(19)国家知识产权局

(12)发明专利

(10)授权公告号 CN 117610543 B
(45)授权公告日 2024.08.02

(21)申请号 202311476751.6

(22)申请日 2023.11.08

(65)同一申请的已公布的文献号
申请公布号 CN 117610543 A

(43)申请公布日 2024.02.27

(73)专利权人 华南理工大学
地址 510640 广东省广州市天河区五山路381号
专利权人 人工智能与数字经济广东省实验室(广州)

(72)发明人 张通 邓忠易

(74)专利代理机构 广州市华学知识产权代理有限公司 44245
专利代理师 霍健兰

(51)Int.Cl.
G06F 40/216(2020.01)
G06F 40/151(2020.01)
G06F 16/901(2019.01)

(56)对比文件
CN 110020189 A,2019.07.16
CN 113435207 A,2021.09.24

审查员 吴姝泓

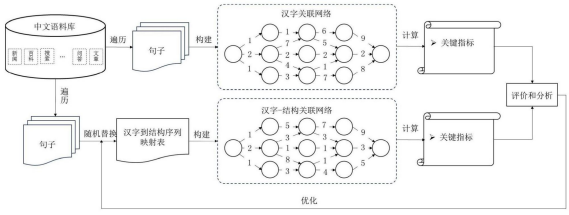
权利要求书3页 说明书10页 附图5页

(54)发明名称

基于图网络的汉字及结构的关联分析方法、介质及设备

(57)摘要

本发明提供了一种基于图网络的汉字及结构的关联分析方法、介质及设备;其中方法为:构建字频共现矩阵;遍历字频共现矩阵,构建以有向图为网络结构的汉字关联网络;构建汉字到结构序列的映射表;根据映射表中汉字到结构序列的映射关系,对中文语料库的每个句子采用结构序列随机替换部分汉字,在汉字关联网络中引入结构节点,得到汉字-结构关联网络;当汉字-结构关联网络的关键指标有提升则视为有效操作;输出每个汉字的权重以及汉字-结构关联网络。该方法更关注于字与字之间关联,利用不同汉字间存在相同结构的数据特点有效地缓解词性失衡问题对分析结果的干扰,有效提升中文语料库学习效率和学习准确率。



CN 117610543 B

1.一种基于图网络的汉字及结构的关联分析方法,其特征在于:包括如下步骤:

S1、遍历中文语料库,构建字频共现矩阵;遍历字频共现矩阵,以共现的汉字作为相邻节点,汉字的前后顺序作为边的指向,共现频率作为边的权重,构建以有向图为网络结构的汉字关联网络;

S2、根据数据处理流程从数据源获取汉字的结构序列,构建汉字到结构序列的映射表;

S3、根据映射表中汉字到结构序列的映射关系,对中文语料库的每个句子以设定概率采用结构序列随机替换部分汉字,在汉字关联网络中引入结构节点,得到汉字-结构关联网络;

S4、分别计算汉字关联网络和汉字-结构关联网络的关键指标,并进行比较:

如果汉字-结构关联网络的关键指标较汉字关联网络提升,则步骤S3的操作视为有效操作,执行步骤S5;

如果汉字-结构关联网络的关键指标较汉字关联网络没有提升,或者没有达到满意的程度,将分析和对比结果作为反馈信号,优化步骤S3的随机替换流程,并重复步骤S4;

S5、输出当前中文语料库中每个汉字的权重以及当前中文语料库的汉字-结构关联网络;

利用每个汉字的权重以及汉字-结构关联网络进行关联分析。

2.根据权利要求1所述的基于图网络的汉字及结构的关联分析方法,其特征在于:所述步骤S1,是指:

建立空的字频共现矩阵;

遍历中文语料库的每个句子,遍历每个句子的每个汉字,构建字频共现矩阵;所述字频共现矩阵的构建方法是:遍历每个句子,将句子中的每个汉字和下一个汉字拼接作为字频共现矩阵的键k;若字频共现矩阵已存在键k,则键k值累加1;否则将键k存入字频共现矩阵,键k值设为1;当中文语料库的所有句子均遍历后,完成字频共现矩阵的构建;

遍历字频共现矩阵,将各个键k分别重新拆分为两个汉字,作为相邻节点;汉字的前后顺序作为边的指向,键k值作为边的权重,构建以有向图为网络结构的汉字关联网络。

3.根据权利要求1所述的基于图网络的汉字及结构的关联分析方法,其特征在于:所述步骤S2,是指:

读取Unicode标准区中所有收录汉字的象形结构序列;构建汉字到结构序列的映射表;映射表中,键为汉字,值初步设定为对应的象形结构序列;

遍历映射表中的各个映射对,判断各个映射对的键和值是否相等:

若当前映射对的键和值相等,则对象形结构序列进行去重操作;

否则标记当前键为特殊汉字,读取特殊汉字的笔画分解序列,并进行去重操作;计算序列中无意义笔画的占比p:当占比p＜占比阈值时,将笔画分解序列替换原映射表中当前键的键值,当占比p≥占比阈值时,保持原映射表中当前键的键值;

当映射表中的各个映射对均遍历后,得到汉字到结构序列的映射表。

4.根据权利要求1所述的基于图网络的汉字及结构的关联分析方法,其特征在于:所述步骤S3,包括如下步骤:

S31、设置替换百分比r的值和概率p的值;

S32、遍历当前中文语料库的每个句子;

设置当前句子替换汉字的个数上限m,m＝句子长度×r;从当前句子中随机选取m个汉字作为替换候选集合;

遍历当前句子替换候选集合的每个汉字,在区间[0,1]随机选择一个小数x;若x<概率p,把当前汉字加入到替换候选集合;否则无操作;

在当前句子替换候选集合遍历结束后,跳至下一句子,重复执行步骤S32,直至当前中文语料库的每个句子遍历结束;

S33、遍历替换候选集合的每个汉字,从汉字到结构序列的映射表中读取当前汉字对应的结构序列;将汉字与对应结构序列中的每个结构作为相邻节点,从汉字到结构作为“汉字-结构”边的指向,将汉字所有边的权重取均值作为“汉字-结构”边的权重,将结构节点加入到汉字关联网络中,形成汉字-结构关联网络。

5.根据权利要求1所述的基于图网络的汉字及结构的关联分析方法,其特征在于:所述步骤S4中关键指标,是指:平均最短路径长度、半径与直径、强连通分量中的任一项或两项以上;

所述平均最短路径长度的计算方法是:

对于汉字关联网络和汉字-结构关联网络,将两两汉字节点之间的最短路径长度进行算术平均,得到平均最短路径长度:

$$S = \frac{\sum_{i,j} d_{i,j}}{n} \quad \{i,j\} \in G, i \neq j$$

其中,G表示汉字关联网络/汉字-结构关联网络;n表示汉字关联网络/汉字-结构关联网络中汉字节点的个数;i、j表示汉字关联网络/汉字-结构关联网络中互不相同的任意两个汉字节点;$d_{i,j}$表示汉字节点i和j之间的最短路径;

所述半径r与直径d分别是汉字关联网络/汉字-结构关联网络中最小和最大的节点偏心率E_p;

$r＝Min(E_p)$,p∈G

$d＝Max(E_p)$,p∈G

所述强连通分量的计算方法是:

假设sg是汉字关联网络/汉字-结构关联网络G的一个子图,其中sg的每一个节点都能通过某条路径到达另一个节点,则sg具备强连通性;如果此时从G中其余的节点任取一个加入sg,都会破坏sg的强连通性,则sg为G的一个强连通分量。

6.根据权利要求1所述的基于图网络的汉字及结构的关联分析方法,其特征在于:所述步骤S5中,每个汉字的权重是指:汉字关联网络或汉字-结构关联网络中,节点的偏心率、节点的近性中心度和节点的介性中心度中的任一项;

