(19)国家知识产权局

(12)发明专利

(10)授权公告号 CN 113569055 B
(45)授权公告日 2023.09.22

(21)申请号 202110843700.7

(22)申请日 2021.07.26

(65)同一申请的已公布的文献号
申请公布号 CN 113569055 A

(43)申请公布日 2021.10.29

(73)专利权人 东北大学
地址 110819 辽宁省沈阳市和平区文化路3号巷11号

(72)发明人 马连博 尹海源 王经纬 王兴伟 黄敏

(74)专利代理机构 沈阳东大知识产权代理有限公司 21109
专利代理师 梁焱

(51)Int.Cl.
G06F 16/36(2019.01)
G06F 40/295(2020.01)
G06N 3/086(2023.01)
G06N 3/126(2023.01)

(56)对比文件
CN 109243172 A,2019.01.18
CN 111341386 A,2020.06.26
CN 111522965 A,2020.08.11
CN 111813950 A,2020.10.23
CN 112288075 A,2021.01.29
US 2020370110 A1,2020.11.26
CN 109543047 A,2019.03.29
李振 等.自适应学习系统中知识图谱的人机协同构建方法与应用研究.《现代教育技术》.2019,第29卷(第10期),第80-86页.
Darkunde Mayur Ashok 等.Sarcasm Detection using Genetic Optimization on LSTM with CNN.《2020 International Conference for Emerging Technology (INCET)》.2020,第1-2页.

审查员 乔晋

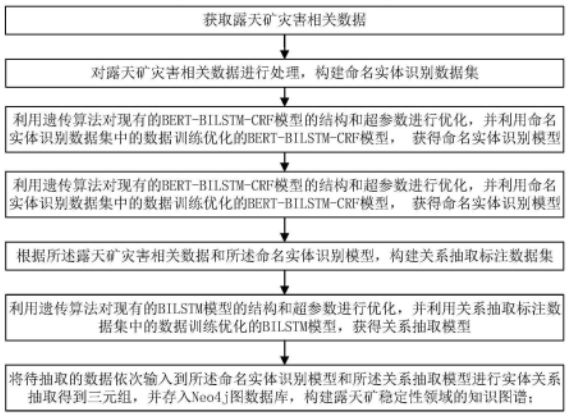
权利要求书3页 说明书8页 附图3页

(54)发明名称

基于遗传算法优化神经网络的露天矿知识图谱构建方法

(57)摘要

本发明公开了一种基于遗传算法优化神经网络的露天矿知识图谱构建方法,属于露天矿山稳定性评估技术领域。包括:获取露天矿灾害相关数据并对其进行处理构建命名实体识别数据集;利用遗传算法对现有BERT-BILSTM-CRF模型进行优化,并利用命名实体识别数据集中的数据训练优化的BERT-BILSTM-CRF模型,获得命名实体识别模型;构建关系抽取标注数据集;利用遗传算法对现有BILSTM模型进行优化,并利用关系抽取标注数据集中的数据训练优化的BILSTM模型,获得关系抽取模型;将待抽取的露天矿灾害相关数据依次输入到命名实体识别模型和关系抽取模型进行实体关系抽取得到三元组,并存入Neo4j图数据库,构建露天矿知识图谱。该方法能够兼顾上下文信息,能够搭建露天矿稳定性领域高质量的知识图谱。



CN 113569055 B

1.一种基于遗传算法优化神经网络的露天矿知识图谱构建方法,其特征在于,该方法包括如下步骤:

步骤1:获取露天矿灾害相关数据,包括灾害发生的原因、灾害发生的形式、灾害治理措施,灾害预防措施;

步骤2:对露天矿灾害相关数据进行处理,构建命名实体识别数据集;

步骤3:利用遗传算法对现有的BERT-BILSTM-CRF模型的结构和超参数进行优化,并利用所述命名实体识别数据集中的数据训练优化的BERT-BILSTM-CRF模型,获得命名实体识别模型;

步骤4:根据所述露天矿灾害相关数据和所述命名实体识别模型,构建关系抽取标注数据集;

步骤5:利用遗传算法对现有的BILSTM模型的结构和超参数进行优化,并利用所述关系抽取标注数据集中的数据训练优化的BILSTM模型,获得关系抽取模型;

步骤6:将待抽取的露天矿灾害相关数据依次输入到所述命名实体识别模型和所述关系抽取模型进行实体关系抽取得到三元组,并存入Neo4j图数据库,构建露天矿知识图谱;

所述步骤3包括如下具体步骤:

步骤3.1:为现有的BERT－BILSTM－CRF模型的每一个超参数设置初始化范围;

步骤3.2:设置遗传算法的最大迭代次数与设置种群规模;

步骤3.3:初始化个体:随机从上述每一个超参数的初始化范围内生成一个数值,利用所有生成的超参数的数值组成一个集合,表示一个个体;

步骤3.4:通过构建与每个个体相对应的BERT-BILSTM-CRF模型确定每个个体的适应度;

步骤3.5:根据每个个体的适应度,采用锦标赛选择算法选择出预设数量的优秀个体进入下一代;

步骤3.6:每次从优秀个体中选择两个个体进行交叉操作;

步骤3.7:对交叉操作后得到的新个体进行突变操作;

步骤3.8:重复执行步骤3.4至步骤3.7,直到达到最大迭代次数,选取最大的适应度对应的BERT-BILSTM-CRF模型,获得命名实体识别模型;

所述步骤4包括如下具体步骤:

步骤4.1:将露天矿灾害相关数据中的文本拆分成单个句子得到相应的句子数据集Sentence＝{sentence1,sentence2,sentence3,.....,sentencem},m表示句子的个数;

步骤4.2:通过调用命名实体识别模型识别句子数据集Sentence中的实体得到实体数据集Entity;

步骤4.3:手动抽取句子数据集Sentence中实体之间的关系得到关系数据集Relation;

步骤4.4:将Entity,Relation,Sentence合并得到关系抽取标注数据集Relation-Data＝{[entity1,relation1,sentence1],[entity2,relation2,sentence2],[entity3,relation3,senten-ce3],……,[entitym,relationm,sentencem]},其中m表示句子的个数;

