alpha_n^{(old)})<rv'^{(k)},rv'^{(n)}>$

[0174] $b_2^{(new)}=b_2^{(old)}-E_n-y^{(k)}(\alpha_k^{(new)}-\alpha_k^{(old)})<rv'^{(k)},rv'^{(n)}>-y^{(n)}(\alpha_n^{(new)}-\alpha_n^{(old)})<rv'^{(n)},rv'^{(n)}>$

[0175] 其中,$b^{(n)}$为更新后评论相量$rv'^{(n)}$对应的阈值b的值;

[0176] $b_1^{(old)}$、$b_2^{(old)}$为之前保留的待选参数b_1和b_2;

[0177] 第二确定单元646,根据更新后各评论向量的拉格朗日参数及其对应的阈值确定变量参数w和b,其中,

[0178] $$w=\sum_{n=1}^{G}\alpha_n y^{(n)} rv'^{(n)}$$

[0179] $$\arg\min_{w,b}\frac{\|w\|}{2}$$，且$y^{(n)}(w\cdot rv'^{(n)}-b)\geq 1$。

[0180] 利用上述第二获得单元640确定分类模型的变量的方法包括：

[0181] 初始化具有评分的各评论向量的拉格朗日参数和阈值，设置指定精度和容差；

[0182] 计算上述评论向量对应的E函数值；

[0183] 筛选出满足条件的评论向量，其中，所述条件为$y^{(n)}E_n<-tol$且$\alpha_n<C$，或者$y^{(n)}E_n>tol$且$\alpha_n>0$；

[0184] 将满足上述条件的评论向量中任意两个评论向量进行配对，更新每一个评论向量的拉格朗日参数；

[0185] 更新上述每一个评论向量对应的阈值；

[0186] 根据更新后各评论向量的拉格朗日参数及其对应的阈值确定变量参数w和b。

[0187] 综上所述，参照附图以示例的方式描述了根据本发明提出的基于词向量的针对电影评论信息的情感分析方法及系统。但是，本领域技术人员应当理解，对于上述本发明所提出的系统及方法，还可以在不脱离本发明内容的基础上做出各种改进。因此，本发明的保护范围应当由所附的权利要求书的内容确定。