节点的偏心率E_p的计算方法是:

对于汉字关联网络或汉字-结构关联网络中的任意一个汉字节点p,计算汉字节点p与其他汉字节点i的最短路径长度,将最大的距离作为汉字节点p的偏心率:

$E_p＝Max(d_{p,i})$,{i,p}∈G,i≠p

其中,G表示汉字关联网络/汉字-结构关联网络;$d_{p,i}$表示汉字节点p和i之间的最短路径;

节点的近性中心度C_p的计算方法是:

$$C_p = \frac{1}{d_p} = \frac{n-1}{\sum_i d_{p,i}}$$

其中,n表示汉字关联网络/汉字-结构关联网络中的汉字节点数量;

节点的介性中心度的计算方法是:

$$B_p = \sum_{i,j} \frac{d_{i,j}^p}{d_{i,j}}, \qquad \{i,j,p\} \in G, i \neq j \neq p$$

其中,p、i、j表示互不相同的任意一个汉字节点;$d_{i,j}^p$表示汉字节点i和j之间经过汉字节点p的最短路径。

7.一种可读存储介质,其特征在于,其中所述存储介质存储有计算机程序,所述计算机程序当被处理器执行时使所述处理器执行权利要求1-6中任一项所述的基于图网络的汉字及结构的关联分析方法。

8.一种计算机设备,包括处理器以及用于存储处理器可执行程序的存储器,其特征在于,所述处理器执行存储器存储的程序时,实现权利要求1-6中任一项所述的基于图网络的汉字及结构的关联分析方法。

基于图网络的汉字及结构的关联分析方法、介质及设备

技术领域

[0001] 本发明涉及现代汉语分析技术领域,更具体地说,涉及一种基于图网络的汉字及结构的关联分析方法、介质及设备。

背景技术

[0002] 语法分析是自然语言处理中一个比较重要的任务,其目标是分析句子的语法结构,反映一个完整句子中的字词关联。由于自然语言分布特性,不同领域的语料库在知识侧重方面也大有不同,例如,在开放域的语料库中,虚词的出现频率明显高于实词;在垂直域的语料库中,专业名词的出现频率明显高于非专业名词;相同的汉字在不同的语境下也蕴含截然相反的意思。简单地通过汉字出现的频率来衡量其重要程度明显缺乏合理性。在汉语言研究领域的现有工作中,尚未发现一种科学的分析方法能够在不同类型的中文语料库中因地制宜的准确量化每个汉字的贡献度。

[0003] 现代汉语的常规分析方法通常采用语法树对完整语句进行结构分析。语法树就是按照某一语言规则进行推导时所形成的能够反映句子语法结构的图形表示。一棵语法树包括了一个句型的所有可能的推导过程。语法树的构建和设计以人工经验为主导,蕴含了丰富且专业的语言学知识。

[0004] 但是以人工经验为主导的语法树分析方式在数据层面上局限于词粒度。中文语句转化成语法树,通常包括名词、动词、形容词、数词、量词、代词这6种实词,以及副词、介词、连词、助词、拟声词、叹词这6种虚词,汉字在语法树中大部分以词组的形式出现,小部分以单个字的形式出现。因此,语法树很大程度上反映的是字与词、词与词之间的关联,忽略了字与字之间的关联,没有深入到字粒度的层面进行分析。

[0005] 自然语言处理领域通常采用序列标注的方式挖掘句子中的字词关联。序列标注是一个序列到序列的过程,即给定一个输入序列,使用模型对该序列的每一个位置标注相应的标签,可以认为是分类问题的推广。分词、词性标注、命名实体识别等都属于序列标注的范畴。词性标注是指将系统接收到的人类语言转换为序列数据,根据语法结构进行切分,同时确定分词结果中每个词的词性,如名词、动词、形容词、介词等;命名实体识别是指通过捕捉汉字间的依赖关系识别词组边界,从语句中发现特定的实体,如人命、地名及特定领域的专属名词等。序列标注任务的标签粒度需要细分到字符级别,即对语料库中每个句子的每个字进行标注。

[0006] 但是,以深度学习为主导的序列标注方法容易受到语料库词性失衡的影响。深度学习常用的梯度下降学习方法对所有训练数据一视同仁,在训练过程中出现频率越高的数据越容易被模型记住,因此语料库中的高频字词会得到更加充分的学习,容易导致过拟合的现象,影响模型的准确率。现有的序列标注建模方案中,针对词性失衡问题的处理方式通常分为数据层面和模型层面两种。数据层面的优化是指通过对低频词过采样或对高频词降采样,使得语料库的不同词性的词频变得均衡,但这种方法的缺陷是人为改变了自然语言的分布规律。模型层面的优化是指通过对多分类损失函数设置不同权重,使低频词类别在

梯度下降的训练过程中比高频词占据更大的梯度值,间接弱化高频词对模型的影响。这种方法的不足是引入了如何分配损失函数权重的新问题。

发明内容

[0007] 为克服现有技术中的缺点与不足,本发明的目的在于提供一种基于图网络的汉字及结构的关联分析方法、介质及设备;该方法更关注于字与字之间关联,利用不同汉字间存在相同结构的数据特点有效地缓解词性失衡问题对分析结果的干扰,有效提升中文语料库学习效率和学习准确率。

[0008] 为了达到上述目的,本发明通过下述技术方案予以实现:一种基于图网络的汉字及结构的关联分析方法,包括如下步骤:

[0009] S1、遍历中文语料库,构建字频共现矩阵;遍历字频共现矩阵,以共现的汉字作为相邻节点,汉字的前后顺序作为边的指向,共现频率作为边的权重,构建以有向图为网络结构的汉字关联网络;

[0010] S2、根据数据处理流程从数据源获取汉字的结构序列,构建汉字到结构序列的映射表;

[0011] S3、根据映射表中汉字到结构序列的映射关系,对中文语料库的每个句子以设定概率采用结构序列随机替换部分汉字,在汉字关联网络中引入结构节点,得到汉字-结构关联网络;

[0012] S4、分别计算汉字关联网络和汉字-结构关联网络的关键指标,并进行比较:

[0013] 如果汉字-结构关联网络的关键指标较汉字关联网络提升,则步骤S3的操作视为有效操作,执行步骤S5;

[0014] 如果汉字-结构关联网络的关键指标较汉字关联网络没有提升,或者没有达到满意的程度,将分析和对比结果作为反馈信号,优化步骤S3的随机替换流程,并重复步骤S4;

[0015] S5、计算当前中文语料库中每个汉字的权重,以及输出当前中文语料库的汉字-结构关联网络;

[0016] 利用每个汉字的权重以及汉字-结构关联网络进行关联分析。

[0017] 优选地,所述步骤S1,是指:

[0018] 建立空的字频共现矩阵;

[0019] 遍历中文语料库的每个句子,遍历每个句子的每个汉字,构建字频共现矩阵;所述字频共现矩阵的构建方法是:遍历每个句子,将句子中的每个汉字和下一个汉字拼接作为字频共现矩阵的键k;若字频共现矩阵已存在键k,则键k值累加1;否则将键k存入字频共现矩阵,键k值设为1;当中文语料库的所有句子均遍历后,完成字频共现矩阵的构建;

[0020] 遍历字频共现矩阵,将各个键k分别重新拆分为两个汉字,作为相邻节点;汉字的前后顺序作为边的指向,键k值作为边的权重,构建以有向图为网络结构的汉字关联网络。

[0021] 优选地,所述步骤S2,是指:

[0022] 读取Unicode标准区中所有收录汉字的象形结构序列;构建汉字到结构序列的映射表;映射表中,键为汉字,值初步设定为对应的象形结构序列;