所述步骤5包括如下具体步骤:

步骤5.1:为现有的BILSTM模型的每一个超参数设置初始化范围;

步骤5.2:设置遗传算法的最大迭代次数与设置种群规模;

步骤5.3:初始化个体:随机从上述每一个超参数的初始化范围内生成一个数值,利用所有生成的超参数的数值组成一个集合,表示一个个体;

步骤5.4:通过构建与每个个体相对应的BILSTM模型确定每个个体的适应度;

步骤5.5:根据每个个体的适应度,采用锦标赛选择算法选择出预设数量的优秀个体进入下一代;

步骤5.6:每次从优秀个体中选择两个个体进行交叉操作;

步骤5.7:对交叉操作后得到的新个体进行突变操作;

步骤5.8:重复执行步骤5.4至步骤5.7,直到达到最大迭代次数,选取最大的适应度对应的BERT-BILSTM-CRF模型,获得命名实体识别模型。

2.根据权利要求1所述的基于遗传算法优化神经网络的露天矿知识图谱构建方法,其特征在于,所述对露天矿灾害相关数据进行处理,构建命名实体识别数据集的方法为:首先过滤掉露天矿灾害相关数据中的无效词汇以及敏感词汇;然后将数据处理为单个字存入数据集Word中;再然后采用BIO标注方式对数据集Word中的每个字分别标注,标签存入数据集Label中;最后将数据集Word中的每个字和数据集Label中与数据集Word中的每个字对应的标签分别打包成元组后组合在一起构成命名实体识别数据集。

3.根据权利要求1所述的基于遗传算法优化神经网络的露天矿知识图谱构建方法,其特征在于,所述BERT－BILSTM－CRF模型的超参数包括结构超参数和训练超参数,其中结构超参数包括BILSTM层数和BILSTM隐藏层神经元数量,训练超参数包括时期epochs、批大小batch size、学习率和优化器。

4.根据权利要求1所述的基于遗传算法优化神经网络的露天矿知识图谱构建方法,其特征在于,所述确定每个个体的适应度的方法为:在确定每个个体的适应度时,首先将个体中的数值解析成BERT-BILSTM-CRF模型对应部分的超参数,根据解析出的超参数及其对应的数值构建出与每个个体相对应的BERT-BILSTM-CRF模型;然后,利用命名实体识别数据集中的数据对每个个体所对应的BERT-BILSTM-CRF模型进行训练、验证与测试,每个个体所对应的BERT-BILSTM-CRF模型测试后都会得到F1值,将每个F1值作为对应个体的适应度。

5.根据权利要求1所述的基于遗传算法优化神经网络的露天矿知识图谱构建方法,其特征在于,所述每次从优秀个体中选择两个个体进行交叉操作的方法为:采用两点交叉方式,随机生成两个交叉点作为个体交叉的开始位置和个体交叉的结束位置,然后对在两个交叉点之间的部分染色体进行交叉操作。

6.根据权利要求1所述的基于遗传算法优化神经网络的露天矿知识图谱构建方法,其特征在于,所述对交叉操作后的新个体进行突变操作的方法为:首先设置个体中每个对应位置的超参数的突变概率为P、针对每个对应位置的超参数生成一个[0,1]的随机数,如果生成的随机数小于或者等于P,则对当前位置的超参数进行突变操作,突变方式为从预设的每一个超参的初始化范围内重新生成一个随机数代替当前位置的超参数值;如果生成的随机数大于P,则不对当前位置的超参数进行突变操作。

7.根据权利要求1所述的基于遗传算法优化神经网络的露天矿知识图谱构建方法,其特征在于,所述BILSTM模型的超参数包括结构超参数和训练超参数,其中结构超参数包括BILSTM层数和BILSTM隐藏层神经元数量,训练超参数包括时期epochs、批大小batch size、

学习率和优化器。

基于遗传算法优化神经网络的露天矿知识图谱构建方法

技术领域

[0001] 本发明属于露天矿山稳定性评估技术领域，具体涉及一种基于遗传算法优化神经网络的露天矿知识图谱构建方法。

背景技术

[0002] 露天矿边坡稳定性是露天开采领域研究的关键问题。如何针对不同的露天矿提出合理的边坡设计及稳定性控制方案，是采矿工程科学技术人员亟待解决的问题。露天矿边坡稳定性受多种因素的影响，灾害模式也比较复杂。露天矿边坡安全设计、管理和灾害预警防控目前存在诸多问题。大量的类比案例没有形成数据库和深度分析归纳，缺乏智能化的分析理论，或者给矿山的生产带来安全隐患，或者导致成本增加。因此需要构建案例库、知识库和专家系统，建立一种灾害多因素多模式识别数学模型例如深度学习知识图谱开展分析案例，进行案例聚类和模式匹配。

[0003] 构建露天矿稳定性领域的灾害多因素模式的知识图谱的难点在于如何高效、高质量、快速地搭建。由于目前对矿业信息的获取大多来源于非结构的文本数据，所以想要搭建高质量的知识图谱就需要从非结构的文本数据中获取到准确的实体、关系和属性。准确地获取实体、关系和属性取决于命名实体识别模型和关系抽取模型的好坏。因此，设计出好的神经网络模型是非常关键的。神经网络模型的结构和超参数对模型效果起着关键的作用，大多数神经网络模型的结构和超参数都是通过手动设计的，对于没有专业知识的人员来说，很难设计出最佳的神经网络模型。

发明内容

[0004] 针对现有技术中存在的问题，本发明提供一种基于遗传算法优化神经网络的露天矿知识图谱构建方法。

[0005] 本发明的技术方案是：

[0006] 一种基于遗传算法优化神经网络的露天矿知识图谱构建方法，该方法包括如下步骤：

[0007] 步骤1：获取露天矿灾害相关数据，包括灾害发生的原因、灾害发生的形式、灾害治理措施，灾害预防措施；

[0008] 步骤2：对露天矿灾害相关数据进行处理，构建命名实体识别数据集；

[0009] 步骤3：利用遗传算法对现有的BERT-BILSTM-CRF模型的结构和超参数进行优化，并利用所述命名实体识别数据集中的数据训练优化的BERT-BILSTM-CRF模型，获得命名实体识别模型；

