(19)国家知识产权局

(12)发明专利

(10)授权公告号 CN 115828924 B
(45)授权公告日 2023.07.25

(21)申请号 202211470449.5

(22)申请日 2022.11.23

(65)同一申请的已公布的文献号
申请公布号 CN 115828924 A

(43)申请公布日 2023.03.21

(66)本国优先权数据
202211455587.6 2022.11.21 CN

(73)专利权人 武汉工商学院
地址 430065 湖北省武汉市洪山区黄家湖西路3号

(72)发明人 胡成松 薛莲

(74)专利代理机构 武汉知律知识产权代理事务所(普通合伙) 42307
专利代理师 田常娟

(51)Int.Cl.
G06F 40/30(2020.01)
G06F 16/9532(2019.01)

(56)对比文件
CN 114329225 A,2022.04.12

审查员 王永波

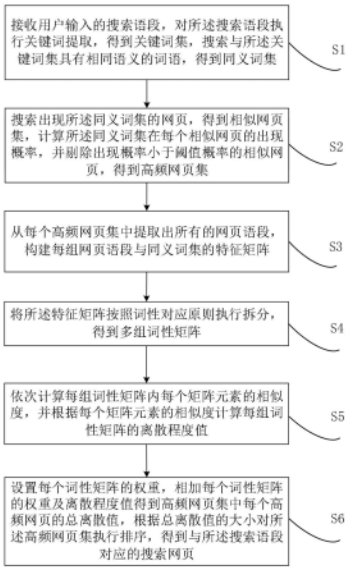
权利要求书4页 说明书11页 附图2页

(54)发明名称

自然语言语义理解方法及装置

(57)摘要

本发明涉及一种自然语言语义理解方法及装置,包括:接收用户的搜索语段,根据所述搜索语段构建同义词集并匹配得到高频网页集,从每个高频网页中提取出所有的网页语段,构建每组网页语段与同义词集的多组词性矩阵,依次计算每组词性矩阵内每个矩阵元素的相似度,并根据每个矩阵元素的相似度计算每组词性矩阵的离散程度值,设置每个词性矩阵的权重,相加每个词性矩阵的权重及离散程度值得到高频网页集中每个高频网页的总离散值,根据总离散值的大小对所述高频网页集执行排序,得到与所述搜索语段对应的搜索网页。本发明可解决未考虑用户输入的搜索语段与网页语段的词性而导致语义理解准确率不高的问题。



CN 115828924 B

1.一种自然语言语义理解方法,其特征在于,所述方法包括:

接收用户输入的搜索语段,对所述搜索语段执行关键词提取,得到关键词集,搜索与所述关键词集具有相同语义的词语,得到同义词集;

搜索出现所述同义词集的网页,得到相似网页集,计算所述同义词集在每个相似网页的出现概率,并剔除出现概率小于阈值概率的相似网页,得到高频网页集,出现概率的计算方法为:

$p_j(w_1,w_2,\cdots,w_n)=p(w_1)p(w2|w1)p(w3|w1,w2)\cdots p(w_n|w_1,w_2\cdots w_{n-1})$其中,$(w_1,w_2,\cdots,w_n)$表示同义词集的词向量组的每个词向量,$p_j(w_1,w_2,\cdots,w_n)$表示同义词集在第j个相似网页的出现概率,|表示条件概率;

从每个高频网页中提取出所有的网页语段,构建每组网页语段与同义词集的特征矩阵:

$$\boldsymbol{N}_{1i}=\boldsymbol{N}_1\times\mathbf{N}_i^{\mathrm{T}}=\begin{pmatrix} w_1w_{i1} & \cdots & w_nw_{i1} \\ \cdots & & \cdots \\ w_1w_{im} & \cdots & w_nw_{im} \end{pmatrix}$$

$N_1=(w_1,w_2,\cdots,w_n)$

$N_i=(w_{i1},w_{i2},\cdots,w_{im})$

其中,N_{1i}表示同义词集与第i组网页语段构建的特征矩阵,N_1为所述同义词集对应的词向量组,n为同义词集对应的词向量组的维度,N_i为第i组网页语段对应的词向量组,m为第i组网页语段的词向量组的维度;

将所述特征矩阵按照词性对应原则执行拆分,得到多组词性矩阵,所述拆分包括:

构建名词空矩阵、动词空矩阵、形容词空矩阵、数词空矩阵和代词空矩阵;

遍历出所述特征矩阵中每个矩阵元素,其中每个矩阵元素均由同义词集的词向量和网页语段的词向量组成;

剔除同义词集的词向量和网页语段的词向量的词性不同的矩阵元素后,判断每个矩阵元素所属于的词性类别;

根据词性类别依次将矩阵元素填写至名词空矩阵、动词空矩阵、形容词空矩阵、数词空矩阵和代词空矩阵中的其中一个,直至遍历完所有的矩阵元素,得到名词矩阵、动词矩阵、形容词矩阵、数词矩阵和代词矩阵共五组词性矩阵;

依次计算每组词性矩阵内每个矩阵元素的相似度,并根据每个矩阵元素的相似度计算每组词性矩阵的离散程度值;

设置每个词性矩阵的权重,相加每个词性矩阵的权重及离散程度值得到高频网页集中每个高频网页的总离散值,其中,总离散值的计算过程为:

$Q_r^{all}=\alpha_N Q_r^N+\alpha_V Q_r^V+\alpha_A Q_r^A+\alpha_M Q_r^M+\alpha_R Q_r^R$

其中,Q_r^{all}表示总离散值,Q_r^N,Q_r^V,Q_r^A,Q_r^M,Q_r^R分别表示名词矩阵、动词矩阵、形容词矩阵、数词矩阵和代词矩阵的离散程度值,α_N,α_V,α_A,α_M,α_R分别表示名词矩阵、动词矩阵、形容词矩阵、数词矩阵和代词矩阵的权重;

根据总离散值的大小对所述高频网页集执行排序,得到与所述搜索语段对应的搜索网页。

2.如权利要求1所述的自然语言语义理解方法,其特征在于,所述$p(w_n|w_1,w_2\cdots w_{n-1})$的

计算方法为：

$$p(w_n \mid w_1, w_2 \dots w_{n-1}) = \frac{C(web, w_1, w_2 \dots w_{n-1})}{C(web, s)}$$

其中，C(web,w_1,w_2…w_{n-1})表示w_1,w_2…w_{n-1}各个词向量在相似网页web中的出现次数，C(web,s)表示w_1,w_2…w_{n-1},w_n各个词向量在相似网页web中的出现次数，s表示(w_1,w_2,…,w_n)。

3.如权利要求2所述的自然语言语义理解方法，其特征在于，所述从每个高频网页中提取出所有的网页语段，构建每组网页语段与同义词集的特征矩阵，包括：

利用Word2Vec算法并按照每个网页语段在对应的高频网页的出现顺序构建词向量组；

将网页语段的词向量组与同义词集的词向量组执行阶乘，得到所述特征矩阵。

4.如权利要求3所述的自然语言语义理解方法，其特征在于，所述依次计算每组词性矩阵内每个矩阵元素的相似度，包括：

按照从左到右、从上到下的原则，依次遍历出每组词性矩阵的每个矩阵元素，其中每个矩阵元素均由w_iw_{ij}类型组成；

计算w_iw_{ij}的相似度，其中相似度的计算方法为：

$$dis = \frac{w_i * w_{ij}}{\sqrt{(w_i)^2}\sqrt{(w_{ij})^2}}$$

其中，w_i表示同义词集对应的词向量组对应的第i个词向量，w_{ij}表示第i组网页语段对应的词向量组中的第j个词向量，dis表示相似度，w_i*w_{ij}表示w_i与w_{ij}的内积，$(w_i)^2$及$(w_{ij})^2$均表示词向量的模的积。