[0023] 遍历映射表中的各个映射对,判断各个映射对的键和值是否相等:

[0024] 若当前映射对的键和值相等,则对象形结构序列进行去重操作;

[0025] 否则标记当前键为特殊汉字,读取特殊汉字的笔画分解序列,并进行去重操作;计算序列中无意义笔画的占比p:当占比p＜占比阈值时,将笔画分解序列替换原映射表中当前键的键值,当占比p≥占比阈值时,保持原映射表中当前键的键值;

[0026] 当映射表中的各个映射对均遍历后,得到汉字到结构序列的映射表。

[0027] 优选地,所述步骤S3,包括如下步骤:

[0028] S31、设置替换百分比r的值和概率p的值;

[0029] S32、遍历当前中文语料库的每个句子;

[0030] 设置当前句子替换汉字的个数上限m,m＝句子长度×r;从当前句子中随机选取m个汉字作为替换候选集合;

[0031] 遍历当前句子替换候选集合的每个汉字,在区间[0,1]随机选择一个小数x;若x＜概率p,把当前汉字加入到替换候选集合;否则无操作;

[0032] 在当前句子替换候选集合遍历结束后,跳至下一句子,重复执行步骤S32,直至当前中文语料库的每个句子遍历结束;

[0033] S33、遍历替换候选集合的每个汉字,从汉字到结构序列的映射表中读取当前汉字对应的结构序列;将汉字与对应结构序列中的每个结构作为相邻节点,从汉字到结构作为“汉字-结构”边的指向,将汉字所有边的权重取均值作为“汉字-结构”边的权重,将结构节点加入到汉字关联网络中,形成汉字-结构关联网络。

[0034] 优选地,所述步骤S4中关键指标,是指:平均最短路径长度、半径与直径、强连通分量中的任一项或两项以上;

[0035] 所述平均最短路径长度的计算方法是:

[0036] 对于汉字关联网络和汉字-结构关联网络,将两两汉字节点之间的最短路径长度进行算术平均,得到平均最短路径长度:

[0037] $$S = \frac{\sum_{i,j} d_{i,j}}{n} \quad \{i,j\} \in G, i \neq j$$

[0038] 其中,G表示汉字关联网络/汉字-结构关联网络;n表示汉字关联网络/汉字-结构关联网络中汉字节点的个数;i、j表示汉字关联网络/汉字-结构关联网络中互不相同的任意两个汉字节点;$d_{i,j}$表示汉字节点i和j之间的最短路径;

[0039] 所述半径r与直径d分别是汉字关联网络/汉字-结构关联网络中最小和最大的节点偏心率E_p;

[0040] $r＝Min(E_p)$,p∈G

[0041] $d＝Max(E_p)$,p∈G

[0042] 所述强连通分量的计算方法是:

[0043] 假设sg是汉字关联网络/汉字-结构关联网络G的一个子图,其中sg的每一个节点都能通过某条路径到达另一个节点,则sg具备强连通性;如果此时从G中其余的节点任取一个加入sg,都会破坏sg的强连通性,则sg为G的一个强连通分量。

[0044] 优选地,所述步骤S5中,每个汉字的权重是指:汉字关联网络或汉字-结构关联网络中,节点的偏心率、节点的近性中心度和节点的介性中心度中的任一项;

[0045] 节点的偏心率E_p的计算方法是:

[0046] 对于汉字关联网络或汉字-结构关联网络中的任意一个汉字节点p,计算汉字节点

p与其他汉字节点i的最短路径长度,将最大的距离作为汉字节点p的偏心率:

[0047] $E_p=Max(d_{p,i}),\{i,p\}\in G,i\neq p$

[0048] 其中,G表示汉字关联网络/汉字-结构关联网络;$d_{p,i}$表示汉字节点p和i之间的最短路径;

[0049] 节点的近性中心度C_p的计算方法是:

[0050] $$C_p=\frac{1}{d_p}=\frac{n-1}{\sum_i d_{p,i}}$$

[0051] 其中,n表示汉字关联网络/汉字-结构关联网络中的汉字节点数量;

[0052] 节点的介性中心度的计算方法是:

[0053] $$B_p=\sum_{i,j}\frac{d_{i,j}^p}{d_{i,j}},\qquad \{i,j,p\}\in G,i\neq j\neq p$$

[0054] 其中,p、i、j表示互不相同的任意一个汉字节点;$d_{i,j}^p$表示汉字节点i和j之间经过汉字节点p的最短路径。

[0055] 一种可读存储介质,其中所述存储介质存储有计算机程序,所述计算机程序当被处理器执行时使所述处理器执行所述基于图网络的汉字及结构的关联分析方法。

[0056] 一种计算机设备,包括处理器以及用于存储处理器可执行程序的存储器,所述处理器执行存储器存储的程序时,实现所述基于图网络的汉字及结构的关联分析方法。

[0057] 与现有技术相比,本发明具有如下优点与有益效果:

[0058] 1、本发明提出了一种基于有向图网络的汉字关联分析方法,通过统计中文语料库的字频共现矩阵获得节点、有向边和权重等构建汉字关联网络所需的要素,并通过图的关键指标评价关联网络的质量和量化汉字的重要程度;相比于语法树的分析方法,基于图网络的分析方法更关注于字与字之间关联,通过有向图网络的多项关键指标精准地量化汉字的贡献度和重要程度,以及衡量当前语料库的质量,增强了中文自然语言处理研究的可解释性;这一套标准的汉字关联网络构建流程和评价方法广泛适用于不同领域、不同类型的中文语料库;

[0059] 2、本发明提出了一种汉字到结构序列的数据处理方法,根据汉字结构特点选择从不同的数据源获取汉字的结构序列,同时结合一定的人工规则对结构序列进行优化;如何对汉字进行结构拆分目前在学术界尚未有统一的标准,本发明提出的汉字到结构序列的数据处理方法,将汉字拆分粒度的选择、结构序列对深度学习模型的适配等个性化的需求整合到统一的流程中,在结合语言学专业知识的基础上,可以吸纳更多合理且有效的人工经验,有效处理不规则的拆分个例,不断地优化汉字到结构序列的映射关系;

[0060] 3、本发明提出了一种基于结构序列的汉字关联网络优化方法,一定概率地使用结构序列取代汉字在语句中的出现,在汉字关联网络中引入汉字与结构序列的关联关系,并通过图网络的多项关键指标变化来判断这种取代方式是否有效,同时为概率值、替换百分比等参数的选取提供指导信息;引入结构序列的汉字关联网络,打破了语法树分析方法在数据层面止步于词粒度的局限,从更细微的结构层次深入探索了汉字间的关联;在有向图网络的关键指标得到提升的前提下,相比于序列标注等以深度学习为核心的方法,不同汉

字间存在相同结构的数据特点能够有效地缓解词性失衡问题对分析结果的干扰;这种汉字结构属性带来的优势同样广泛适用于不同领域、不同类型的中文语料库。

附图说明

[0061] 图1是本发明基于图网络的汉字及结构的关联分析方法的示意图;

[0062] 图2是本发明基于图网络的汉字及结构的关联分析方法的总体工作流程图;

[0063] 图3是本发明基于图网络的汉字及结构的关联分析方法中,汉字关联网络示意图;

[0064] 图4是本发明基于图网络的汉字及结构的关联分析方法中,汉字关联网络的构建流程图;

[0065] 图5是本发明基于图网络的汉字及结构的关联分析方法中,汉字-结构关联网络示意图;

[0066] 图6是本发明基于图网络的汉字及结构的关联分析方法中,数据处理流程图;