[0010] 步骤4：根据所述露天矿灾害相关数据和所述命名实体识别模型，构建关系抽取标注数据集；

[0011] 步骤5：利用遗传算法对现有的BILSTM模型的结构和超参数进行优化，并利用所述关系抽取标注数据集中的数据训练优化的BILSTM模型，获得关系抽取模型；

[0012] 步骤6:将待抽取的露天矿灾害相关数据依次输入到所述命名实体识别模型和所述关系抽取模型进行实体关系抽取得到三元组,并存入Neo4j图数据库,构建露天矿知识图谱。

[0013] 进一步地,根据所述的基于遗传算法优化神经网络的露天矿知识图谱构建方法,所述对露天矿灾害相关数据进行处理,构建命名实体识别数据集的方法为:首先过滤掉露天矿灾害相关数据中的无效词汇以及敏感词汇;然后将数据处理为单个字存入数据集Word中;再然后采用BIO标注方式对数据集Word中的每个字分别标注,标签存入数据集Label中;最后将数据集Word中的每个字和数据集Label中与数据集Word中的每个字对应的标签分别打包成元组后组合在一起构成命名实体识别数据集。

[0014] 进一步地,根据所述的基于遗传算法优化神经网络的露天矿知识图谱构建方法,所述步骤3包括如下具体步骤:

[0015] 步骤3.1:为现有的BERT－BILSTM－CRF模型的每一个超参数设置初始化范围;

[0016] 步骤3.2:设置遗传算法的最大迭代次数与设置种群规模;

[0017] 步骤3.3:初始化个体:随机从上述每一个超参数的初始化范围内生成一个数值,利用所有生成的超参数的数值组成一个集合,表示一个个体;

[0018] 步骤3.4:通过构建与每个个体相对应的BERT-BILSTM-CRF模型确定每个个体的适应度;

[0019] 步骤3.5:根据每个个体的适应度,采用锦标赛选择算法选择出预设数量的优秀个体进入下一代;

[0020] 步骤3.6:每次从优秀个体中选择两个个体进行交叉操作;

[0021] 步骤3.7:对交叉操作后得到的新个体进行突变操作;

[0022] 步骤3.8:重复执行步骤3.4至步骤3.7,直到达到最大迭代次数,选取最大的适应度对应的BERT-BILSTM-CRF模型,获得命名实体识别模型。

[0023] 进一步地,根据所述的基于遗传算法优化神经网络的露天矿知识图谱构建方法,所述BERT－BILSTM－CRF模型的超参数包括结构超参数和训练超参数,其中结构超参数包括BILSTM层数和BILSTM隐藏层神经元数量,训练超参数包括时期epochs、批大小batch size、学习率和优化器。

[0024] 进一步地,根据所述的基于遗传算法优化神经网络的露天矿知识图谱构建方法,所述确定每个个体的适应度的方法为:在确定每个个体的适应度时,首先将个体中的数值解析成BERT-BILSTM-CRF模型对应部分的超参数,根据解析出的超参数及其对应的数值构建出与每个个体相对应的BERT-BILSTM-CRF模型;然后,利用命名实体识别数据集中的数据对每个个体所对应的BERT-BILSTM-CRF模型进行训练、验证与测试,每个个体所对应的BERT-BILSTM-CRF模型测试后都会得到F1值,将每个F1值作为对应个体的适应度。

[0025] 进一步地,根据所述的基于遗传算法优化神经网络的露天矿知识图谱构建方法,所述每次从优秀个体中选择两个个体进行交叉操作的方法为:采用两点交叉方式,随机生成两个交叉点作为个体交叉的开始位置和个体交叉的结束位置,然后对在两个交叉点之间的部分染色体进行交叉操作。

[0026] 进一步地,根据所述的基于遗传算法优化神经网络的露天矿知识图谱构建方法,所述对交叉操作后的新个体进行突变操作的方法为:首先设置个体中每个对应位置的超参

数的突变概率为P、针对每个对应位置的超参数生成一个[0,1]的随机数，如果生成的随机数小于或者等于P，则对当前位置的超参数进行突变操作，突变方式为从预设的每一个超参的初始化范围内重新生成一个随机数代替当前位置的超参数值；如果生成的随机数大于P，则不对当前位置的超参数进行突变操作。

[0027] 进一步地，根据所述的基于遗传算法优化神经网络的露天矿知识图谱构建方法，所述步骤4包括如下具体步骤：

[0028] 步骤4.1：将露天矿灾害相关数据中的文本拆分成单个句子得到相应的句子数据集Sentence＝{sentence1,sentence2,sentence3,.....,sentencem}，m表示句子的个数；

[0029] 步骤4.2：通过调用命名实体识别模型识别句子数据集Sentence中的实体得到实体数据集Entity；

[0030] 步骤4.3：手动抽取句子数据集Sentence中实体之间的关系得到关系数据集Relation；

[0031] 步骤4.4：将Entity,Relation,Sentence合并得到关系抽取标注数据集Relation-Data＝{[entity1,relation1,sentence1],[entity2,relation2,sentence2],[entity3,relation3,senten-ce3],……,[entitym,relationm,sentencem]}，其中m表示句子的个数。

[0032] 进一步地，根据所述的基于遗传算法优化神经网络的露天矿知识图谱构建方法，所述步骤5包括如下具体步骤：

[0033] 步骤5.1：为现有的BILSTM模型的每一个超参数设置初始化范围；

[0034] 步骤5.2：设置遗传算法的最大迭代次数与设置种群规模；

[0035] 步骤5.3：初始化个体：随机从上述每一个超参数的初始化范围内生成一个数值，利用所有生成的超参数的数值组成一个集合，表示一个个体；

[0036] 步骤5.4：通过构建与每个个体相对应的BILSTM模型确定每个个体的适应度；

[0037] 步骤5.5：根据每个个体的适应度，采用锦标赛选择算法选择出预设数量的优秀个体进入下一代；

[0038] 步骤5.6：每次从优秀个体中选择两个个体进行交叉操作；

[0039] 步骤5.7：对交叉操作后得到的新个体进行突变操作；

[0040] 步骤5.8：重复执行步骤5.4至步骤5.7，直到达到最大迭代次数，选取最大的适应度对应的BERT-BILSTM-CRF模型，获得命名实体识别模型。

[0041] 进一步地，根据所述的基于遗传算法优化神经网络的露天矿知识图谱构建方法，所述BILSTM模型的超参数包括结构超参数和训练超参数，其中结构超参数包括BILSTM层数和BILSTM隐藏层神经元数量，训练超参数包括时期epochs、批大小batch size、学习率和优化器。