基于词向量的针对电影评论信息的情感分析系统1000

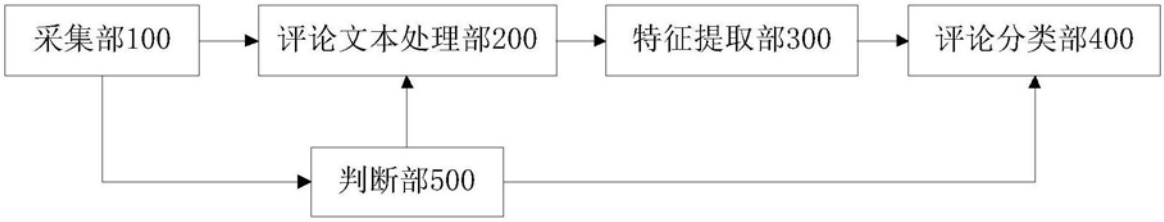


图1

特征提取部300

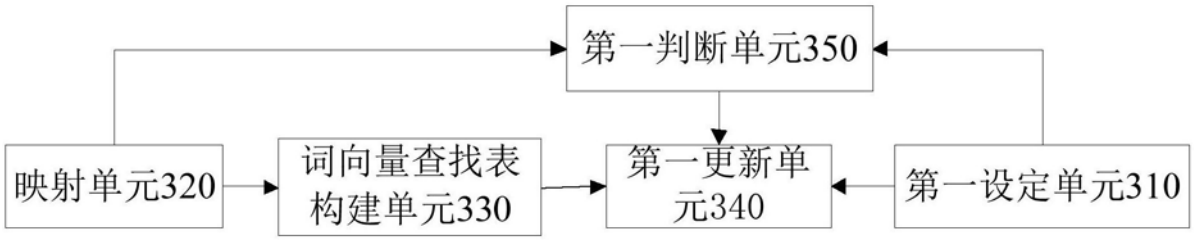


图2

评论文本处理部200

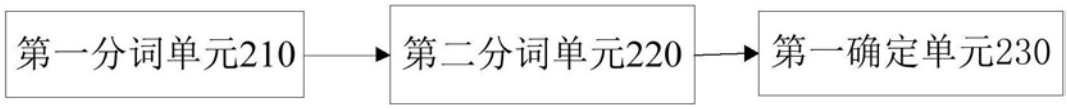


图3

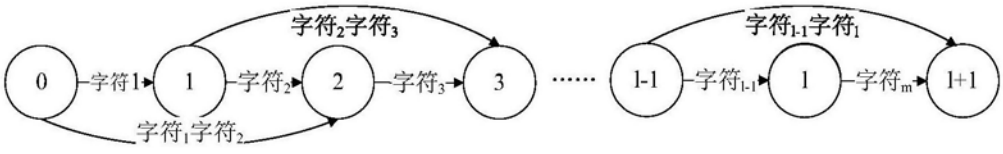


图4

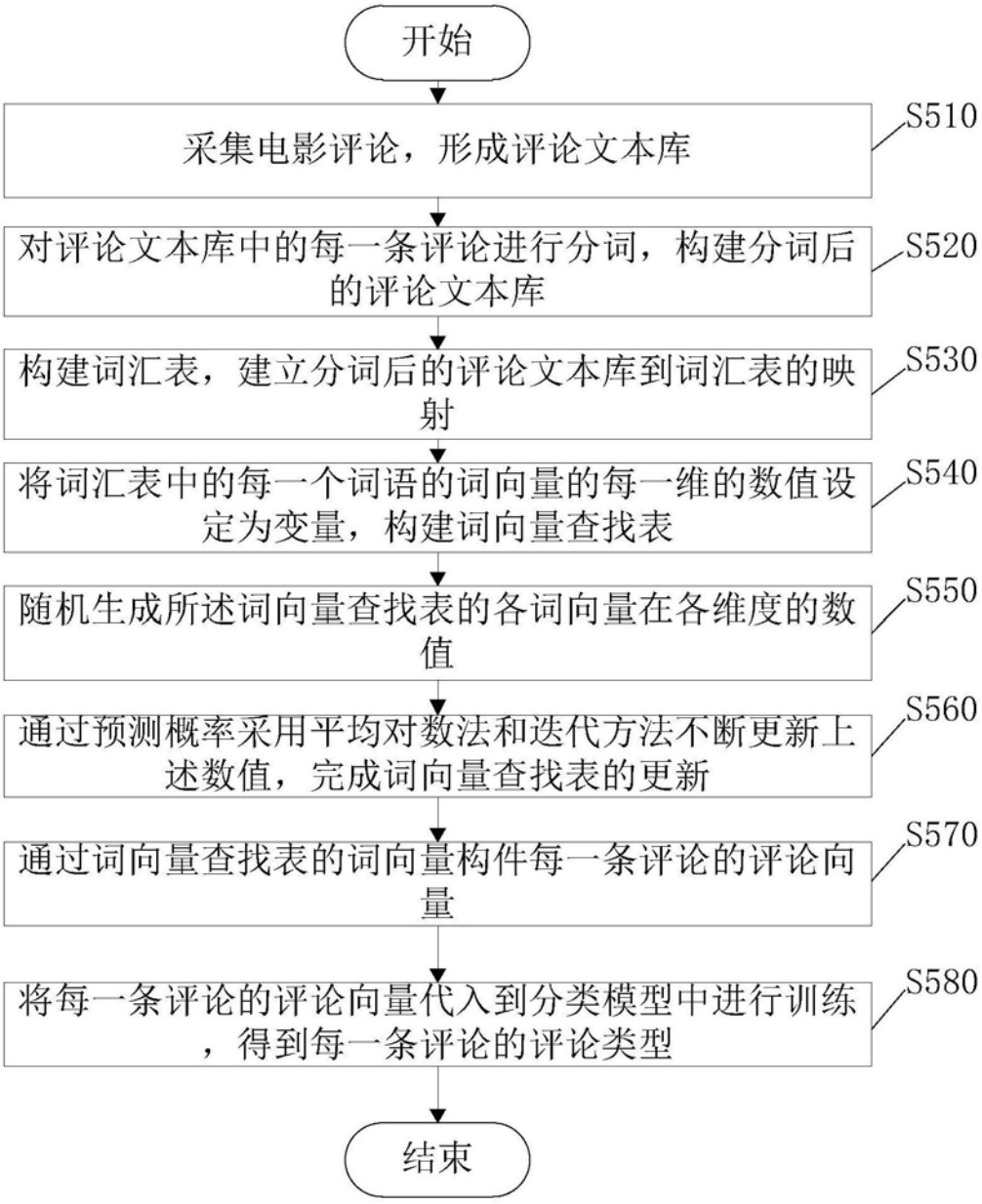

开始
采集电影评论，形成评论文本库
S510
对评论文本库中的每一条评论进行分词，构建分词后的评论文本库
S520
构建词汇表，建立分词后的评论文本库到词汇表的映射
S530
将词汇表中的每一个词语的词向量的每一维的数值设定为变量，构建词向量查找表
S540
随机生成所述词向量查找表的各词向量在各维度的数值
S550
通过预测概率采用平均对数法和迭代方法不断更新上述数值，完成词向量查找表的更新
S560
通过词向量查找表的词向量构件每一条评论的评论向量
S570
将每一条评论的评论向量代入到分类模型中进行训练，得到每一条评论的评论类型
S580
结束