5.如权利要求4所述的自然语言语义理解方法，其特征在于，所述根据每个矩阵元素的相似度计算每组词性矩阵的离散程度值，包括：

将每组词性矩阵所对应的所有的矩阵元素的相似度按照从大到小排列，得到排序相似度集；

从所述排序相似度集中提取末尾25%的相似度值及靠前25%的相似度值；

根据末尾25%的相似度值及靠前25%的相似度值计算对应的词性矩阵的离散程度值：

$Q_r = Q_3 - Q_1$

$$Q_1 = \frac{\sum_{i=1}^{l} dis_i}{\frac{l+1}{4}\arg(\sum_i^l dis_i)}$$

$$Q_3 = \frac{\sum_{i=u*0.75}^{o} dis_i}{\frac{o+1}{4}\arg(\sum_i^o dis_i)}$$

其中，Q_r表示离散程度值，arg表示求平均值，u表示排序相似度集的个数，l表示排序相似度集中靠前25%的相似度值的个数，o表示排序相似度集中末尾25%的相似度值的个数，dis_i表示相似度值。

6.如权利要求5所述的自然语言语义理解方法，其特征在于，所述对所述搜索语段执行关键词提取，得到关键词集，包括：

对所述搜索语段执行分词处理，得到词语集；

剔除所述词语集中的停用词，得到所述关键词集。

7.如权利要求6所述的自然语言语义理解方法，其特征在于，所述搜索与所述关键词集

具有相同语义的词语，得到同义词集，包括：

启动预先构建的语义库，将所述关键词集中的每个关键词作为语义库的搜索词，搜索得到与每个关键词具有相同语义的词语；

汇总每个关键词及对应的相同语义的词语，得到所述同义词集。

8.如权利要求7所述的自然语言语义理解方法，其特征在于，所述阈值概率设定为0.5。

9.一种自然语言语义理解装置，其特征在于，所述装置包括：

同义词集构建模块，用于接收用户输入的搜索语段，对所述搜索语段执行关键词提取，得到关键词集，搜索与所述关键词集具有相同语义的词语，得到同义词集；

高频网页计算模块，用于搜索出现所述同义词集的网页，得到相似网页集，计算所述同义词集在每个相似网页的出现概率，并剔除出现概率小于阈值概率的相似网页，得到高频网页集，出现概率的计算方法为：

$p_j(w_1,w_2,\cdots,w_n)=p(w_1)p(w2|w1)p(w3|w1,w2)\cdots p(w_n|w_1,w_2\cdots w_{n-1})$其中，$(w_1,w_2,\cdots,w_n)$表示同义词集的词向量组的每个词向量，$p_j(w_1,w_2,\cdots,w_n)$表示同义词集在第j个相似网页的出现概率，|表示条件概率；

特征矩阵构建模块，用于从每个高频网页中提取出所有的网页语段，构建每组网页语段与同义词集的特征矩阵：

$$\boldsymbol{N}_{1i}=\boldsymbol{N}_1\times\mathbf{N}_i^{\mathrm{T}}=\begin{pmatrix} w_1w_{i1} & \cdots & w_nw_{i1} \\ \cdots & & \cdots \\ w_1w_{im} & \cdots & w_nw_{im} \end{pmatrix}$$

$N_1=(w_1,w_2,\cdots,w_n)$

$N_i=(w_{i1},w_{i2},\cdots,w_{im})$

其中，N_{1i}表示同义词集与第i组网页语段构建的特征矩阵，N_1为所述同义词集对应的词向量组，n为同义词集对应的词向量组的维度，N_i为第i组网页语段对应的词向量组，m为第i组网页语段的词向量组的维度；

将所述特征矩阵按照词性对应原则执行拆分，得到多组词性矩阵，所述拆分包括：

构建名词空矩阵、动词空矩阵、形容词空矩阵、数词空矩阵和代词空矩阵；

遍历出所述特征矩阵中每个矩阵元素，其中每个矩阵元素均由同义词集的词向量和网页语段的词向量组成；

剔除同义词集的词向量和网页语段的词向量的词性不同的矩阵元素后，判断每个矩阵元素所属于的词性类别；

根据词性类别依次将矩阵元素填写至名词空矩阵、动词空矩阵、形容词空矩阵、数词空矩阵和代词空矩阵中的其中一个，直至遍历完所有的矩阵元素，得到名词矩阵、动词矩阵、形容词矩阵、数词矩阵和代词矩阵共五组词性矩阵；

网页匹配模块，用于依次计算每组词性矩阵内每个矩阵元素的相似度，并根据每个矩阵元素的相似度计算每组词性矩阵的离散程度值，设置每个词性矩阵的权重，相加每个词性矩阵的权重及离散程度值得到高频网页集中每个高频网页的总离散值，其中，总离散值的计算过程为：

$Q_r^{all}=\alpha_N Q_r^N+\alpha_V Q_r^V+\alpha_A Q_r^A+\alpha_M Q_r^M+\alpha_R Q_r^R$

其中，Q_r^{all}表示总离散值，Q_r^N，Q_r^V，Q_r^A，Q_r^M，Q_r^R分别表示名词矩阵、动词矩阵、形容词矩

阵、数词矩阵和代词矩阵的离散程度值，α_N，α_V，α_A，α_M，α_R分别表示名词矩阵、动词矩阵、形容词矩阵、数词矩阵和代词矩阵的权重，根据总离散值的大小对所述高频网页集执行排序，得到与所述搜索语段对应的搜索网页。

自然语言语义理解方法及装置

技术领域

[0001] 本发明涉及自然语言处理技术领域，尤其涉及一种自然语言语义理解方法及装置。

背景技术

[0002] 语义理解在信息检索领域有着广泛应用，特别是网页检索对语义理解的智能化程度极其重要，决定了是否能满足用户利用搜索语段找到目标网页。

[0003] 目前基于网页检索中的语言理解主要使用文本相似度计算方法，即遍历出所有与搜索语段相关的目标网页，通过计算目标网页中的网页语段与搜索语段的相似度达到理解搜索语段的目的。

[0004] 上述方法虽然可实现语义理解，但由于未考虑网页语段和搜索语段不同词性对搜素结果的影响，从而所匹配出的搜索网页准确率有待进一步提高。

发明内容

[0005] 本发明提供一种自然语言语义理解方法、计算机可读存储介质，其主要目的在于解决未考虑用户输入的搜索语段与网页语段的词性而导致语义理解准确率不高的问题。

[0006] 为实现上述目的，本发明提供的一种自然语言语义理解方法，包括：

[0007] 接收用户输入的搜索语段，对所述搜索语段执行关键词提取，得到关键词集，搜索与所述关键词集具有相同语义的词语，得到同义词集；

[0008] 搜索出现所述同义词集的网页，得到相似网页集，计算所述同义词集在每个相似网页的出现概率，并剔除出现概率小于阈值概率的相似网页，得到高频网页集，出现概率的计算方法为：