[0042] 本发明采用以上技术方案，具有以下有益效果：本方法针对手动设计神经网络模型结构和超参数方面的困难，将遗传算法与BERT-BILSTM-CRF模型、BILSTM模型相结合，进行模型超参数自动优化和选择，有效提高了模型的精度，经过多次迭代，得到最优的BERT-BILSTM-CRF模型作为命名实体识别模型和最优的BILSTM模型作为关系抽取模型。根据命名实体识别模型和关系抽取模型实现在露天矿文本数据中抽取实体关系，从而有效建立三元组，进而构建露天矿知识图谱。该方法能够兼顾上下文信息，增强泛化能力，能够搭建露天

矿稳定性领域高质量的知识图谱。

附图说明

[0043] 图1为本实施方式的基于遗传算法优化神经网络的露天矿知识图谱构建方法的流程示意图；

[0044] 图2为本实施方式中利用遗传算法和现有的BERT-BILSTM-CRF模型建立命名实体识别模型的流程示意图；

[0045] 图3为本实施方式的BERT-BILSTM-CRF模型的结构示意图；

[0046] 图4为将知识图谱应用于专家系统的流程图。

具体实施方式

[0047] 下面结合附图和具体实施方式，进一步阐明本发明，应理解这些实例仅用于说明本发明而不用于限制本发明的范围，在阅读了本发明之后，本领域技术人员对本发明的各种等价形式的修改均落于本申请所附权力要求所限定的范围。

[0048] 图1为是本发明实施方式的基于遗传算法优化神经网络的露天矿知识图谱构建方法的流程示意图，所述基于遗传算法优化神经网络的露天矿知识图谱构建方法，具体包括如下步骤：

[0049] 步骤1：获取露天矿灾害相关数据，包括灾害发生的原因、灾害发生的形式、灾害治理措施，灾害预防措施；

[0050] 在本实施例中，首先选择网上露天矿行业网站作为数据源，然后利用网络爬虫技术从相关网站上获取露天矿灾害相关的数据，包括露天矿灾害发生的原因、灾害发生的形式、灾害治理措施，灾害预防措施。

[0051] 步骤2：对露天矿灾害相关数据进行处理，构建命名实体识别数据集；

[0052] 首先过滤掉露天矿灾害相关数据中的无效词汇以及敏感词汇；然后定义数据中每个字构成的数据集Word＝{word1,word2,word3,……,wordnum1}，以及每个字对应的标签构成的数据集,Label＝{label1,label2,label3,……,labelnum2}。其中num1为数据集中字的个数；num2为数据集中标签的个数，Label包含{B-X,I-X,O}三种标签，X表示某种实体类型。在本实施方式中，采用BIO标注方式，将数据集中的每个元素标注为“B-X”、“I-X”或者“O”。其中，“B-X”表示此元素所在的片段属于X类型并且此元素在此片段的开头，“I-X”表示此元素所在的片段属于X类型并且此元素在片段的中间位置，“O”表示此元素所在片段不属于任何类型，即非实体；再然后遍历预处理后的数据，将数据处理为单个字存入数据集Word中，若所述单个字对应的词为X类型实体，则将实体的第一个字标注为“B-X”，并将该标签存入数据集Label中，实体剩余字标注为“I-X”，标注标签存入数据集Label中，将所有非实体标注为“O”标注标签存入数据集Label中；最后将数据集Word中的每个字，和数据集Label中与数据集Word中的每个字对应的标签构成的分别打包成元组后组合在一起构成命名实体识别数据集，Word、Label＝{[word1,label1],[word2,label2],[word3,label3],……,[wordn,labeln]}。其中n为预处理后的数据中字的个数，wordn为数据集中第n个字，labeln为数据集中第n个字的标签。

[0053] 步骤3：利用遗传算法对现有的BERT-BILSTM-CRF模型的结构和超参数进行优化，

并利用步骤2中得到的命名实体识别数据集中的数据训练优化的BERT-BILSTM-CRF模型，获得命名实体识别模型。所述步骤3如图2所示包括如下具体步骤：

[0054] 步骤3.1：为现有的BERT－BILSTM－CRF模型的每一个超参数设置初始化范围；

[0055] 根据神经网络模型的结构特点，BERT－BILSTM－CRF模型的超参数包括结构超参数和训练超参数，其中结构超参数包括BILSTM层数和BILSTM隐藏层神经元数量，训练超参数包括时期(epochs)、批大小(batch size)、学习率和优化器；在本实施例中，为每一个超参数设置的初始化范围如下：BILSTM层数[1,4]，BILSTM隐藏层神经元数量[200,400]，时期[1,100]，批大小[8,64]，学习率[0.00001,0.00003]和优化器[1,5]。其中，优化器的初始化范围表示相应优化器类型，其中1表示随机梯度下降算法(SGD)、2表示基于动量的算法(Momentum)、3表示自适应梯度算法(Adagrad)、4表示Adam算法、5表示前向均方根梯度下降算法(RMSprop)。

[0056] 步骤3.2：设置遗传算法的最大迭代次数与设置种群规模；

[0057] 在本实施例中，本步骤中的遗传算法的最大迭代次数设置为30代；在本实施例中，种群规模设置为50。

[0058] 步骤3.3：初始化个体，方法为：随机从上述每一个超参数的初始化范围内生成一个数值，利用所有生成的超参数的数值组成一个集合，表示一个个体；

[0059] 在本实施例中，种群中的个体采用实数编码，随机从上述每一个超参数的初始化范围内生成一个数值，利用所有生成的超参数的数值组成一个集合，表示一个个体。例如，随机生成BILSTM层数为2层，根据BILSTM层数，随机生成神经元数量为200，表示每个隐藏层神经元数量，然后生成时期为50，批大小为32，学习率为0.00001，优化器为2，则个体表示为{2,200,50,32,0.00001,2}；