[0067] 图7是本发明基于图网络的汉字及结构的关联分析方法中,汉字-结构关联网络的构建流程图。

具体实施方式

[0068] 下面结合附图与具体实施方式对本发明作进一步详细的描述。

[0069] 实施例一

[0070] 本专利提出的基于图网络的汉字及其结构的关联分析方法以汉字作为研究对象,通过统计具体中文语料库的字频共现矩阵构建汉字及其结构的关联网络,并使用关联网络的分析结果量化汉字在语料库中的贡献度。该方法的核心是关联网络的构建和分析两个阶段,其中关联网络采用有向图作为网络结构,构建有向图的节点、边的方向和权重等要素均从语料库的字频共现矩阵中获取;分析阶段则选取有向图的平均最短路径长度、有向图的强连通分量、有向图的半径与直径、点的偏心率、点的近性中心度和点的介性中心度等能够准确反映图网络质量的指标进行计算和分析。在构建和分析的基础上,采取结构序列随机替换汉字的方式,在汉字关联网络中引入结构节点,并根据指标分析结果优化关联网络的构建过程。在汉字-结构关联网络构建完成的前提下,各项关键指标的提升证明了不同汉字间存在相同结构这一汉字构造属性能够有效缓解语料库中的词性失衡问题。

[0071] 本实施例一种基于图网络的汉字及结构的关联分析方法,如图1和图2所示,包括如下步骤:

[0072] S1、遍历中文语料库,构建字频共现矩阵;遍历字频共现矩阵,以共现的汉字作为相邻节点,汉字的前后顺序作为边的指向,共现频率作为边的权重,构建以有向图为网络结构的汉字关联网络。

[0073] 汉字关联网络以有向图作为网络结构,图网络的节点、边的方向以及权重等要素需要从中文语料库的字频共现矩阵获取。字频共现矩阵是一种反映汉字共同出现频率的数学形式。将当前语料库中的所有汉字映射成数字id,作为字频共现矩阵的横纵坐标轴,根据语句中共同出现的两个汉字的id定位到矩阵中的具体坐标点,两个汉字共同出现的频率作为该点的值。遍历当前中文语料库得到的统计结果就组成了一个完整的字频共现矩阵。在字频共现矩阵的基础上,以共现的汉字作为有向图的相邻节点,汉字在语句中的顺序作为

边的指向,共现频率作为边的权重,构建该语料库的汉字关联网络。不同类型、不同领域的中文语料库具有不同的字频共现矩阵,对应的有向图也是不一样的。汉字关联网络构建阶段如图3所示。

[0074] 由于中文语料库一般具有较大的规模和丰富的内容,统计得到的字频共现矩阵在维度上可能达到数百万甚至上千万,对计算机内存的压力很大,同时直接导致有向图节点和边的数量指数级增加,带来无法接受的计算复杂度和时间消耗。因此,在编程实现过程中,本专利采用字典的数据结构实现字频共现矩阵的动态构建,只统计真实存在的共现汉字。

[0075] 具体构建流程如图4所示,步骤S1具体是指:

[0076] 建立空的字频共现矩阵;

[0077] 遍历中文语料库的每个句子,遍历每个句子的每个汉字,构建字频共现矩阵;所述字频共现矩阵的构建方法是:遍历每个句子,将句子中的每个汉字和下一个汉字拼接作为字频共现矩阵的键k;若字频共现矩阵已存在键k,则键k值累加1;否则将键k存入字频共现矩阵,键k值设为1;当中文语料库的所有句子均遍历后,完成字频共现矩阵的构建;

[0078] 遍历字频共现矩阵,将各个键k分别重新拆分为两个汉字,作为相邻节点;汉字的前后顺序作为边的指向,键k值作为边的权重,构建以有向图为网络结构的汉字关联网络。

[0079] S2、根据数据处理流程从数据源获取汉字的结构序列,构建汉字到结构序列的映射表。

[0080] S3、根据映射表中汉字到结构序列的映射关系,对中文语料库的每个句子以设定概率采用结构序列随机替换部分汉字,在汉字关联网络中引入结构节点,得到汉字-结构关联网络。

[0081] 本专利提出一种通过汉字结构缓解中文语料库词性失衡问题的方法,首先需要构建汉字到结构序列的映射关系,然后在遍历语料库构建字频共现矩阵的过程中,随机将每个句子的部分汉字替换成对应的结构序列,最后在汉字关联网络中引入结构节点和汉字到结构的关联,实现网络各项指标的优化。方法如图5所示。

[0082] 构建汉字到结构序列的映射关系需要考虑实际需求,例如中文语料经过本专利提出的方法处理后,可能会用于深度学习、数据挖掘等研究,因此对汉字拆分到什么程度、拆分后的结构序列是否进行额外处理等实际情况的处理要有所依据。在本方法中,汉字的结构序列从不同的数据源获取,一个是象形结构数据源,一个是笔画数据源。象形结构数据源在对汉字进行拆分时,遵循尽可能保留大结构的原则,笔画数据源则更倾向于拆分为简单的笔画。显然,相对于象形结构序列,笔画序列可能存在大量无意义的笔画结构,因此,从汉字的发展和知识出发,在拆分过程中应该首选将汉字拆分为象形结构。本方法针对深度学习场景的需求,制定一系列数据处理流程,首先读取当前汉字的象形结构序列,根据读取结果判断当前汉字是否可以拆分,如果无法拆分为象形结构序列,则继续读取其笔画序列,并设置对笔画序列的质量验证要求,如果通过满足要求,则将笔画序列作为该汉字的结构序列;最后对结构序列进行去重操作。

[0083] 数据处理流程如图6所示,步骤S2具体是指:

[0084] 读取Unicode标准区中所有收录汉字的象形结构序列;构建汉字到结构序列的映射表;映射表中,键为汉字,值初步设定为对应的象形结构序列;

[0085] 遍历映射表中的各个映射对,判断各个映射对的键和值是否相等:

[0086] 若当前映射对的键和值相等,则对象形结构序列进行去重操作;

[0087] 否则标记当前键为特殊汉字,读取特殊汉字的笔画分解序列,并进行去重操作;计算序列中无意义笔画的占比p:当占比p<占比阈值时,将笔画分解序列替换原映射表中当前键的键值,当占比p≥占比阈值时,保持原映射表中当前键的键值;

[0088] 当映射表中的各个映射对均遍历后,得到汉字到结构序列的映射表。

[0089] 汉字-结构关联网络同样采用有向图作为网络结构,构建过程与汉字关联网络类似,区别在于通过对句子中的汉字随机替换成结构序列的方式引入结构节点。假设关联网络中两个相邻汉字AB,边的方向是A指向B,加入A的结构节点后,增加两条边,方向是A指向结构节点和结构节点指向B。汉字-结构关联网络的核心在于随机替换过程,该过程涉及两个关键参数:替换百分比r和替换概率p。替换百分比r决定一个句子中等待替换的汉字个数,计算方式为r乘上句子长度。替换概率p决定一个汉字被替换成结构序列的概率,计算方式为从[0,1]区间选取一个随机数,若大于p则对该汉字执行替换操作。因此,一个句子中最终有哪些汉字会被替换,同时取决于r和p。虽然这两个关键参数的初始值通常是基于人工经验赋值,但是在本专利提出的汉字-结构关联网络构建流程中,可以将关联网络的指标变化作为反馈,指导关键参数的更新以构建更优的关联网络,具体流程如图7所示。

[0090] 步骤S3包括如下步骤:

[0091] S31、设置替换百分比r的值和概率p的值;

[0092] S32、遍历当前中文语料库的每个句子;

[0093] 设置当前句子替换汉字的个数上限m,m=句子长度×r;从当前句子中随机选取m个汉字作为替换候选集合;

[0094] 遍历当前句子替换候选集合的每个汉字,在区间[0,1]随机选择一个小数x;若x<概率p,把当前汉字加入到替换候选集合;否则无操作;

