(19)中华人民共和国国家知识产权局

(12)发明专利申请

(10)申请公布号 CN 111881677 A
(43)申请公布日 2020.11.03

(21)申请号 202010738017.2
(22)申请日 2020.07.28
(71)申请人 武汉大学
地址 430072 湖北省武汉市武昌区八一路299号
(72)发明人 亢孟军 刘越 苏世亮 翁敏 林玥 叶蕾
(74)专利代理机构 湖北武汉永嘉专利代理有限公司 42102
代理人 许美红
(51)Int.Cl.
*G06F 40/289*(2020.01)
*G06F 40/30*(2020.01)
*G06F 16/29*(2019.01)
*G06F 16/903*(2019.01)
*G06K 9/62*(2006.01)
*G06N 3/04*(2006.01)

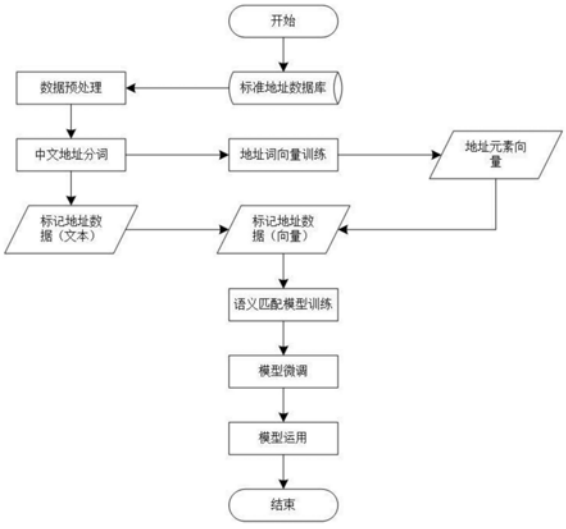
权利要求书4页 说明书10页 附图2页

(54)发明名称

基于深度学习模型的地址匹配算法

(57)摘要

本发明涉及一种基于深度学习模型的地址匹配算法,首先利用结巴(jieba)中文分词库对语料库中的地址进行分词;然后利用词向量(Word2vec)模型,进行地址词向量训练;最后利用增强序列推理模型(Enhanced Sequential Inference Model,ESIM)进行地址文本语义相似度计算,并输出匹配结果。该方法不同于传统的地址匹配算法侧重于利用匹配地址的字面重叠直接进行相似度计算与文本匹配,该算法侧重研究地址文本在语义上的相似程度,并以此为基础完成匹配任务,提供了一种适用于当今海量的多源异构地址数据匹配任务的深度学习算法。




CN 111881677 A

1.一种基于深度学习模型的地址匹配算法,其特征在于,包括以下步骤:

步骤1、对地址语料库进行数据预处理,包括去除语料库中的重复地址、空格及特殊符号,以及校正错别字改;所述地址语料库为标准地址库,其数据结构如下表1所示,其中,每一个待查询地址address_a分别对应1个正样本和1个负样本,采用UTF-8编码,正样本为匹配的address_b,负样本为不匹配的address_b;

表1 标准地址库数据结构

| 元素 | 描述 |
|---|---|
| address_a | 待查询地址 |
| address_b | 标准地址库地址 |
| label | 匹配标记,1为匹配,0为不匹配 |

步骤2、对预处理后的地址语料库进行中文分词,将地址文本中的词语与词语之间加上标记;

步骤3、对经过中文分词后的地址进行词向量训练,生成词表及其对应的多维词向量;

步骤4、地址文本语义匹配,包括以下子步骤:

步骤4.1.在进行模型训练之前,对实验数据集进行一系列预处理,以满足模型输入的要求,具体的数据预处理操作如下:

(1)将实验数据集中的地址文本进行分词;

(2)载入步骤3中生成的词表,将分词后的地址文本转化为词表ID序列;

(3)将label转化为独热编码One-Hot,设置正样本索引为1,负样本索引为2,则label为1时独热编码为[1,0],label为0时独热编码为[0,1];

步骤4.2.增强序列推理模型ESIM训练,具体包括:

(1)采用小批处理进行训练,并添加随机失活层,使每一个小批都使用随机丢弃了一部分神经网络节点的深度神经网络进行训练;

(2)在词嵌入层中采用动态词向量,将载入的预训练词向量设置为可训练模式,模型在训练过程中根据输入文本对预训练词向量进行更新,加速收敛;

(3)在计算模型损失时采用L2正则化,在损失函数后添加正则化项$\lambda\theta_L^2/2$对网络权值进行约束,λ为L2正则化参数;

(4)根据超参数的重要性水平,对模型的学习率、隐层节点数和小批大小进行调参,得到训练后的深度学习模型;

步骤5、将需要匹配的地址直接输入到训练后的深度学习模型,输出匹配后的结果。

2.根据权利要求1所述的基于深度学习模型的地址匹配算法,其特征在于,增强序列推理模型ESIM调参后采用的超参数设置如表2所示:

表2 增强序列推理模型ESIM超参数描述及实施例设置

| 超参数 | 描述 | 实验设置 |
| --- | --- | --- |
| 学习率 | 每个训练回合中参数更新的幅度，范围在 0-1 之间 | 0.0001 |
| 隐层节点数 | 神经网络隐层神经元数 | 50 |
| 小批（mini batch）大小 | 每个训练回合中使用的训练数据集的大小，取值在 10-100 范围内 | 50 |
| 随机失活（dropout）的保留概率 | 每个 dropout 过程中神经网络节点被保留下来的概率，范围在 0-1 之间 | 0.5 |
| L2 正则化参数 | L2 正则化项的值 | 0.01 |
| 训练回合数 | 训练集的迭代次数 | 20 |

。

3.根据权利要求1所述的基于深度学习模型的地址匹配算法，其特征在于，步骤2中，采用结巴jieba中文分词库对语料库中的地址进行分词，分词模式为精确模式；先用统计词典中所有可能成词的词语构建前缀词典，再根据该前缀词典得到输入文本的所有可能切分方式，并基于所有可能切分形成该输入文本的一个有向无环图，最后采用动态规划算法从后向前计算概率最大的文本切分形式。

4.根据权利要求3所述的基于深度学习模型的地址匹配算法，其特征在于，在分词过程中加载搜狗输入法词库和清华大学开放中文词库的地名词典作为分词的自定义词典。

5.根据权利要求1所述的基于深度学习模型的地址匹配算法，其特征在于，步骤3中，采用主题模型工具包自然语言处理库中的词向量模型对分词后的语料库进行词向量训练，训练过程中采用的模型为连续词袋模型CBOW，训练方法为负采样；训练时过滤词频小于5的词，并设置窗口大小为10、随机梯度下降的最大迭代次数为10，其余参数均设为默认值，最后生成该语料库的词表及其对应的256维词向量。

6.一种基于深度学习模型的地址匹配系统，其特征在于，包括：

语料库预处理模块，对地址语料库进行数据预处理，包括去除语料库中的重复地址、空格及特殊符号，以及校正错别字改；所述地址语料库为标准地址库，其数据结构如下表1所示，其中，每一个待查询地址address_a分别对应1个正样本和1个负样本，采用UTF-8编码，正样本为匹配的address_b，负样本为不匹配的address_b；

表1 标准地址库数据结构

| 元素 | 描述 |
| --- | --- |
| address_a | 待查询地址 |
| address_b | 标准地址库地址 |
| label | 匹配标记，1为匹配，0为不匹配 |

语料库分词模块，用于对预处理后的地址语料库进行中文分词，将地址文本中的词语与词语之间加上标记；

词向量训练模块，用于对经过中文分词后的地址进行词向量训练，生成词表及其对应的多维词向量；

实验数据集地址文本语义匹配模块，包括实验数据集预处理子模块和增强序列推理模

型(ESIM)训练子模块,其中,

实验数据集预处理子模块,用于实验数据集进行一系列预处理,以满足模型输入的要求,具体的数据预处理操作如下:

(1)将实验数据集中的地址文本进行分词;

(2)载入步骤3中生成的词表,将分词后的地址文本转化为词表ID序列;