[0009] $p_j(w_1,w_2,\cdots,w_n)=p(w_1)p(w2|w1)p(w3|w1,w2)\cdots p(w_n|w_1,w_2\cdots w_{n-1})$

[0010] 其中，$(w_1,w_2,\cdots,w_n)$表示同义词集的词向量组的每个词向量，$p_j(w_1,w_2,\cdots,w_n)$表示同义词集在第j个相似网页的出现概率，|表示条件概率；

[0011] 从每个高频网页中提取出所有的网页语段，构建每组网页语段与同义词集的特征矩阵：

[0012] $$\boldsymbol{N}_{1i}=\boldsymbol{N}_1\times\mathbf{N}_i^{\mathrm{T}}=\begin{pmatrix} w_1w_{i1} & \cdots & w_nw_{i1} \\ \cdots & & \cdots \\ w_1w_{im} & \cdots & w_nw_{im} \end{pmatrix}$$

[0013] $N_1=(w_1,w_2,\cdots,w_n)$

[0014] $N_i=(w_{i1},w_{i2},\cdots,w_{im})$

[0015] 其中，N_{1i}表示同义词集与第i组网页语段构建的特征矩阵，N_1为所述同义词集对应的词向量组，n为同义词集对应的词向量组的维度，N_i为第i组网页语段对应的词向量组，m为第i组网页语段的词向量组的维度；

[0016] 将所述特征矩阵按照词性对应原则执行拆分，得到多组词性矩阵，所述拆分包括：

[0017] 构建名词空矩阵、动词空矩阵、形容词空矩阵、数词空矩阵和代词空矩阵；

[0018] 遍历出所述特征矩阵中每个矩阵元素，其中每个矩阵元素均由同义词集的词向量和网页语段的词向量组成；

[0019] 剔除同义词集的词向量和网页语段的词向量的词性不同的矩阵元素后，判断每个矩阵元素所属于的词性类别；

[0020] 根据词性类别依次将矩阵元素填写至名词空矩阵、动词空矩阵、形容词空矩阵、数词空矩阵和代词空矩阵中的其中一个，直至遍历完所有的矩阵元素，得到名词矩阵、动词矩阵、形容词矩阵、数词矩阵和代词矩阵共五组词性矩阵；

[0021] 依次计算每组词性矩阵内每个矩阵元素的相似度，并根据每个矩阵元素的相似度计算每组词性矩阵的离散程度值；

[0022] 设置每个词性矩阵的权重，相加每个词性矩阵的权重及离散程度值得到高频网页集中每个高频网页的总离散值，，其中，总离散值的计算过程为：

[0023] $Q_r^{all}=\alpha_N Q_r^N+\alpha_V Q_r^V+\alpha_A Q_r^A+\alpha_M Q_r^M+\alpha_R Q_r^R$

[0024] 其中，Q_r^{all}表示总离散值，Q_r^N，Q_r^V，Q_r^A，Q_r^M，Q_r^R分别表示名词矩阵、动词矩阵、形容词矩阵、数词矩阵和代词矩阵的离散程度值，α_N，α_V，α_A，α_M，α_R分别表示名词矩阵、动词矩阵、形容词矩阵、数词矩阵和代词矩阵的权重；

[0025] 根据总离散值的大小对所述高频网页集执行排序，得到与所述搜索语段对应的搜索网页。

[0026] 可选地，所述$p(w_n|w_1,w_2\cdots w_{n-1})$的计算方法为：

[0027] $$p(w_n \mid w_1, w_2 \ldots w_{n-1}) = \frac{C(web, w_1, w_2 \ldots w_{n-1})}{C(web, s)}$$

[0028] 其中，$C(web,w_1,w_2\cdots w_{n-1})$表示$w_1,w_2\cdots w_{n-1}$各个词向量在相似网页web中的出现次数，$C(web,s)$表示$w_1,w_2\cdots w_{n-1},w_n$各个词向量在相似网页web中的出现次数，s表示$(w_1,w_2,\cdots,w_n)$。

[0029] 可选地，所述从每个高频网页中提取出所有的网页语段，构建每组网页语段与同义词集的特征矩阵，包括：

[0030] 利用Word2Vec算法并按照每个网页语段在对应的高频网页的出现顺序构建词向量组；

[0031] 将网页语段的词向量组与同义词集的词向量组执行阶乘，得到所述特征矩阵。

[0032] 可选地，所述依次计算每组词性矩阵内每个矩阵元素的相似度，包括：

[0033] 按照从左到右、从上到下的原则，依次遍历出每组词性矩阵的每个矩阵元素，其中每个矩阵元素均由w_iw_{ij}类型组成；

[0034] 计算w_iw_{ij}的相似度，其中相似度的计算方法为：

[0035] $$dis = \frac{w_i * w_{ij}}{\sqrt{(w_i)^2}\sqrt{(w_{ij})^2}}$$

[0036] 其中，w_i表示同义词集对应的词向量组对应的第i个词向量，w_{ij}表示第i组网页语段对应的词向量组中的第j个词向量，dis表示相似度，w_i*w_{ij}表示w_i与w_{ij}的内积，$(w_i)^2$及$(w_{ij})^2$均表示词向量的模的积。

[0037] 可选地，所述根据每个矩阵元素的相似度计算每组词性矩阵的离散程度值，包括：

[0038] 将每组词性矩阵所对应的所有的矩阵元素的相似度按照从大到小排列，得到排序

相似度集；

[0039] 从所述排序相似度集中提取末尾25%的相似度值及靠前25%的相似度值；

[0040] 根据末尾25%的相似度值及靠前25%的相似度值计算对应的词性矩阵的离散程度值：

[0041] $Q_r = Q_3 - Q_1$

[0042] $$Q_1 = \frac{\sum_{i=1}^{l} dis_i}{\frac{l+1}{4}\arg(\sum_i^l dis_i)}$$

[0043] $$Q_3 = \frac{\sum_{i=u*0.75}^{o} dis_i}{\frac{o+1}{4}\arg(\sum_i^o dis_i)}$$

[0044] 其中，Q_r表示离散程度值，arg表示求平均值，u表示排序相似度集的个数，l表示排序相似度集中靠前25%的相似度值的个数，o表示排序相似度集中末尾25%的相似度值的个数，dis_i表示相似度值。

[0045] 可选地，所述对所述搜索语段执行关键词提取，得到关键词集，包括：

[0046] 对所述搜索语段执行分词处理，得到词语集；

[0047] 剔除所述词语集中的停用词，得到所述关键词集。

[0048] 可选地，所述搜索与所述关键词集具有相同语义的词语，得到同义词集，包括：

[0049] 启动预先构建的语义库，将所述关键词集中的每个关键词作为语义库的搜索词，搜索得到与每个关键词具有相同语义的词语；

[0050] 汇总每个关键词及对应的相同语义的词语，得到所述同义词集。

[0051] 可选地，所述阈值概率设定为0.5。

[0052] 为实现上述目的，本发明还提供一种自然语言语义理解装置，包括：

[0053] 同义词集构建模块，用于接收用户输入的搜索语段，对所述搜索语段执行关键词提取，得到关键词集，搜索与所述关键词集具有相同语义的词语，得到同义词集；

[0054] 高频网页计算模块，用于搜索出现所述同义词集的网页，得到相似网页集，计算所述同义词集在每个相似网页的出现概率，并剔除出现概率小于阈值概率的相似网页，得到高频网页集，出现概率的计算方法为：