图5

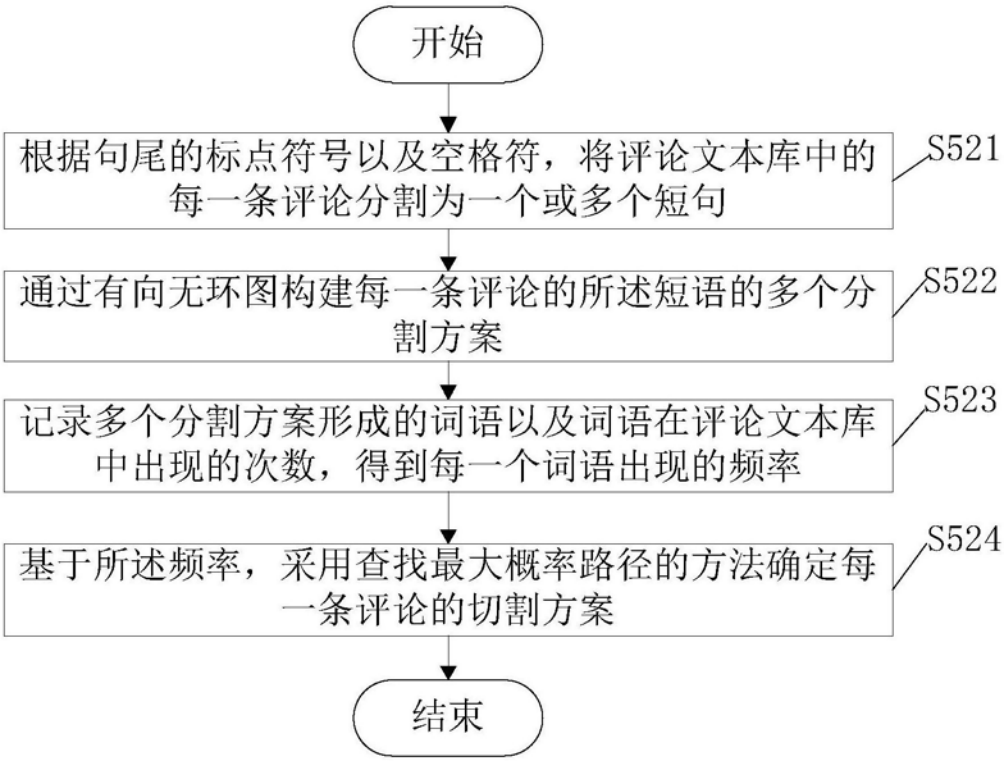

开始
根据句尾的标点符号以及空格符，将评论文本库中的每一条评论分割为一个或多个短句
S521
通过有向无环图构建每一条评论的所述短语的多个分割方案
S522
记录多个分割方案形成的词语以及词语在评论文本库中出现的次数，得到每一个词语出现的频率
S523
基于所述频率，采用查找最大概率路径的方法确定每一条评论的切割方案
S524
结束

图6

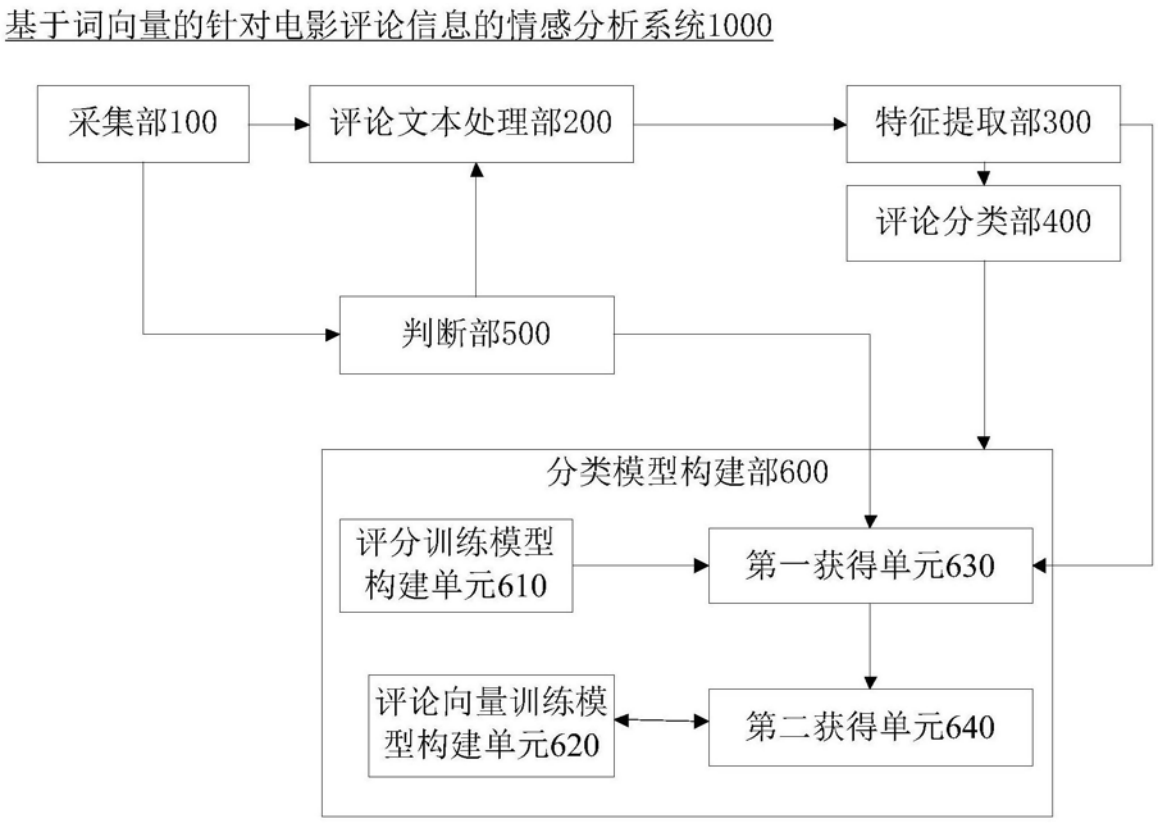

基于词向量的针对电影评论信息的情感分析系统1000
采集部100
评论文本处理部200
特征提取部300
评论分类部400
判断部500
分类模型构建部600
评分训练模型构建单元610
第一获得单元630
评论向量训练模型构建单元620
第二获得单元640