[0095] 在当前句子替换候选集合遍历结束后,跳至下一句子,重复执行步骤S32,直至当前中文语料库的每个句子遍历结束;

[0096] S33、遍历替换候选集合的每个汉字,从汉字到结构序列的映射表中读取当前汉字对应的结构序列;将汉字与对应结构序列中的每个结构作为相邻节点,从汉字到结构作为“汉字-结构”边的指向,将汉字所有边的权重取均值作为“汉字-结构”边的权重,将结构节点加入到汉字关联网络中,形成汉字-结构关联网络。

[0097] S4、分别计算汉字关联网络和汉字-结构关联网络的关键指标,并进行比较:

[0098] 如果汉字-结构关联网络的关键指标较汉字关联网络提升,则步骤S3的操作视为有效操作,执行步骤S5;

[0099] 如果汉字-结构关联网络的关键指标较汉字关联网络没有提升,或者没有达到满意的程度,将分析和对比结果作为反馈信号,优化步骤S3的随机替换流程,并重复步骤S4。

[0100] 关键指标是指:平均最短路径长度、半径与直径、强连通分量中的任一项或两项以上;

[0101] 所述平均最短路径长度的计算方法是:

[0102] 对于汉字关联网络和汉字-结构关联网络,将两两汉字节点之间的最短路径长度进行算术平均,得到平均最短路径长度,可以用来衡量关联网络中汉字节点之间的平均距

离:

[0103] $$S = \frac{\sum_{i,j} d_{i,j}}{n} \quad \{i,j\} \in G, i \neq j$$

[0104] 其中,G表示汉字关联网络/汉字-结构关联网络;n表示汉字关联网络/汉字-结构关联网络中汉字节点的个数;i、j表示汉字关联网络/汉字-结构关联网络中互不相同的任意两个汉字节点;$d_{i,j}$表示汉字节点i和j之间的最短路径;

[0105] 所述半径r与直径d分别是汉字关联网络/汉字-结构关联网络中最小和最大的节点偏心率E_p;

[0106] $r=Min(E_p), p \in G$

[0107] $d=Max(E_p), p \in G$

[0108] 所述强连通分量的计算方法是:

[0109] 假设sg是汉字关联网络/汉字-结构关联网络G的一个子图,其中sg的每一个节点都能通过某条路径到达另一个节点,则sg具备强连通性;如果此时从G中其余的节点任取一个加入sg,都会破坏sg的强连通性,则sg为G的一个强连通分量。利用强连通分量可以把关联网络中强连通的子网络缩为一个点,减少图操作的计算量。

[0110] 除了上述提到的关键指标外,还可以根据实际需要继续增加其他的图网络指标,如点度中心性、特征向量中心性、pagerank值等,从不同的角度对汉字及其结构关联网络做出评价。

[0111] S5、计算当前中文语料库中每个汉字的权重,以及输出当前中文语料库的汉字-结构关联网络。

[0112] 每个汉字的权重是指:汉字关联网络或汉字-结构关联网络中,节点的偏心率、节点的近性中心度和节点的介性中心度中的任一项;

[0113] 节点的偏心率E_p的计算方法是:

[0114] 对于汉字关联网络或汉字-结构关联网络中的任意一个汉字节点p,计算汉字节点p与其他汉字节点i的最短路径长度,将最大的距离作为汉字节点p的偏心率:

[0115] $E_p=Max(d_{p,i}), \{i,p\} \in G, i \neq p$

[0116] 其中,G表示汉字关联网络/汉字-结构关联网络;$d_{p,i}$表示汉字节点p和i之间的最短路径;

[0117] 节点的近性中心度C_p的计算方法是:

[0118] $$C_p = \frac{1}{d_p} = \frac{n-1}{\sum_i d_{p,i}}$$

[0119] 其中,n表示汉字关联网络/汉字-结构关联网络中的汉字节点数量;

[0120] 近性中心度衡量的是某个汉字节点和他汉字节点的联通能力,数学上定义为给定汉字节点到其他所有汉字节点的平均最短距离的倒数。一个汉字节点的近性中心度较高,说明该汉字到网络中其他汉字节点的最短距离都很小,更靠近关联网络的几何中心。

[0121] 节点的介性中心度的计算方法是:

[0122] $$B_p = \sum_{i,j} \frac{d_{i,j}^p}{d_{i,j}}, \qquad \{i,j,p\} \in G, i \neq j \neq p$$

[0123] 其中,p、i、j表示互不相同的任意一个汉字节点;$d_{i,j}^p$表示汉字节点i和j之间经过汉字节点p的最短路径。

[0124] 介性中心度衡量的是某个汉字节点的相邻节点之间的联通程度,数学上定义为经过给定汉字节点的最短路径数和经过关联网络中其他汉字节点的最短路径总数之比。一个汉字节点的介性中心度较高,说明该汉字对相邻汉字之间的联通具备很强的影响力。

[0125] 利用每个汉字的权重以及汉字-结构关联网络进行关联分析。

[0126] 本发明提出了为一种汉字到结构序列的数据处理方法,对一般汉字和疑难汉字的结构拆分依据兼顾了语言学专业知识和人工经验,同时将汉字拆分粒度的选择、结构序列对深度学习模型的适配等个性化的需求整合到统一的流程中。

[0127] 本发明提出了一种基于有向图网络的汉字关联分析方法,包括汉字关联网络的构建和分析两个部分。其中,汉字关联网络的构建需要统计语料库的字频共现矩阵,以共现的汉字作为节点,汉字在语句中的顺序作为边的指向,共现频率作为边的权重。汉字关联网络的分析通过计算有向图的平均最短路径长度、强连通分量、半径与直径以及点的偏心率、中心度等指标量化每个汉字在关联网络中的重要程度和衡量关联网络的整体质量,关联网络的质量同样反映了当前语料库的质量。两部分组合成一套广泛适用于不同领域、不同类型中文语料库的标准构建流程和评价方法。

[0128] 本发明提出了一种基于结构序列的汉字关联网络优化方法,一定概率地使用结构序列取代汉字在语句中的出现,在汉字关联网络中引入结构序列,从更细微的结构粒度对汉字间的关联进行分析,并利用网络关键指标的前后变化为概率值的选取提供指导信息。在关键指标得到提升的前提下,将结构序列的引入视为有效,利用不同汉字间存在相同结构的数据特点有效地缓解词性失衡问题对分析结果的干扰。这种汉字结构属性带来的优势同样广泛适用于不同领域、不同类型的中文语料库。

[0129] 本发明可应用于NLP大模型的中文预训练等场景中。NLP大模型预训练的核心任务是屏蔽词任务,即输入一个语句,随机屏蔽掉语句中的单个汉字或者词组,通过模型预测被屏蔽的汉字。在该任务的训练过程中,大模型对每个屏蔽字都赋予同等权重,然而,在中文的实际应用中,不同汉字的重要程度是有区别的,例如,在一般场景中,介词、虚词、助词的重要程度通常小于名词,在专业场景中,专业名词的重要程度通常大于普通名词,这种情况显然不适合对每个汉字赋予相同权重,因此,需要一种能够针对不同中文语料库特点来准确衡量汉字重要程度的科学方法。

[0130] 本专利提出的基于图网络的汉字及结构的关联分析方法,其研究对象可以是各种类型的中文语料库(文本集合),如百度百科、新浪新闻、大众点评、中医药文献……等等。通过构建汉字及其结构的有向图网络等方式,计算当前语料库中每个汉字的权重,该权重即为汉字在当前语料库的重要程度。因此,本发明能够科学计算出不同领域、不同类型的中文语料库中的汉字权重,并将其应用于NLP大模型的屏蔽词预训练任务,让大模型更合理地学习语料库的知识分布,提高语料库学习效率和学习准确率。