[0060] 步骤3.4：通过构建与每个个体相对应的BERT-BILSTM-CRF模型确定每个个体的适应度。

[0061] 由上述可知，种群内的个体是根据BERT-BILSTM-CRF模型所需要的超参数组成的集合，本实施例在确定每个个体的适应度时，首先将个体中的数值解析成BERT-BILSTM-CRF模型对应部分的超参数，根据解析出的超参数及其对应的数值构建出与每个个体相对应的可训练的BERT-BILSTM-CRF模型，如图3所示。然后，将命名实体识别数据集按照5:3:2比例分割成命名实体识别训练集、命名实体识别验证集和命名实体识别测试集。使用命名实体识别训练集对每个个体所对应的BERT-BILSTM-CRF模型进行训练，使用命名实体识别验证集进行模型验证，使用命名实体识别测试集进行模型测试，每个个体所对应的BERT-BILSTM-CRF模型测试后都会得到F1值，将每个F1值作为对应个体的适应度进行保存。

[0062] 步骤3.5：根据每个个体的适应度，采用锦标赛选择算法选择出预设数量的优秀个体进入下一代，适应度越高的个体越优秀。

[0063] 步骤3.6：每次从优秀个体中选择两个个体进行交叉操作，采用两点交叉方式，随机生成两个交叉点作为个体交叉的开始位置和个体交叉的结束位置，然后对在两个交叉点之间的部分染色体进行交叉操作。

[0064] 步骤3.7：对交叉操作后得到的新个体进行突变操作。

[0065] 在本实施例中，对交叉操作后的新个体进行突变操作的方法为：首先设置个体中每个对应位置的超参数的突变概率为0.2、针对每个对应位置的超参数设置生成一个[0,1]

的随机数，如果生成的随机数小于或者等于0.2，则对当前位置的超参数进行突变操作。突变方式为从预设的每一个超参的初始化范围内重新生成一个随机数代替当前位置的超参数值。如果生成的随机数大于0.2，则不对当前位置超参数进行突变操作。

[0066] 步骤3.8：重复执行步骤3.4至步骤3.7，直到达到最大迭代次数，选取最大的F1值也即最大的适应度对应的BERT-BILSTM-CRF模型，获得命名实体识别模型。

[0067] 步骤4：根据所述露天矿灾害相关数据和所述命名实体识别模型，构建关系抽取标注数据集；

[0068] 步骤4.1：将露天矿灾害相关数据中的文本拆分成单个句子得到相应的句子数据集Sentence＝{sentence1,sentence2,sentence3,.....,sentencem}，m表示句子的个数；

[0069] 步骤4.2：通过调用命名实体识别模型识别句子数据集Sentence中的实体得到实体数据集Entity；

[0070] 步骤4.3：手动抽取句子数据集Sentence中实体之间的关系得到关系数据集Relation；

[0071] 步骤4.4：将Entity，Relation，Sentence合并得到关系抽取标注数据集Relation-Data＝{[entity1,relation1,sentence1],[entity2,relation2,sentence2],[entity3,relation3,senten-ce3],……,[entitym,relationm,sentencem]}，其中m表示句子的个数。

[0072] 步骤5：利用遗传算法对现有的BILSTM模型的结构和超参数进行优化，并利用步骤4中得到的关系抽取标注数据集中的数据训练优化的BILSTM模型，获得关系抽取模型。

[0073] 步骤5.1：为现有的BILSTM模型的每一个超参数设置初始化范围；

[0074] 根据神经网络模型的结构特点，BILSTM模型的超参数包括结构超参数和训练超参数，其中结构超参数包括BILSTM层数和BILSTM隐藏层神经元数量，训练超参数包括时期(epochs)、批大小(batch size)、学习率和优化器。

[0075] 在本实施例中，为每一个超参数设置的初始化范围如下：BILSTM层数[1,4]、BILSTM隐藏层神经元数量[200,400]、时期[1,100]、批大小[8,64]、学习率[0.00001,0.00003]、优化器[1,5]。其中，优化器的初始化范围表示相应优化器类型，其中1表示随机梯度下降算法(SGD)、2表示基于动量的算法(Momentum)、3表示自适应梯度算法(Adagrad)、4表示Adam算法、5表示前向均方根梯度下降算法(RMSprop)。

[0076] 步骤5.2：设置遗传算法的最大迭代次数与设置种群规模；

[0077] 在本实施例中，本步骤中的遗传算法的最大迭代次数设置为50代。当个体数量到达预设的种群规模形成种群，在本实施例中，种群规模设置为50。

[0078] 步骤5.3：初始化个体：随机从上述每一个超参数的初始化范围内生成一个数值，利用所有生成的超参数的数值组成一个集合，表示一个个体；

[0079] 在本实施例中，种群中的个体采用实数编码，随机从上述每一个超参数的初始化范围内生成一个数值，利用所有生成的超参数的数值组成一个集合，表示一个个体。例如，随机生成BILSTM层数为2层，根据BILSTM层数，随机生成神经元数量为200，表示每个隐藏层神经元数量，然后生成时期为50，批大小为32，学习率为0.00001，优化器为2，则个体表示为{2,200,50,32,0.00001,2}。

[0080] 步骤5.4：通过构建与每个个体相对应的BILSTM模型确定每个个体的适应度；

[0081] 由上述可知,种群内的个体是根据BILSTM模型所需要的超参数组成的集合,本实施例在确定每个个体的适应度时,首先将个体中的数值解析成BILSTM模型对应部分的超参数,根据解析出的超参数及其对应的数值构建出与每个个体相对应的可训练的BILSTM模型。然后,将关系抽取标注数据集按照5:3:2比例分割成关系抽取标注训练集、关系抽取标注验证集和关系抽取标注测试集。使用关系抽取标注训练集对每个个体所对应的BILSTM模型进行训练,使用关系抽取标注验证集进行模型验证,使用关系抽取标注测试集进行模型测试,每个个体所对应的BILSTM模型测试后都会得到模型F1值,将每个F1值作为对应个体的适应度进行保存。