图7

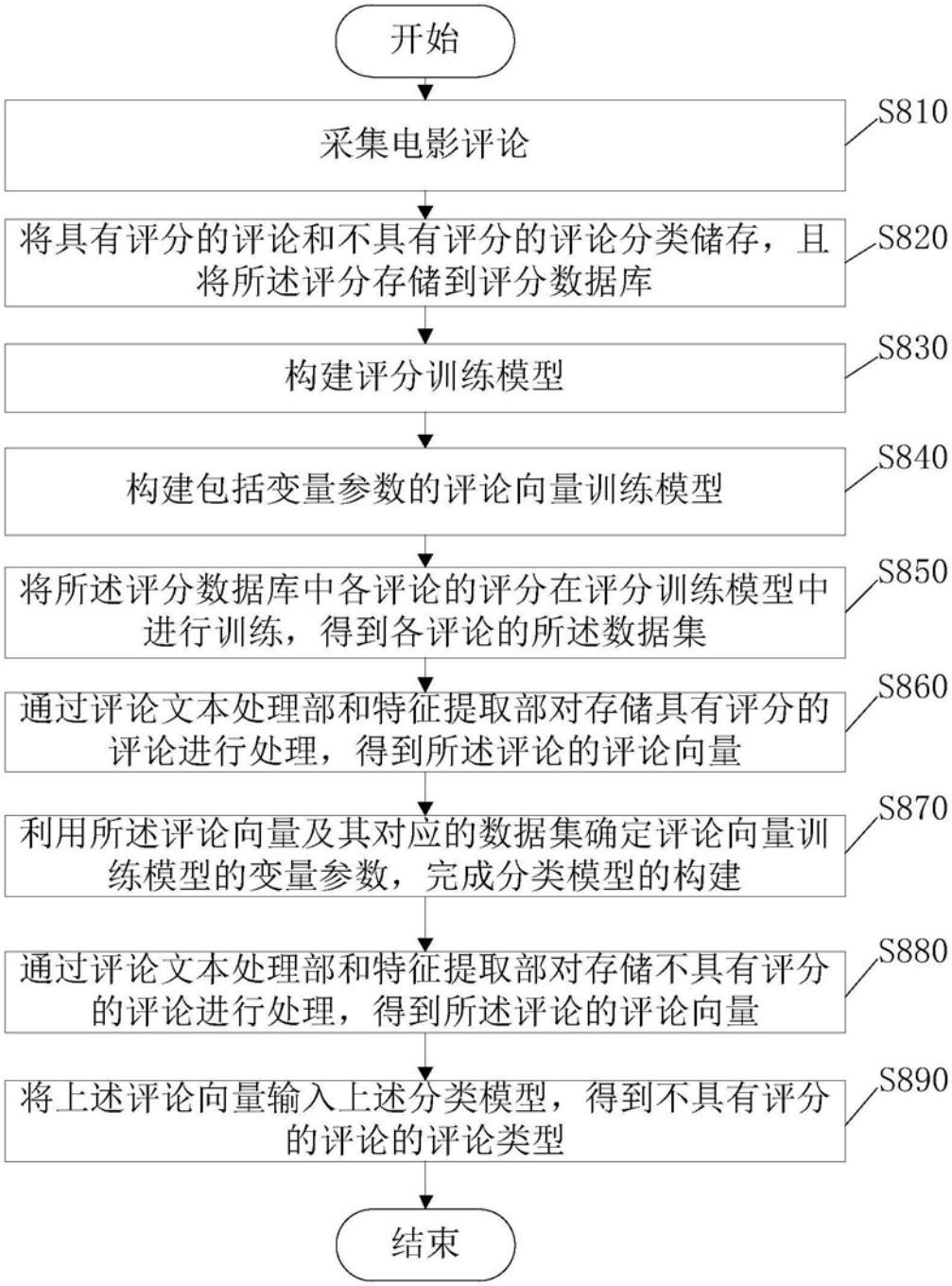

开始
采集电影评论
S810
将具有评分的评论和不具有评分的评论分类储存，且将所述评分存储到评分数据库
S820
构建评分训练模型
S830
构建包括变量参数的评论向量训练模型
S840
将所述评分数据库中各评论的评分在评分训练模型中进行训练，得到各评论的所述数据集
S850
通过评论文本处理部和特征提取部对存储具有评分的评论进行处理，得到所述评论的评论向量
S860