[0055] $p_j(w_1, w_2, \cdots, w_n) = p(w_1)p(w2|w1)p(w3|w1,w2)\cdots p(w_n|w_1, w_2 \cdots w_{n-1})$

[0056] 其中，$(w_1, w_2, \cdots, w_n)$表示同义词集的词向量组的每个词向量，$p_j(w_1, w_2, \cdots, w_n)$表示同义词集在第j个相似网页的出现概率，|表示条件概率；

[0057] 特征矩阵构建模块，用于从每个高频网页中提取出所有的网页语段，构建每组网页语段与同义词集的特征矩阵：

[0058] $$\boldsymbol{N}_{1i} = \boldsymbol{N}_1 \times \mathbf{N}_i^{\mathrm{T}} = \begin{pmatrix} w_1 w_{i1} & \cdots & w_n w_{i1} \\ \cdots & & \cdots \\ w_1 w_{im} & \cdots & w_n w_{im} \end{pmatrix}$$

[0059] $N_1 = (w_1, w_2, \cdots, w_n)$

[0060] $N_i = (w_{i1}, w_{i2}, \cdots, w_{im})$

[0061] 其中，N_{1i}表示同义词集与第i组网页语段构建的特征矩阵，N_1为所述同义词集对应的词向量组，n为同义词集对应的词向量组的维度，N_i为第i组网页语段对应的词向量组，m

为第i组网页语段的词向量组的维度；

[0062] 将所述特征矩阵按照词性对应原则执行拆分，得到多组词性矩阵，所述拆分包括：

[0063] 构建名词空矩阵、动词空矩阵、形容词空矩阵、数词空矩阵和代词空矩阵；

[0064] 遍历出所述特征矩阵中每个矩阵元素，其中每个矩阵元素均由同义词集的词向量和网页语段的词向量组成；

[0065] 剔除同义词集的词向量和网页语段的词向量的词性不同的矩阵元素后，判断每个矩阵元素所属于的词性类别；

[0066] 根据词性类别依次将矩阵元素填写至名词空矩阵、动词空矩阵、形容词空矩阵、数词空矩阵和代词空矩阵中的其中一个，直至遍历完所有的矩阵元素，得到名词矩阵、动词矩阵、形容词矩阵、数词矩阵和代词矩阵共五组词性矩阵；

[0067] 网页匹配模块，用于依次计算每组词性矩阵内每个矩阵元素的相似度，并根据每个矩阵元素的相似度计算每组词性矩阵的离散程度值，设置每个词性矩阵的权重，相加每个词性矩阵的权重及离散程度值得到高频网页集中每个高频网页的总离散值，其中，总离散值的计算过程为：

[0068] $Q_r^{all}=\alpha_N Q_r^N+\alpha_V Q_r^V+\alpha_A Q_r^A+\alpha_M Q_r^M+\alpha_R Q_r^R$

[0069] 其中，Q_r^{all}表示总离散值，Q_r^N，Q_r^V，Q_r^A，Q_r^M，Q_r^R分别表示名词矩阵、动词矩阵、形容词矩阵、数词矩阵和代词矩阵的离散程度值，α_N，α_V，α_A，α_M，α_R分别表示名词矩阵、动词矩阵、形容词矩阵、数词矩阵和代词矩阵的权重，根据总离散值的大小对所述高频网页集执行排序，得到与所述搜索语段对应的搜索网页。

[0070] 为了解决上述问题，本发明还提供一种电子设备，所述电子设备包括：

[0071] 存储器，存储至少一个指令；及

[0072] 处理器，执行所述存储器中存储的指令以实现上述所述的自然语言语义理解方法。

[0073] 为了解决上述问题，本发明还提供一种计算机可读存储介质，所述计算机可读存储介质中存储有至少一个指令，所述至少一个指令被电子设备中的处理器执行以实现上述所述的自然语言语义理解方法。

[0074] 本发明实施例为解决背景技术所述问题，接收用户输入的搜索语段，对所述搜索语段执行关键词提取，得到关键词集，搜索与所述关键词集具有相同语义的词语，得到同义词集，搜索出现所述同义词集的网页，得到相似网页集，计算所述同义词集在每个相似网页的出现概率，并剔除出现概率小于阈值概率的相似网页，得到高频网页集，可见本发明实施例为提高根据搜索语段而提高语义理解准确率，先进行网页剔除操作，即通过扩大搜索语段同义词的前提下，剔除出现概率小于阈值概率的相似网页，进一步地，从每个高频网页中提取出所有的网页语段，构建每组网页语段与同义词集的特征矩阵，其中每个特征矩阵均可拆分为名词矩阵、动词矩阵、形容词矩阵、数词矩阵和代词矩阵共五组词性矩阵，且每组词性矩阵的权重可根据搜索语段调节，从而有效解决未考虑词性而导致语义理解准确率偏低的问题。因此本发明提出的自然语言语义理解方法、电子设备及计算机可读存储介质，可以解决未考虑用户输入的搜索语段与网页语段的词性而导致语义理解准确率不高的问题。

附图说明

[0075] 图1为本发明一实施例提供的自然语言语义理解方法的流程示意图；

[0076] 图2为本发明一实施例提供的自然语言语义理解装置的功能模块图；

[0077] 图3为本发明一实施例提供的实现所述自然语言语义理解方法的电子设备的结构示意图。

[0078] 本发明目的的实现、功能特点及优点将结合实施例，参照附图做进一步说明。

具体实施方式

[0079] 应当理解，此处所描述的具体实施例仅仅用以解释本发明，并不用于限定本发明。

[0080] 本申请实施例提供一种自然语言语义理解方法。所述自然语言语义理解方法的执行主体包括但不限于服务端、终端等能够被配置为执行本申请实施例提供的该方法的电子设备中的至少一种。换言之，所述自然语言语义理解方法可以由安装在终端设备或服务端设备的软件或硬件来执行，所述软件可以是区块链平台。所述服务端包括但不限于：单台服务器、服务器集群、云端服务器或云端服务器集群等。

[0081] 参照图1所示，为本发明一实施例提供的自然语言语义理解方法的流程示意图。在本实施例中，所述自然语言语义理解方法包括：

[0082] S1、接收用户输入的搜索语段，对所述搜索语段执行关键词提取，得到关键词集，搜索与所述关键词集具有相同语义的词语，得到同义词集；

[0083] 本发明实施例中，所述搜索语段一般包括一至多个句子，用于匹配更加符合用户意向的网页。示例性的，如用户输入“专利分类技术有哪些，特别是基于深度学习的模型”。

[0084] 进一步地，所述对所述搜索语段执行关键词提取，得到关键词集，包括：

[0085] 对所述搜索语段执行分词处理，得到词语集；

[0086] 剔除所述词语集中的停用词，得到所述关键词集。

[0087] 示例性的，“专利分类技术有哪些，特别是基于深度学习的模型”执行分词处理后得到“专利”、“分类”、…、“模型”，并剔除“哪些”、“基于”、“的”等实际意义的停用词，从而得到关键词集。