[0082] 步骤5.5:根据每个个体的适应度,采用锦标赛选择算法选择出预设数量的优秀个体进入下一代,适应度越高的个体越优秀。

[0083] 步骤5.6:每次从优秀个体中选择两个个体进行交叉操作,采用两点交叉方式,随机生成两个交叉点作为个体交叉的开始位置和个体交叉的结束位置,然后对在两个交叉点之间的部分染色体进行交叉操作;

[0084] 步骤5.7:对交叉操作后得到的新个体进行突变操作;

[0085] 在本实施例中,对交叉操作后的新个体进行突变操作的方法为:首先设置个体中每个对应位置的超参数的突变概率为0.2、针对每个对应位置的超参数设置生成一个[0,1]的随机数,如果生成的随机数小于或者等于0.2,则对当前位置的超参数进行突变操作。突变方式为从预设的每一个超参的初始化范围内重新生成一个随机数代替当前位置的超参数值。如果生成的随机数大于0.2,则不对当前位置超参数进行突变操作;

[0086] 步骤5.8:重复执行步骤5.4至步骤5.7,直到达到最大迭代次数,选取最大的F1值也即最大的适应度对应的BILSTM模型,获得关系抽取模型。

[0087] 步骤6:将待抽取的露天矿灾害相关数据依次输入到所述命名实体识别模型和所述关系抽取模型进行实体关系抽取得到三元组,并存入Neo4j图数据库,构建露天矿稳定性领域的知识图谱;

[0088] 首先调用命名实体识别模型和关系抽取模型抽取待抽取文本,得到包含实体、属性、属性值的三元组triple=｛Item1,Relationship,Item2｝;然后将得到的实体、属性、属性值三元组triple存入三元组数据集Triples得到Triples=｛triple1,triple2,triple3,……,triples｝,其中s表示三元组个数;将获得的三元组数据集Triples存入到Neo4j图数据库,构建露天矿稳定性领域的知识图谱;

[0089] 本发明创造性的提出了遗传算法与BERT-BILSTM-CRF模型、BILSTM模型相结合,进行模型超参数自动优化和选择,经过多次迭代,得到最优的BERT-BILSTM-CRF模型作为命名实体识别模型和最优的BILSTM模型作为关系抽取模型。根据命名实体识别模型和关系抽取模型实现在露天矿文本数据中抽取实体关系,从而有效建立三元组,进而构建露天矿稳定性领域高质量的知识图谱。

[0090] 本发明构建的露天矿稳定性领域的知识图谱可以用于专家系统中,通过建立人机交互平台,挖掘用户的输入信息与所述露天矿稳定性领域知识图谱之间的关系,将挖掘出的专家信息返回给用户,具体是,如图4所示,通过人机交互平台获取用户的输入信息,利用所述命名实体识别模型识别用户输入中的实体;将识别出的实体与所述露天矿稳定性领域知识图谱中的实体根据余弦相似度进行对齐;对提取出的每个实体进行权重赋值。如果某

个实体跟越多的其他实体关联，说明这个实体越关键，权重越高。权重的大小与该实体和其他实体在知识图谱中的距离成反比；在所述的露天矿稳定性领域知识图谱中检索与全部实体距离最近的实体；将检索到的结果返回。

[0091] 这样就解决了传统神经网络模型的结构和超参数的选择需要有经验的专家进行设计，而非专业人士在神经网络模型结构和超参数的设计方面具有很大困难的问题。

[0092] 以上所述仅为本发明的实施例子而已，并不用于限制本发明。凡在本发明的原则之内，所做的等同替换，均应在本发明的保护范围之内，本发明未作详细阐述的内容属于本专业领域技术人员公知的已有技术。