[0131] 实施例二

[0132] 本实施例一种可读存储介质,其中所述可读存储介质存储有计算机程序,所述计算机程序当被处理器执行时使所述处理器执行实施例一所述的基于图网络的汉字及结构的关联分析方法。

[0133] 实施例三

[0134] 本实施例一种计算机设备,包括处理器以及用于存储处理器可执行程序的存储器,所述处理器执行存储器存储的程序时,实现实施例一所述的基于图网络的汉字及结构的关联分析方法。

[0135] 上述实施例为本发明较佳的实施方式,但本发明的实施方式并不受上述实施例的限制,其他的任何未背离本发明的精神实质与原理下所作的改变、修饰、替代、组合、简化,均应为等效的置换方式,都包含在本发明的保护范围之内。

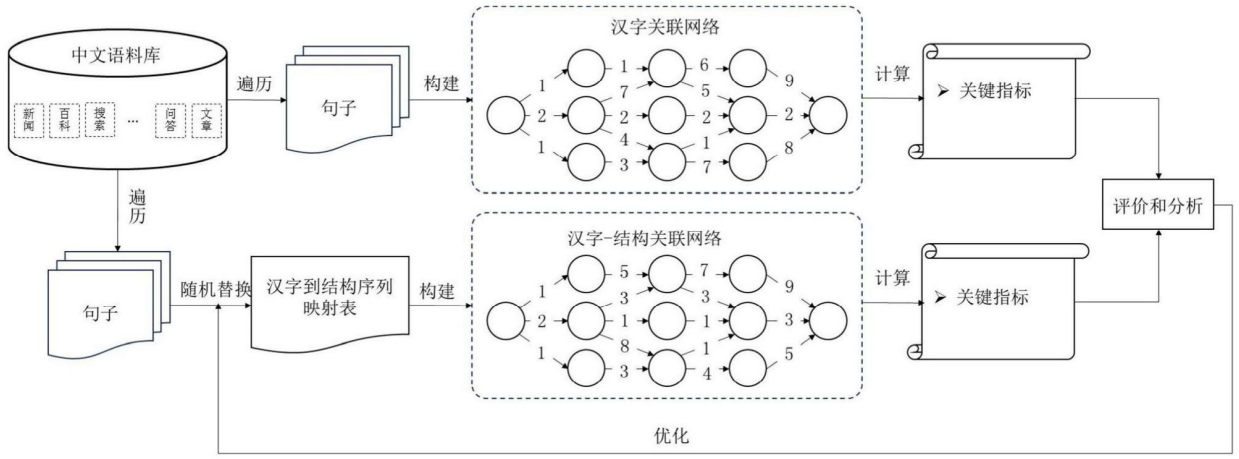

中文语料库
新闻
百科
搜索
...
问答
文章
遍历
句子
构建
汉字关联网络
计算
关键指标
评价和分析
遍历
句子
随机替换
汉字到结构序列映射表
构建
汉字-结构关联网络
计算
关键指标
优化

图1

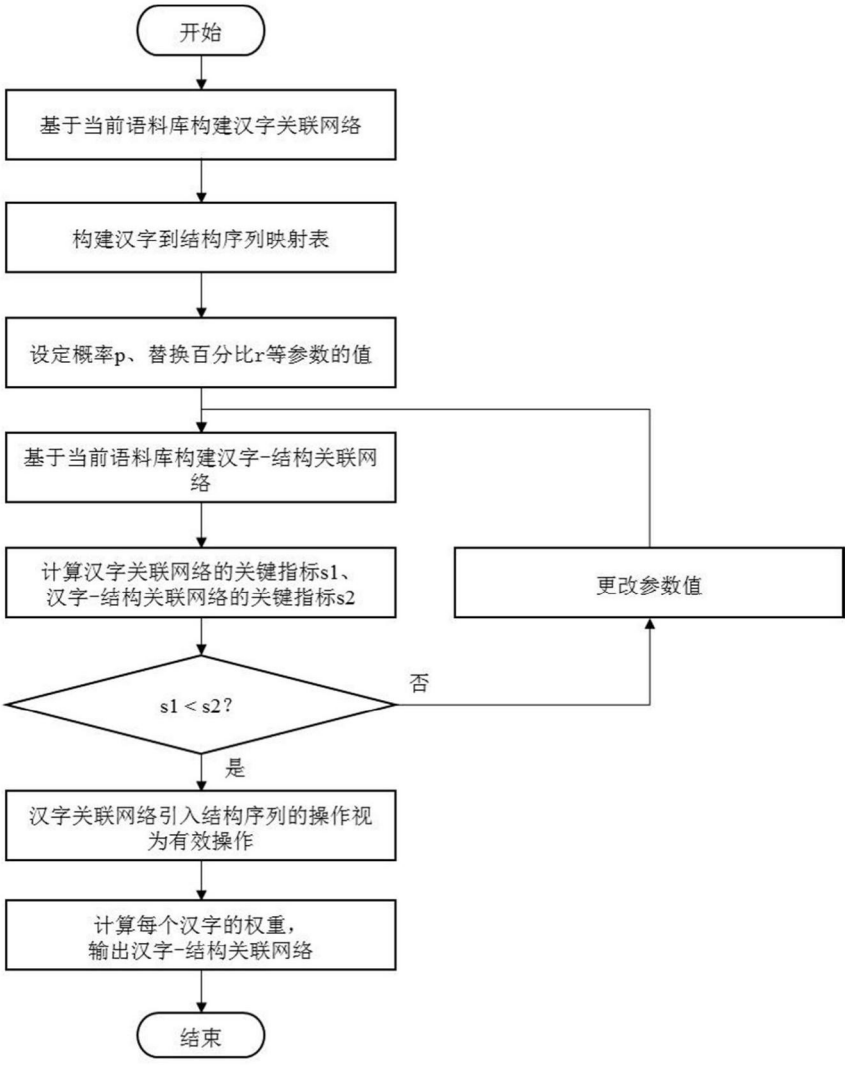

开始
基于当前语料库构建汉字关联网络
构建汉字到结构序列映射表
设定概率p、替换百分比r等参数的值
基于当前语料库构建汉字-结构关联网络
计算汉字关联网络的关键指标s1、汉字-结构关联网络的关键指标s2
更改参数值
s1 < s2?
否
是
汉字关联网络引入结构序列的操作视为有效操作
计算每个汉字的权重，输出汉字-结构关联网络
结束

图2

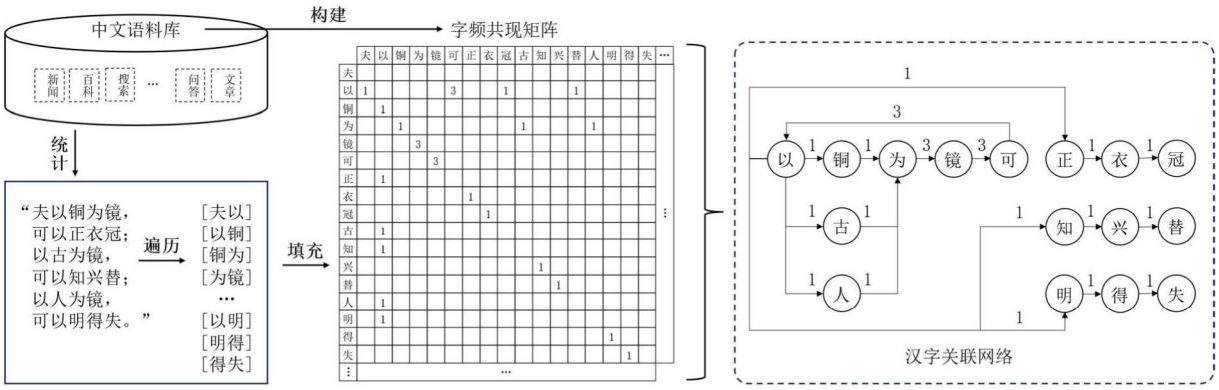

中文语料库
新闻
百科
搜索
...
问答
文章
构建
字频共现矩阵
统计
“夫以铜为镜，
可以正衣冠；
以古为镜，
可以知兴替；
以人为镜，
可以明得失。”
遍历
[夫以]
[以铜]
[铜为]
[为镜]
…
[以明]
[明得]
[得失]
填充
汉字关联网络