[0088] 本发明实施例为提高语义理解的准确率，会根据关键词匹配出与其相近的其他词语，详细地，所述搜索与所述关键词集具有相同语义的词语，得到同义词集，包括：

[0089] 启动预先构建的语义库，将所述关键词集中的每个关键词作为语义库的搜索词，搜索得到与每个关键词具有相同语义的词语；

[0090] 汇总每个关键词及对应的相同语义的词语，得到所述同义词集。

[0091] 示例性的，本发明实施例通过绑定各类型的中文词库从而构建得到较为健全的语义库，如与“分类”具有相同语义的还包括“归类”、“区分”等。

[0092] S2、搜索出现所述同义词集的网页，得到相似网页集，计算所述同义词集在每个相似网页的出现概率，并剔除出现概率小于阈值概率的相似网页，得到高频网页集；

[0093] 本发明实施例中，可利用已公开的搜索引擎搜索与同义词集相似的网页，得到相似网页集，在此不再赘述。

[0094] 进一步地，所述计算所述同义词集在每个相似网页的出现概率，包括：

[0095] 利用Word2Vec算法并按照每个同义词集在所述搜索语段的出现顺序构建词向量

组；

[0096] 根据如下计算方法计算所述词向量组的出现概率：

[0097] $p_j(w_1,w_2,\cdots,w_n)=p(w_1)p(w2|w1)p(w3|w1,w2)\cdots p(w_n|w_1,w_2\cdots w_{n-1})$

[0098] 其中，$(w_1,w_2,\cdots,w_n)$表示同义词集的词向量组的每个词向量，$p_j(w_1,w_2,\cdots,w_n)$表示同义词集在第j个相似网页的出现概率，|表示条件概率。

[0099] 需解释的是，$p(w_n|w_1,w_2\cdots w_{n-1})$的计算方法为：

[0100] $$p(w_n \mid w_1,w_2 \ldots w_{n-1})=\frac{C(web,w_1,w_2 \ldots w_{n-1})}{C(web,s)}$$

[0101] 其中，$C(web,w_1,w_2\cdots w_{n-1})$表示$w_1,w_2\cdots w_{n-1}$各个词向量在相似网页web中的出现次数，$C(web,s)$表示$w_1,w_2\cdots w_{n-1},w_n$各个词向量在相似网页web中的出现次数，s表示$(w_1,w_2,\cdots,w_n)$。

[0102] 示例性的，若根据搜索引擎搜索共有10个相似网页，其中通过上述方法计算得到每个同义词集所构建的词向量组在10个相似网页中的出现概率分别为0.91、0.23、0.17、…、0.87，设定的阈值概率为0.5，则对应剔除出现概率小于0.5的相似网页，得到高频网页集。

[0103] S3、从每个高频网页中提取出所有的网页语段，构建每组网页语段与同义词集的特征矩阵；

[0104] 详细地，所述从每个高频网页中提取出所有的网页语段，构建每组网页语段与同义词集的特征矩阵，包括：

[0105] 利用Word2Vec算法并按照每个网页语段在对应的高频网页的出现顺序构建词向量组；

[0106] 将网页语段的词向量组与同义词集的词向量组执行阶乘，得到所述特征矩阵，其中所述特征矩阵的表达式为：

[0107] $$\boldsymbol{N}_{1i}=\boldsymbol{N}_1\times\mathbf{N}_i^{\mathrm{T}}=\begin{pmatrix} w_1w_{i1} & \cdots & w_nw_{i1} \\ \cdots & & \cdots \\ w_1w_{im} & \cdots & w_nw_{im} \end{pmatrix}$$

[0108] $N_1=(w_1,w_2,\cdots,w_n)$

[0109] $N_i=(w_{i1},w_{i2},\cdots,w_{im})$

[0110] 其中，N_{1i}表示同义词集与第i组网页语段构建的特征矩阵，N_1为所述同义词集对应的词向量组，n为同义词集对应的词向量组的维度，N_i为第i组网页语段对应的词向量组，m为第i组网页语段的词向量组的维度。

[0111] 需解释的是，特征矩阵的作用是为了方便后续从高频网页集中提取出更符合用户意图的网页。

[0112] S4、将所述特征矩阵按照词性对应原则执行拆分，得到多组词性矩阵；

[0113] 需解释的是，不同词性的词语在搜索语段及网页中的重要性不同，明显的，名称、动词的重要性要高于数词及代词。本发明实施例为了提高网页筛选的准确性，将词性按照名词、动词、形容词、数词和代词拆分特征矩阵。

[0114] 详细地，所述将所述特征矩阵按照词性对应原则执行拆分，得到多组词性矩阵，包括：

[0115] 构建名词空矩阵、动词空矩阵、形容词空矩阵、数词空矩阵和代词空矩阵；

[0116] 遍历出所述特征矩阵中每个矩阵元素，其中每个矩阵元素均由同义词集的词向量和网页语段的词向量组成；

[0117] 剔除同义词集的词向量和网页语段的词向量的词性不同的矩阵元素后，判断每个矩阵元素所属于的词性类别；

[0118] 根据词性类别依次将矩阵元素填写至名词空矩阵、动词空矩阵、形容词空矩阵、数词空矩阵和代词空矩阵中的其中一个，直至遍历完所有的矩阵元素，得到名词矩阵、动词矩阵、形容词矩阵、数词矩阵和代词矩阵，其中，词性矩阵包括名词矩阵、动词矩阵、形容词矩阵、数词矩阵和代词矩阵。

[0119] 示例性的，第i组网页语段和同义词集所组成的特征矩阵为

[0120] $$\boldsymbol{N}_{1i}=\begin{pmatrix} w_1w_{i1} & \cdots & w_nw_{i1} \\ \cdots & & \cdots \\ w_1w_{im} & \cdots & w_nw_{im} \end{pmatrix}$$

[0121] 其中特征矩阵N_{1i}的第i行第j列的矩阵元素为w_iw_{ij}，先判断w_iw_{ij}的词性是否相同，如w_i的词性为名词，w_{ij}的词性为形容词，则剔除，如w_i的词性为名词，w_{ij}的词性也为名词，则将w_iw_{ij}填入至名词空矩阵中，直至遍历完成所有的特征矩阵N_{1i}的矩阵元素。

[0122] S5、依次计算每组词性矩阵内每个矩阵元素的相似度，并根据每个矩阵元素的相似度计算每组词性矩阵的离散程度值。

[0123] 详细地，所述依次计算每组词性矩阵内每个矩阵元素的相似度，包括：

[0124] 按照从左到右、从上到下的原则，依次遍历出每组词性矩阵的每个矩阵元素，其中每个矩阵元素均由w_iw_{ij}类型组成；

[0125] 计算w_iw_{ij}的相似度，其中相似度的计算方法为：

[0126] $$dis=\frac{w_i*w_{ij}}{\sqrt{(w_i)^2}\sqrt{(w_{ij})^2}}$$

[0127] 其中，dis表示相似度，w_i*w_{ij}表示w_i与w_{ij}的内积，$(w_i)^2$及$(w_{ij})^2$均表示词向量的模的积。

[0128] 示例性的，共有五组词性矩阵：名词矩阵、动词矩阵、形容词矩阵、数词矩阵和代词矩阵，假设第i组网页语段与同义词集的名词矩阵的维度为4*4，即4行4列总有16组矩阵元素，则可以分别计算w_1*w_{i1}、w_1*w_{i2}、…、$w_{16}*w_{i16}$的相似度，共得到16组相似度。