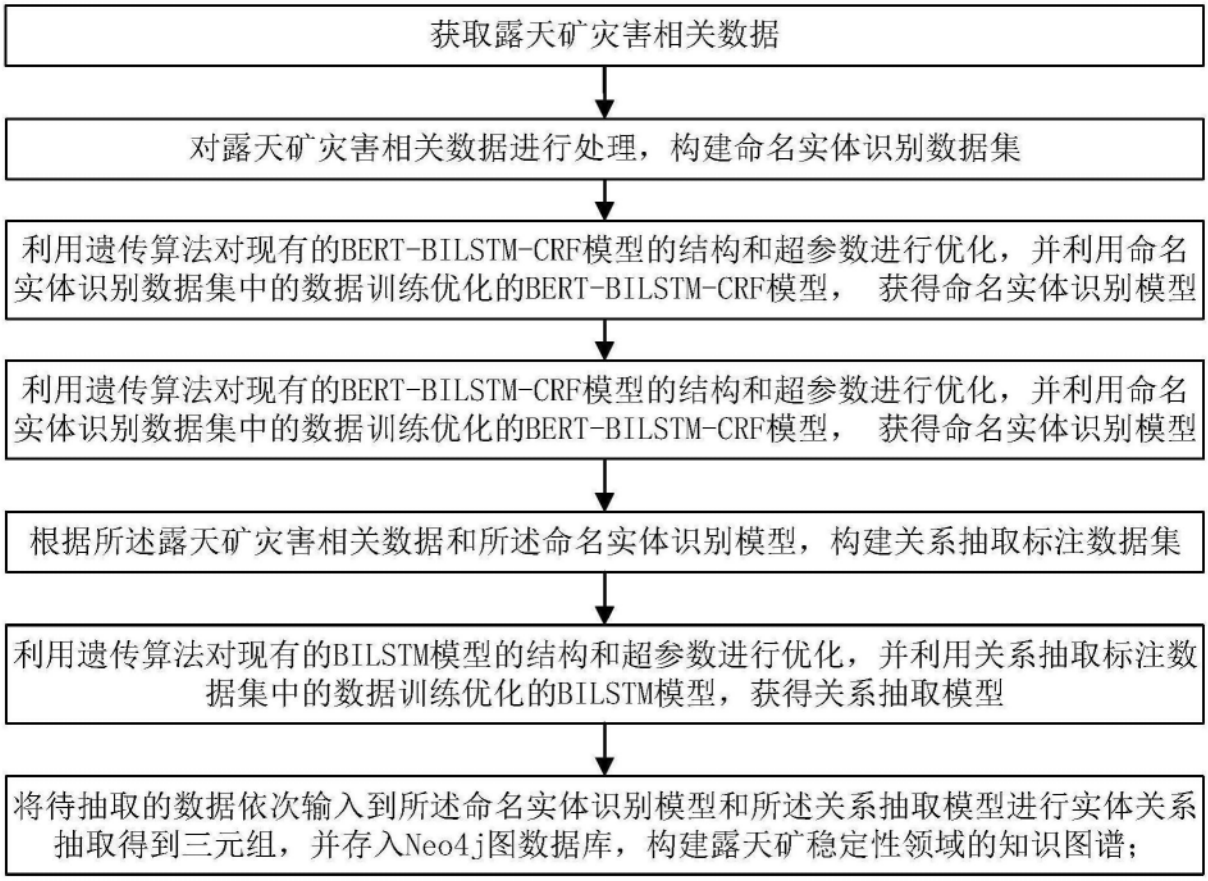

获取露天矿灾害相关数据
对露天矿灾害相关数据进行处理，构建命名实体识别数据集
利用遗传算法对现有的BERT-BILSTM-CRF模型的结构和超参数进行优化，并利用命名实体识别数据集中的数据训练优化的BERT-BILSTM-CRF模型， 获得命名实体识别模型
利用遗传算法对现有的BERT-BILSTM-CRF模型的结构和超参数进行优化，并利用命名实体识别数据集中的数据训练优化的BERT-BILSTM-CRF模型， 获得命名实体识别模型
根据所述露天矿灾害相关数据和所述命名实体识别模型，构建关系抽取标注数据集
利用遗传算法对现有的BILSTM模型的结构和超参数进行优化，并利用关系抽取标注数据集中的数据训练优化的BILSTM模型，获得关系抽取模型
将待抽取的数据依次输入到所述命名实体识别模型和所述关系抽取模型进行实体关系抽取得到三元组，并存入Neo4j图数据库，构建露天矿稳定性领域的知识图谱；