图3

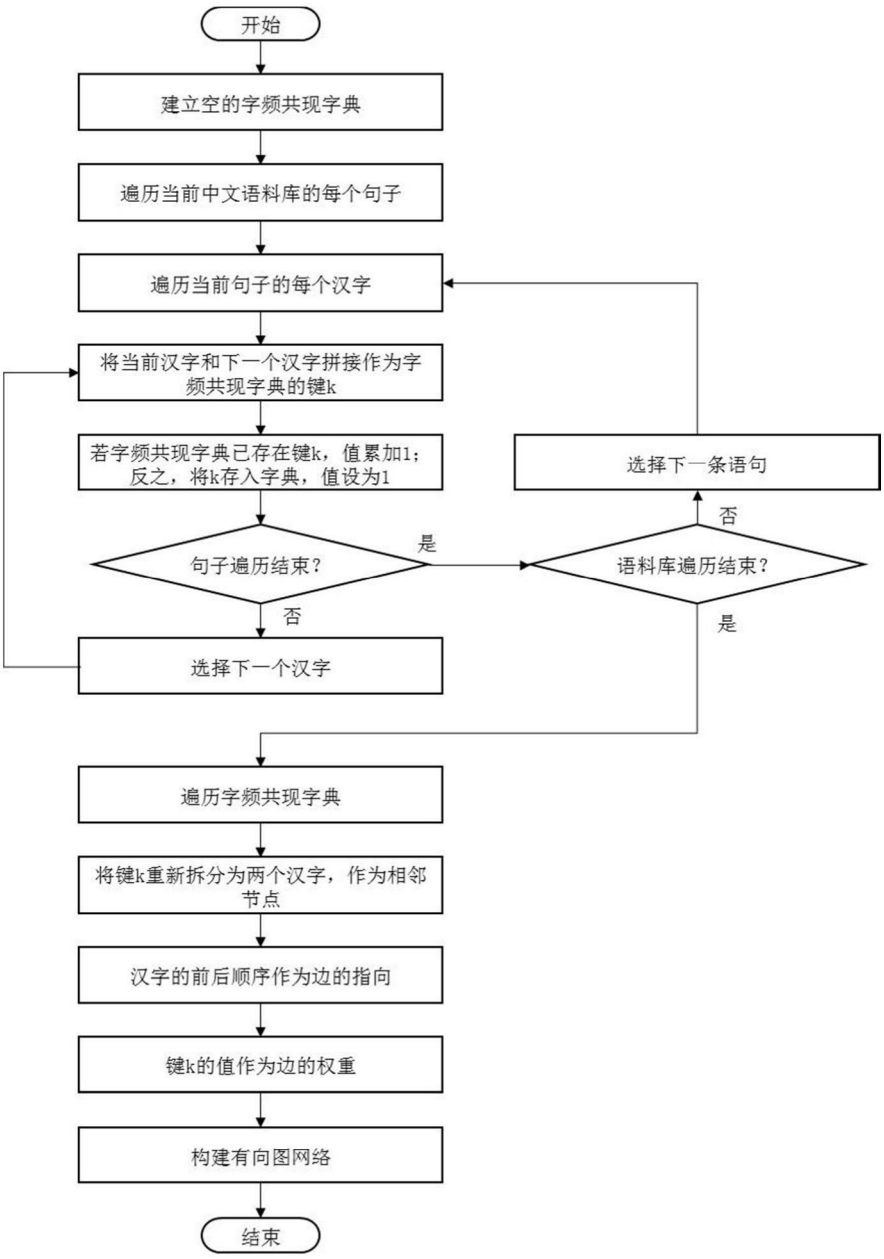

开始
建立空的字频共现字典
遍历当前中文语料库的每个句子
遍历当前句子的每个汉字
将当前汉字和下一个汉字拼接作为字频共现字典的键k
若字频共现字典已存在键k，值累加1；反之，将k存入字典，值设为1
句子遍历结束？
是
否
选择下一个汉字
选择下一条语句
语料库遍历结束？
否
是
遍历字频共现字典
将键k重新拆分为两个汉字，作为相邻节点
汉字的前后顺序作为边的指向
键k的值作为边的权重
构建有向图网络
结束

图4

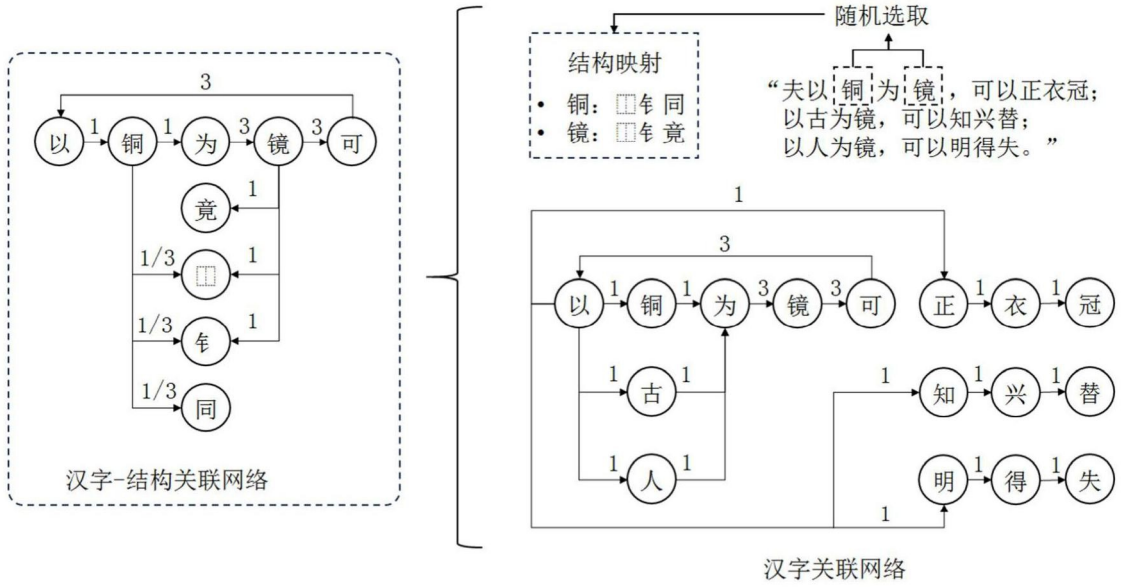

汉字-结构关联网络
以
铜
为
镜
可
竟
⿰
钅
同
随机选取
结构映射
铜：⿰钅同
镜：⿰钅竟
“夫以铜为镜，可以正衣冠；
以古为镜，可以知兴替；
以人为镜，可以明得失。”
正
衣
冠
古
人
知
兴
替
明
得
失
汉字关联网络