(3)将label转化为独热编码,设置正样本索引为1,负样本索引为2,则label为1时独热编码为[1,0],label为0时独热编码为[0,1];

增强序列推理模型训练子模块,具体用于:

(1)采用小批处理进行训练,并添加随机失活层,使每一个小批都使用随机丢弃了一部分神经网络节点的深度神经网络进行训练;

(2)在词嵌入层中采用动态词向量,将载入的预训练词向量设置为可训练模式,模型在训练过程中根据输入文本对预训练词向量进行更新,加速收敛;

(3)在计算模型损失时采用L2正则化,在损失函数后添加正则化项$\lambda\theta_E^2/2$对网络权值进行约束,λ为L2正则化参数;

(4)根据超参数的重要性水平,对模型的学习率、隐层节点数和小批大小进行调参,得到训练后的深度学习模型;

匹配模块,用于将需要匹配的地址直接输入到训练后的深度学习模型,输出匹配后的结果。

7.根据权利要求6所述的基于深度学习模型的地址匹配系统,其特征在于,增强序列推理模型调参后采用的超参数设置如表2所示:

表2 增强序列推理模型ESIM超参数描述及实施例设置

| 超参数 | 描述 | 实验设置 |
|---|---|---|
| 学习率 | 每个训练回合中参数更新的幅度,范围在 0-1 之间 | 0.0001 |
| 隐层节点数 | 神经网络隐层神经元数 | 50 |
| 小批(mini batch)大小 | 每个训练回合中使用的训练数据集的大小,取值在 10-100 范围内 | 50 |
| 随机失活(dropout)的保留概率 | 每个 dropout 过程中神经网络节点被保留下来的概率, 范围在 0-1 之间 | 0.5 |
| L2 正则化参数 | L2 正则化项的值 | 0.01 |
| 训练回合数 | 训练集的迭代次数 | 20 |

。

8.根据权利要求6所述的基于深度学习模型的地址匹配系统,其特征在于,语料库分词模块具体采用结巴中文分词库对语料库中的地址进行分词,分词模式为精确模式;先用统计词典中所有可能成词的词语构建前缀词典,再根据该前缀词典得到输入文本的所有可能切分方式,并基于所有可能切分形成该输入文本的一个有向无环图,最后采用动态规划算法从后向前计算概率最大的文本切分形式。

9.根据权利要求6所述的基于深度学习模型的地址匹配系统，其特征在于，词向量训练模块具体采用主题模型工具包中的词向量模型对分词后的语料库进行词向量训练，训练过程中采用的模型为连续词袋模型，训练方法为负采样；训练时过滤词频小于5的词，并设置窗口大小为10、随机梯度下降的最大迭代次数为10，其余参数均设为默认值，最后生成该语料库的词表及其对应的256维词向量。

10.一种计算机存储介质，其特征在于，其内存储有可被处理器执行的计算机程序，该计算机程序执行如权利要求1-5中任一项所述的基于深度学习模型的地址匹配算法。

# 基于深度学习模型的地址匹配算法

## 技术领域

[0001] 本发明涉及一种计算机深度学习领域，尤其是涉及一种地址匹配的深度学习方法。

## 背景技术

[0002] 随着信息技术的快速发展，医疗、通信、物流等各个行业产生的时空数据日益增多。据统计，人类活动与城市信息中有80％以上是与地理空间位置有关的，其主要纽带就是地址信息，因此，地址在人们生活中扮演着越来越重要的作用。现有的行业数据(如医疗、公安等)通常将空间位置属性以地址文本的形式进行存储，要实现其在地理空间上的集中管理、分析与信息共享，则必须先将其转化为空间数据。要完成这一过程，就需要通过地址匹配查找并获得数据对应的地理坐标。地址匹配是地理编码的关键环节，其作用主要是将输入的地址文本与标准地址库中的地址进行配对，最终转化为空间地理坐标。目前，地址匹配已在城市管理、医疗服务、快递及邮政服务、灾害预警等方面获得了广泛应用。