利用所述评论向量及其对应的数据集确定评论向量训练模型的变量参数，完成分类模型的构建
S870
通过评论文本处理部和特征提取部对存储不具有评分的评论进行处理，得到所述评论的评论向量
S880
将上述评论向量输入上述分类模型，得到不具有评分的评论的评论类型
S890
结束

图8

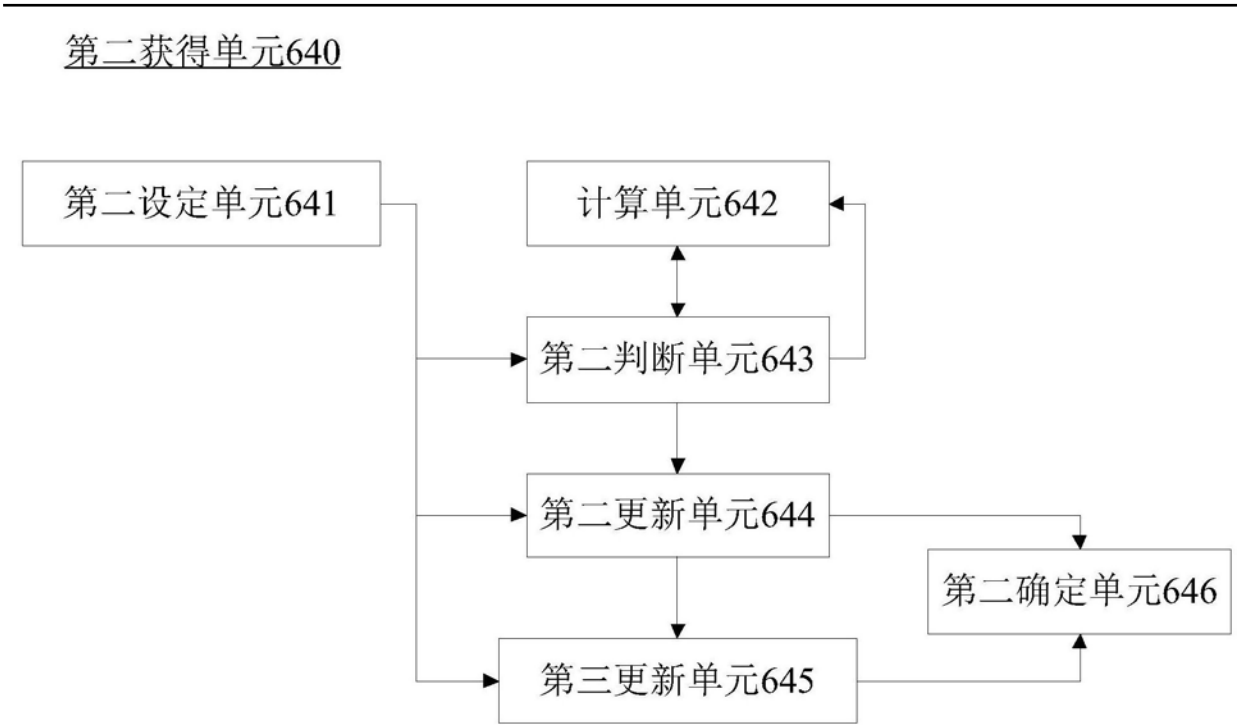

第二获得单元640
第二设定单元641
计算单元642
第二判断单元643
第二更新单元644
第二确定单元646
第三更新单元645

图9