[0129] 进一步地，所述根据每个矩阵元素的相似度计算每组词性矩阵的离散程度值，包括：

[0130] 将每组词性矩阵所对应的所有的矩阵元素的相似度按照从大到小排列，得到排序相似度集；

[0131] 从所述排序相似度集中提取末尾25%的相似度值及靠前25%的相似度值；

[0132] 根据末尾25%的相似度值及靠前25%的相似度值计算对应的词性矩阵的离散程度值：

[0133] $Q_r=Q_3-Q_1$

[0134] $$Q_1=\frac{\sum_{i=1}^{l}dis_i}{\frac{l+1}{4}\arg(\sum_i^l dis_i)}$$

[0135] $$Q_3=\frac{\sum_{i=u*0.75}^{o} dis_i}{\frac{o+1}{4}\arg(\sum_i^o dis_i)}$$

[0136] 其中，Q_r表示离散程度值，arg表示求平均值，u表示排序相似度集的个数，l表示排序相似度集中靠前25%的相似度值的个数，o表示排序相似度集中末尾25%的相似度值的个数，dis_i表示相似度值。

[0137] 需解释的是，离散程度值越大表示网页语段中的词向量与同义词集的词向量具有越大的变动性，即离散程度值越小的网页语段词向量更接近于同义词集的词向量。

[0138] 示例性的，名词矩阵共有16组相似度值，则排序得到从大到小的排序相似度集，因此提取末尾25%及靠前25%的相似度值，即分别有4组，并代入上式可得到离散程度值。

[0139] S6、设置每个词性矩阵的权重，相加每个词性矩阵的权重及离散程度值得到所述高频网页集中每个高频网页的总离散值，根据总离散值的大小对所述高频网页集执行排序，得到与所述搜索语段对应的搜索网页。

[0140] 可理解的是，不同搜索语段所侧重的词性不一样，因此所对应的词性矩阵的权重也不同，如上述搜索语段：“专利分类技术有哪些，特别是基于深度学习的模型”，更侧重于名词“分类技术”和“模型”；但若其他类型的搜索语段：“专利不保护哪些方法”，则侧重于动词“不保护”。

[0141] 详细地，所述相加每个词性矩阵的权重及离散程度值得到所述高频网页集中每个高频网页的总离散值，包括：

[0142] $$Q_r^{all}=\alpha_N Q_r^N+\alpha_V Q_r^V+\alpha_A Q_r^A+\alpha_M Q_r^M+\alpha_R Q_r^R$$

[0143] 其中，Q_r^{all}表示总离散值，Q_r^N，Q_r^V，Q_r^A，Q_r^M，Q_r^R分别表示名词矩阵、动词矩阵、形容词矩阵、数词矩阵和代词矩阵的离散程度值，α_N，α_V，α_A，α_M，α_R分别表示名词矩阵、动词矩阵、形容词矩阵、数词矩阵和代词矩阵的权重。

[0144] 可以见得，当依次计算出每个高频网页的总离散值以后，可选择总离散值最小的高频网页作为与搜索语段最匹配的搜索网页。

[0145] 本发明实施例为解决背景技术所述问题，接收用户输入的搜索语段，对所述搜索语段执行关键词提取，得到关键词集，搜索与所述关键词集具有相同语义的词语，得到同义词集，搜索出现所述同义词集的网页，得到相似网页集，计算所述同义词集在每个相似网页的出现概率，并剔除出现概率小于阈值概率的相似网页，得到高频网页集，可见本发明实施例为提高根据搜索语段而提高语义理解准确率，先进行网页剔除操作，即通过扩大搜索语段同义词的前提下，剔除出现概率小于阈值概率的相似网页，进一步地，从每个高频网页中提取出所有的网页语段，构建每组网页语段与同义词集的特征矩阵，其中每个特征矩阵均可拆分为名词矩阵、动词矩阵、形容词矩阵、数词矩阵和代词矩阵共五组词性矩阵，且每组词性矩阵的权重可根据搜索语段调节，从而有效解决未考虑词性而导致语义理解准确率偏低的问题。因此本发明提出的自然语言语义理解方法、电子设备及计算机可读存储介质，可以解决未考虑用户输入的搜索语段与网页语段的词性而导致语义理解准确率不高的问题。

[0146] 如图2所示，是本发明一实施例提供的自然语言语义理解装置的功能模块图。

[0147] 本发明所述自然语言语义理解装置100可以安装于电子设备中。根据实现的功能，所述自然语言语义理解装置100可以包括同义词集构建模块101、高频网页计算模块102、特征矩阵构建模块103及网页匹配模块104。本发明所述模块也可以称之为单元，是指一种能

够被电子设备处理器所执行，并且能够完成固定功能的一系列计算机程序段，其存储在电子设备的存储器中。

[0148] 所述同义词集构建模块101，用于接收用户输入的搜索语段，对所述搜索语段执行关键词提取，得到关键词集，搜索与所述关键词集具有相同语义的词语，得到同义词集；

[0149] 所述高频网页计算模块102，用于搜索出现所述同义词集的网页，得到相似网页集，计算所述同义词集在每个相似网页的出现概率，并剔除出现概率小于阈值概率的相似网页，得到高频网页集，出现概率的计算方法为：

[0150] $p_j(w_1,w_2,\cdots,w_n)=p(w_1)p(w2|w1)p(w3|w1,w2)\cdots p(w_n|w_1,w_2\cdots w_{n-1})$

[0151] 其中，$(w_1,w_2,\cdots,w_n)$表示同义词集的词向量组的每个词向量，$p_j(w_1,w_2,\cdots,w_n)$表示同义词集在第j个相似网页的出现概率，|表示条件概率；

[0152] 所述特征矩阵构建模块103，用于用于从每个高频网页中提取出所有的网页语段，构建每组网页语段与同义词集的特征矩阵：

[0153] $$\boldsymbol{N}_{1i}=\boldsymbol{N}_1\times\mathbf{N}_i^{\mathrm{T}}=\begin{pmatrix} w_1w_{i1} & \cdots & w_nw_{i1} \\ \cdots & & \cdots \\ w_1w_{im} & \cdots & w_nw_{im} \end{pmatrix}$$

[0154] $N_1=(w_1,w_2,\cdots,w_n)$

[0155] $N_i=(w_{i1},w_{i2},\cdots,w_{im})$

[0156] 其中，N_{1i}表示同义词集与第i组网页语段构建的特征矩阵，N_1为所述同义词集对应的词向量组，n为同义词集对应的词向量组的维度，N_i为第i组网页语段对应的词向量组，m为第i组网页语段的词向量组的维度；

[0157] 将所述特征矩阵按照词性对应原则执行拆分，得到多组词性矩阵，所述拆分包括：

[0158] 构建名词空矩阵、动词空矩阵、形容词空矩阵、数词空矩阵和代词空矩阵；

[0159] 遍历出所述特征矩阵中每个矩阵元素，其中每个矩阵元素均由同义词集的词向量和网页语段的词向量组成；

[0160] 剔除同义词集的词向量和网页语段的词向量的词性不同的矩阵元素后，判断每个矩阵元素所属于的词性类别；

[0161] 根据词性类别依次将矩阵元素填写至名词空矩阵、动词空矩阵、形容词空矩阵、数词空矩阵和代词空矩阵中的其中一个，直至遍历完所有的矩阵元素，得到名词矩阵、动词矩阵、形容词矩阵、数词矩阵和代词矩阵共五组词性矩阵；