图5

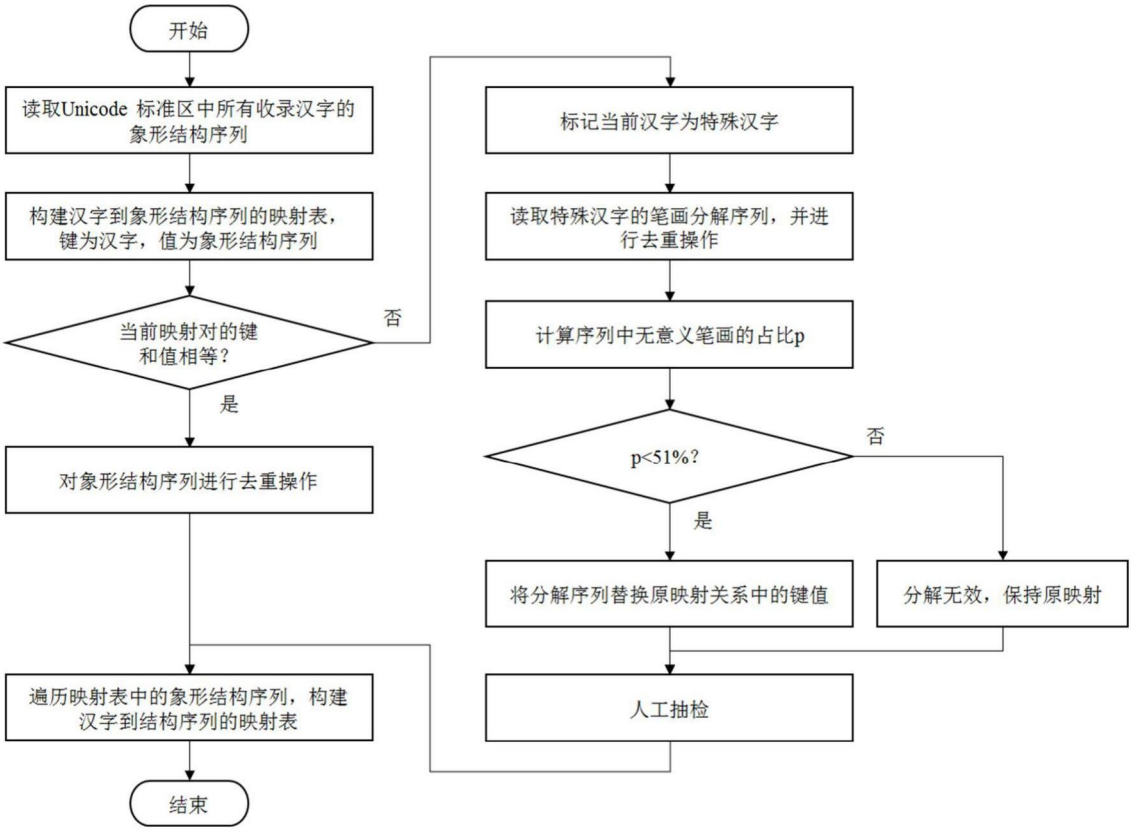

开始
读取Unicode 标准区中所有收录汉字的象形结构序列
构建汉字到象形结构序列的映射表，键为汉字，值为象形结构序列
当前映射对的键和值相等？
否
是
对象形结构序列进行去重操作
遍历映射表中的象形结构序列，构建汉字到结构序列的映射表
结束
标记当前汉字为特殊汉字
读取特殊汉字的笔画分解序列，并进行去重操作
计算序列中无意义笔画的占比p
p<51%?
否
是
将分解序列替换原映射关系中的键值
分解无效，保持原映射
人工抽检

图6

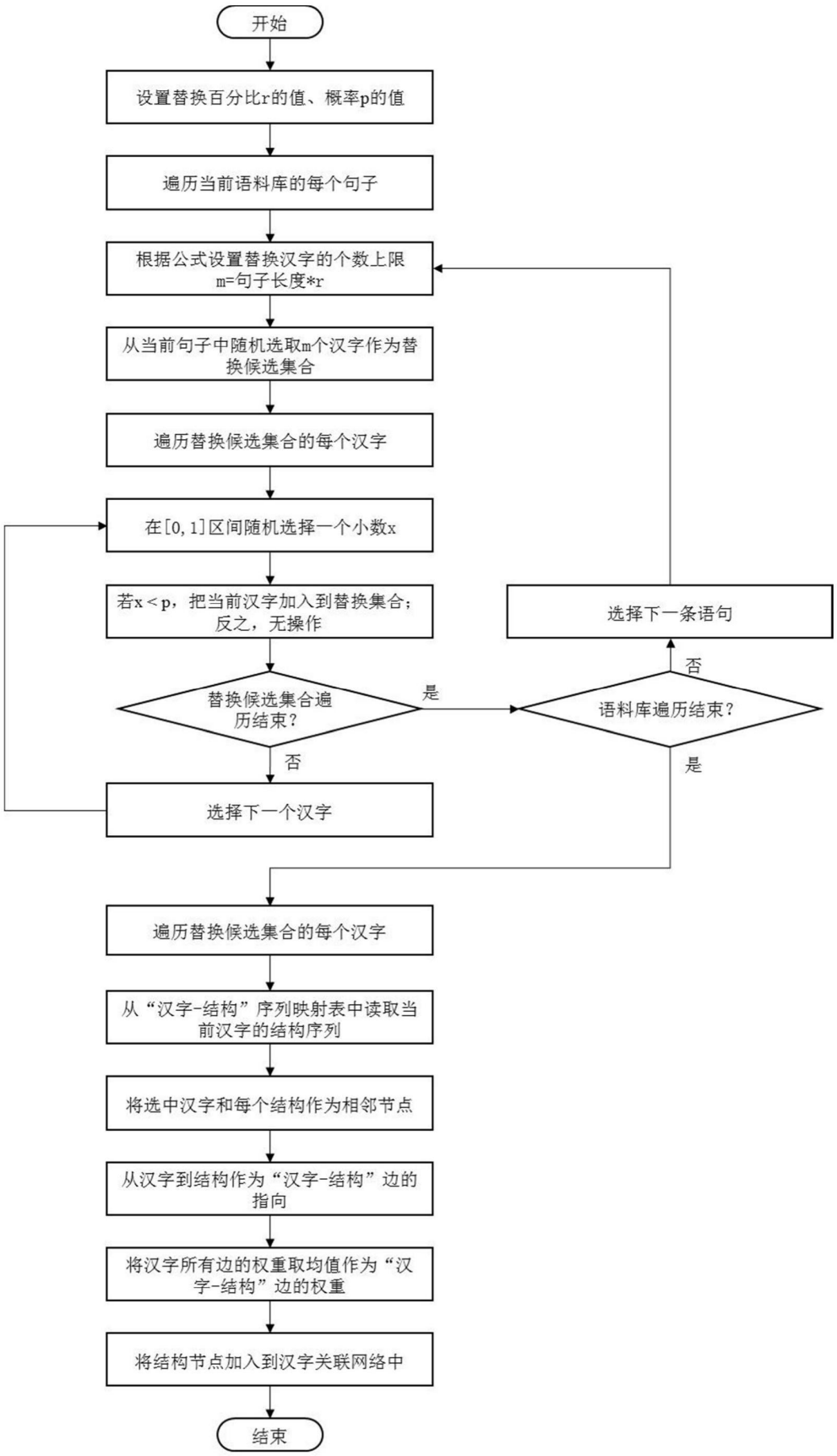

开始
设置替换百分比r的值、概率p的值
遍历当前语料库的每个句子
根据公式设置替换汉字的个数上限
m=句子长度*r
从当前句子中随机选取m个汉字作为替换候选集合
遍历替换候选集合的每个汉字
在[0,1]区间随机选择一个小数x
若x＜p，把当前汉字加入到替换集合；反之，无操作
替换候选集合遍历结束？
是
否
选择下一个汉字
选择下一条语句
语料库遍历结束？
否
是
遍历替换候选集合的每个汉字
从“汉字-结构”序列映射表中读取当前汉字的结构序列
将选中汉字和每个结构作为相邻节点
从汉字到结构作为“汉字-结构”边的指向
将汉字所有边的权重取均值作为“汉字-结构”边的权重
将结构节点加入到汉字关联网络中
结束

图7