图1

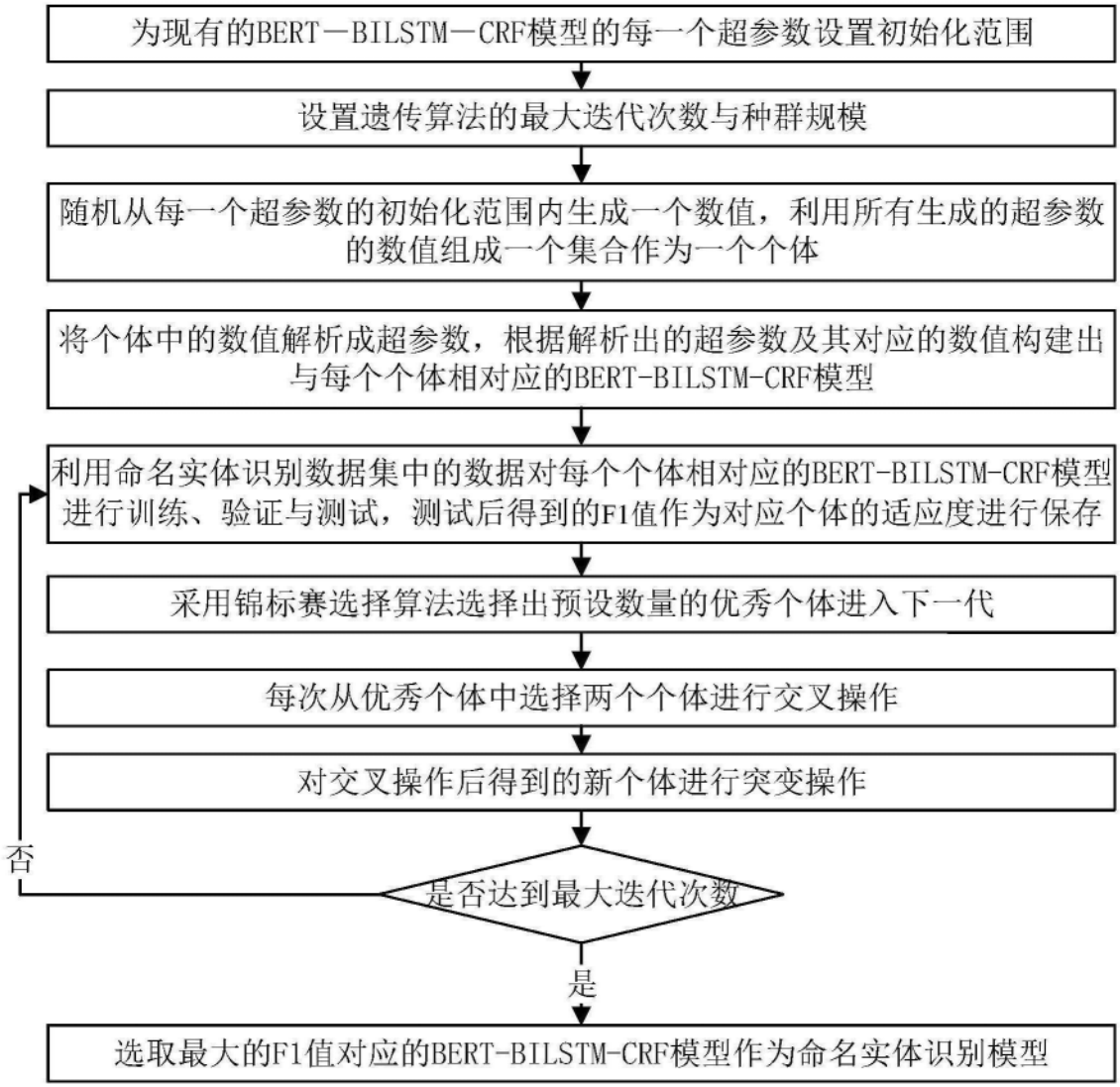

为现有的BERT－BILSTM－CRF模型的每一个超参数设置初始化范围
设置遗传算法的最大迭代次数与种群规模
随机从每一个超参数的初始化范围内生成一个数值，利用所有生成的超参数的数值组成一个集合作为一个个体
将个体中的数值解析成超参数，根据解析出的超参数及其对应的数值构建出与每个个体相对应的BERT-BILSTM-CRF模型
利用命名实体识别数据集中的数据对每个个体相对应的BERT-BILSTM-CRF模型进行训练、验证与测试，测试后得到的F1值作为对应个体的适应度进行保存
采用锦标赛选择算法选择出预设数量的优秀个体进入下一代
每次从优秀个体中选择两个个体进行交叉操作
对交叉操作后得到的新个体进行突变操作
是否达到最大迭代次数
否
是
选取最大的F1值对应的BERT-BILSTM-CRF模型作为命名实体识别模型

图2

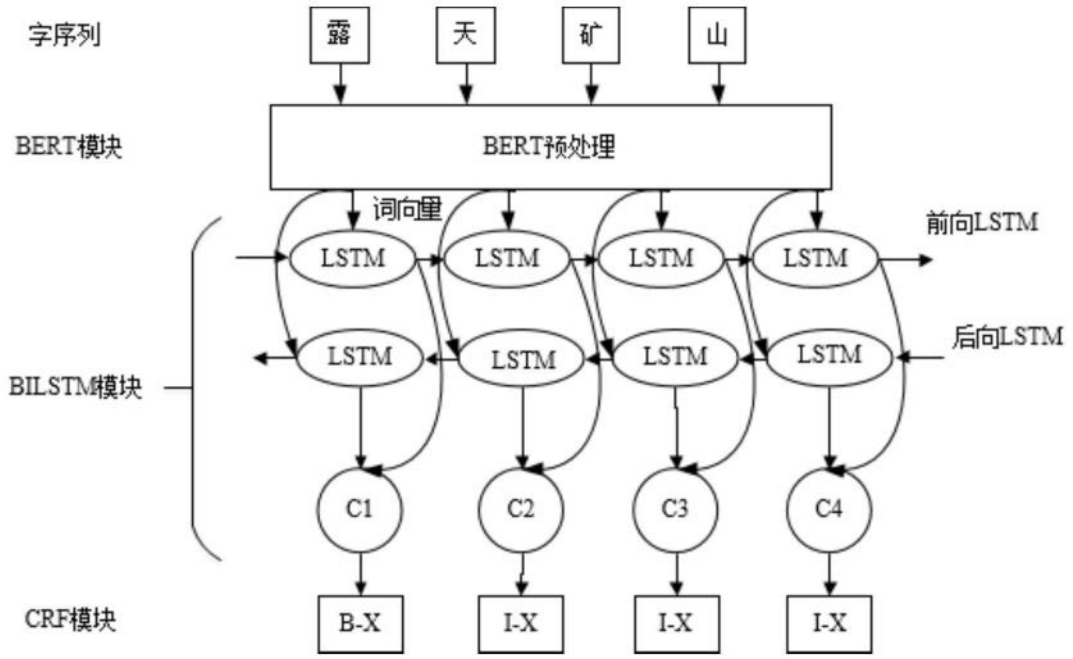

字序列
露
天
矿
山
BERT模块
BERT预处理
词向量
前向LSTM
LSTM
后向LSTM
BILSTM模块
C1
C2
C3
C4
CRF模块
B-X
I-X
I-X
I-X

图3

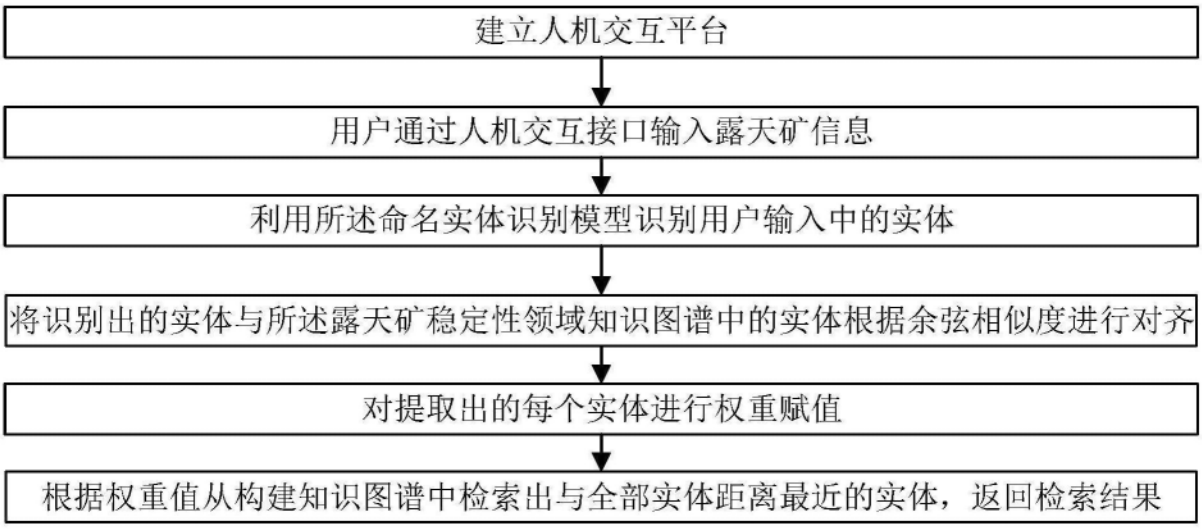

建立人机交互平台
用户通过人机交互接口输入露天矿信息
利用所述命名实体识别模型识别用户输入中的实体
将识别出的实体与所述露天矿稳定性领域知识图谱中的实体根据余弦相似度进行对齐
对提取出的每个实体进行权重赋值
根据权重值从构建知识图谱中检索出与全部实体距离最近的实体，返回检索结果

图4