[0162] 所述网页匹配模块104，用于依次计算每组词性矩阵内每个矩阵元素的相似度，并根据每个矩阵元素的相似度计算每组词性矩阵的离散程度值，设置每个词性矩阵的权重，相加每个词性矩阵的权重及离散程度值得到高频网页集中每个高频网页的总离散值，其中，总离散值的计算过程为：

[0163] $Q_r^{all}=\alpha_N Q_r^N+\alpha_V Q_r^V+\alpha_A Q_r^A+\alpha_M Q_r^M+\alpha_R Q_r^R$

[0164] 其中，Q_r^{all}表示总离散值，Q_r^N，Q_r^V，Q_r^A，Q_r^M，Q_r^R分别表示名词矩阵、动词矩阵、形容词矩阵、数词矩阵和代词矩阵的离散程度值，α_N，α_V，α_A，α_M，α_R分别表示名词矩阵、动词矩阵、形容词矩阵、数词矩阵和代词矩阵的权重，根据总离散值的大小对所述高频网页集执行排序，得到与所述搜索语段对应的搜索网页。

[0165] 详细地，本发明实施例中所述自然语言语义理解装置100中的所述各模块在使用

时采用与上述的图1中所述的基于区块链的产品供应链管理方法一样的技术手段，并能够产生相同的技术效果，这里不再赘述。

[0166] 如图3所示，是本发明一实施例提供的实现自然语言语义理解方法的电子设备的结构示意图。

[0167] 所述电子设备1可以包括处理器10、存储器11和总线12，还可以包括存储在所述存储器11中并可在所述处理器10上运行的计算机程序，如自然语言语义理解方法程序。

[0168] 其中，所述存储器11至少包括一种类型的可读存储介质，所述可读存储介质包括闪存、移动硬盘、多媒体卡、卡型存储器(例如:SD或DX存储器等)、磁性存储器、磁盘、光盘等。所述存储器11在一些实施例中可以是电子设备1的内部存储单元，例如该电子设备1的移动硬盘。所述存储器11在另一些实施例中也可以是电子设备1的外部存储设备，例如电子设备1上配备的插接式移动硬盘、智能存储卡(Smart Media Card,SMC)、安全数字(Secure Digital,SD)卡、闪存卡(Flash Card)等。进一步地，所述存储器11还可以既包括电子设备1的内部存储单元也包括外部存储设备。所述存储器11不仅可以用于存储安装于电子设备1的应用软件及各类数据，例如自然语言语义理解方法程序的代码等，还可以用于暂时地存储已经输出或者将要输出的数据。

[0169] 所述处理器10在一些实施例中可以由集成电路组成，例如可以由单个封装的集成电路所组成，也可以是由多个相同功能或不同功能封装的集成电路所组成，包括一个或者多个中央处理器(Central Processing unit,CPU)、微处理器、数字处理芯片、图形处理器及各种控制芯片的组合等。所述处理器10是所述电子设备的控制核心(Control Unit)，利用各种接口和线路连接整个电子设备的各个部件，通过运行或执行存储在所述存储器11内的程序或者模块(例如自然语言语义理解方法程序等)，以及调用存储在所述存储器11内的数据，以执行电子设备1的各种功能和处理数据。

[0170] 所述总线12可以是外设部件互连标准(peripheral component interconnect，简称PCI)总线或扩展工业标准结构(extended industry standard architecture，简称EISA)总线等。该总线12可以分为地址总线、数据总线、控制总线等。所述总线12被设置为实现所述存储器11以及至少一个处理器10等之间的连接通信。

[0171] 图3仅示出了具有部件的电子设备，本领域技术人员可以理解的是，图3示出的结构并不构成对所述电子设备1的限定，可以包括比图示更少或者更多的部件，或者组合某些部件，或者不同的部件布置。

[0172] 例如，尽管未示出，所述电子设备1还可以包括给各个部件供电的电源(比如电池)，优选地，电源可以通过电源管理装置与所述至少一个处理器10逻辑相连，从而通过电源管理装置实现充电管理、放电管理、以及功耗管理等功能。电源还可以包括一个或一个以上的直流或交流电源、再充电装置、电源故障检测电路、电源转换器或者逆变器、电源状态指示器等任意组件。所述电子设备1还可以包括多种传感器、蓝牙模块、Wi-Fi模块等，在此不再赘述。

[0173] 对于本领域技术人员而言，显然本发明不限于上述示范性实施例的细节，而且在不背离本发明的精神或基本特征的情况下，能够以其他的具体形式实现本发明。

[0174] 因此，无论从哪一点来看，均应将实施例看作是示范性的，而且是非限制性的，本发明的范围由所附权利要求而不是上述说明限定，因此旨在将落在权利要求的等同要件的

含义和范围内的所有变化涵括在本发明内。不应将权利要求中的任何附关联图标记视为限制所涉及的权利要求。

[0175] 此外，显然“包括”一词不排除其他单元或步骤，单数不排除复数。系统权利要求中陈述的多个单元或装置也可以由一个单元或装置通过软件或者硬件来实现。第二等词语用来表示名称，而并不表示任何特定的顺序。

[0176] 最后应说明的是，以上实施例仅用以说明本发明的技术方案而非限制，尽管参照较佳实施例对本发明进行了详细说明，本领域的普通技术人员应当理解，可以对本发明的技术方案进行修改或等同替换，而不脱离本发明技术方案的精神和范围。

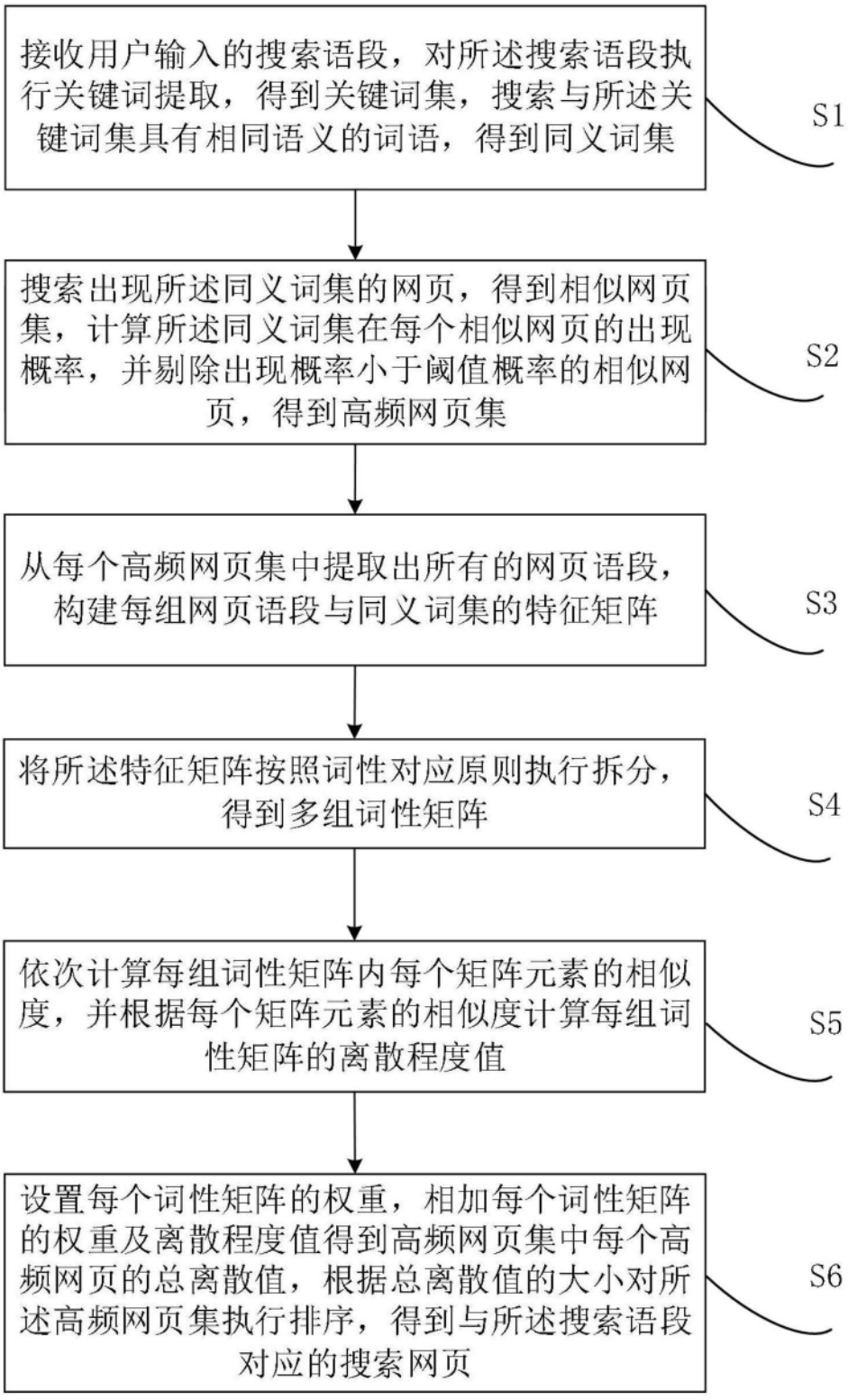

接收用户输入的搜索语段，对所述搜索语段执行关键词提取，得到关键词集，搜索与所述关键词集具有相同语义的词语，得到同义词集
S1
搜索出现所述同义词集的网页，得到相似网页集，计算所述同义词集在每个相似网页的出现概率，并剔除出现概率小于阈值概率的相似网页，得到高频网页集
S2
从每个高频网页集中提取出所有的网页语段，构建每组网页语段与同义词集的特征矩阵
S3
将所述特征矩阵按照词性对应原则执行拆分，得到多组词性矩阵
S4
依次计算每组词性矩阵内每个矩阵元素的相似度，并根据每个矩阵元素的相似度计算每组词性矩阵的离散程度值
S5
设置每个词性矩阵的权重，相加每个词性矩阵的权重及离散程度值得到高频网页集中每个高频网页的总离散值，根据总离散值的大小对所述高频网页集执行排序，得到与所述搜索语段对应的搜索网页
S6

图1

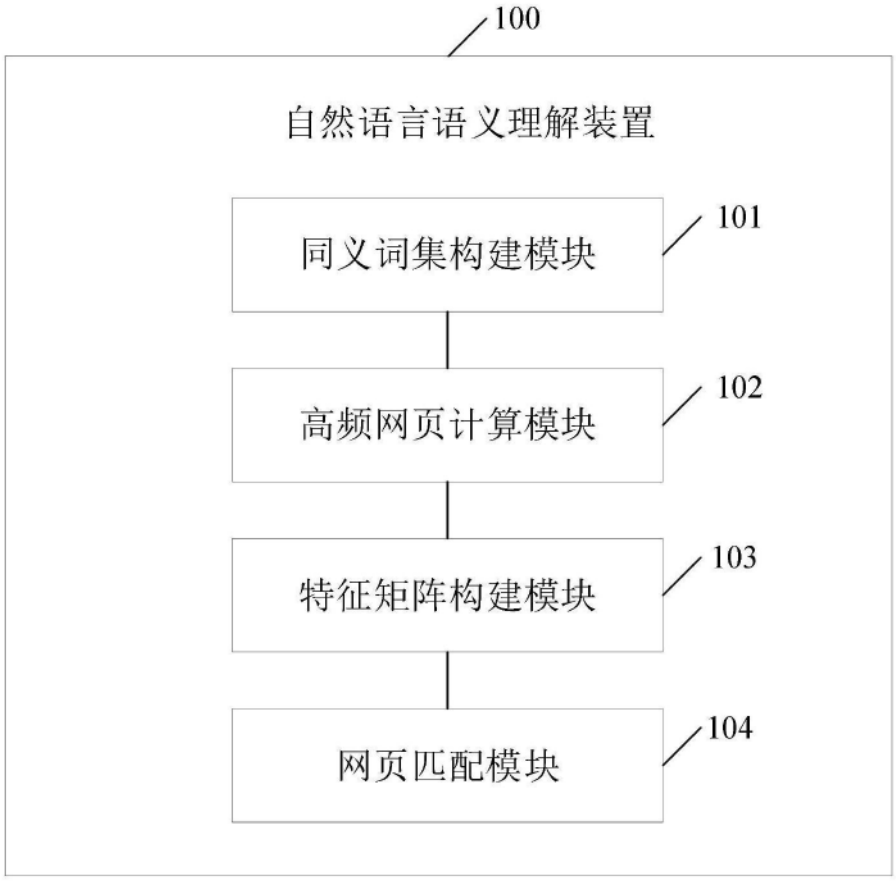

100
自然语言语义理解装置
同义词集构建模块
101
高频网页计算模块
102
特征矩阵构建模块
103
网页匹配模块
104

图2

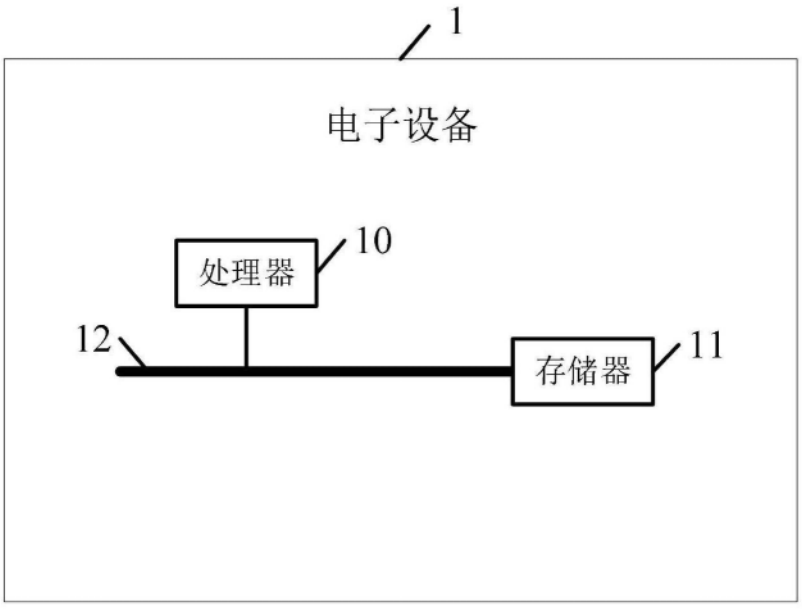

1
电子设备
处理器
10
12
存储器
11

图3