[0003] 至今，已有大量的方法和技术被用于解决地址匹配问题，它们大致可以分为以下三类：基于字符串比较的地址匹配算法、基于要素层级模型的地址匹配算法和基于空间推理的地址匹配算法。基于字符串比较的地址匹配算法的特点为可不经过分词处理，直接利用编辑距离(Edit distance，或Levenshtein距离)等度量地址文本之间的相似性，衡量其匹配程度。此类方法对于非标准地址处理能力较差。基于要素层级模型的地址匹配算法在构建地址要素层级模型的基础上建立匹配规则库，在地址匹配过程中，通过规则库中的规则对匹配加以约束。一定程度上，规则库脱离了算法层面，具有较强的灵活性，方便做出修改和调整，但是由于中文地址的复杂性，匹配规则库需要考虑到匹配过程中遇到的多种可能性，若要有效提高地址匹配精度，其建立难度较大。基于空间推理的地址匹配算法强调了地址文本中隐含的空间位置关系，并根据空间关系进行相应的推理匹配。此类算法对地址质量要求较高，在实际应用中作用有限。

[0004] 近年来，随着移动设备及位置服务的普及，以地址为空间信息基础的海量行业数据开始大量涌现。由于中文地址在表达上的复杂性，多数地址仅在语义具有一定的空间指向性，但缺乏标准的地址结构。传统的地址匹配方法主要关注地址文本间字与字的匹配关系，无法准确识别不同表达方式下不同地址的同一指向关系：例如，“福永龙腾阁1巷2502”与“宝安区福永街道白石厦社区德丰路龙腾阁一巷2号”虽指向同一地理位置，但在字面表达上并无过多重叠，直接在字词层面进行匹配很难准确识别其相关关系。因此在这一背景下，要对多源异构的海量地址数据进行匹配，传统的地址匹配方法已不再适用。

## 发明内容

[0005] 本发明的目的在于提供一种基于深度学习模型的地址匹配算法，其充分利用地址中丰富的语义信息，很大程度上减小地址表达方式与结构对地址匹配度的影响，有效提高地址匹配准确率。

[0006] 本发明为达上述目的所采用的技术方案是：

[0007] 提供一种基于深度学习模型的地址匹配算法，包括以下步骤：

[0008] 步骤1、对地址语料库进行数据预处理，包括去除语料库中的重复地址、空格及特殊符号，以及校正错别字改；所述地址语料库为标准地址库，其数据结构如下表1所示，其中，每一个待查询地址address_a分别对应1个正样本和1个负样本，采用UTF-8编码，正样本为匹配的address_b，负样本为不匹配的address_b；

[0009] 表1标准地址库数据结构

[0010]

| 元素 | 描述 |
|---|---|
| address_a | 待查询地址 |
| address_b | 标准地址库地址 |
| label | 匹配标记，1为匹配，0为不匹配 |

[0011] 步骤2、对预处理后的地址语料库进行中文分词，将地址文本中的词语与词语之间加上标记；

[0012] 步骤3、对经过中文分词后的地址进行词向量训练，生成词表及其对应的多维词向量；

[0013] 步骤4、地址文本语义匹配，包括以下子步骤：

[0014] 步骤4.1.在进行模型训练之前，对实验数据集进行一系列预处理，以满足模型输入的要求，具体的数据预处理操作如下：

[0015] (1)将实验数据集中的地址文本进行分词；

[0016] (2)载入步骤3中生成的词表，将分词后的地址文本转化为词表ID序列；

[0017] (3)将label转化为独热(One-Hot)编码，设置正样本索引为1，负样本索引为2，则label为1时独热(One-Hot)编码为[1,0]，label为0时独热(One-Hot)编码为[0,1]；

[0018] 步骤4.2.增强序列推理模型(ESIM)训练，具体包括：

[0019] (1)采用小批处理(mini batch)进行训练，并添加随机失活(dropout)层，使每一个小批(mini batch)都使用随机丢弃了一部分神经网络节点的深度神经网络进行训练；

[0020] (2)在词嵌入层中采用动态词向量，将载入的预训练词向量设置为可训练模式，模型在训练过程中根据输入文本对预训练词向量进行更新，加速收敛；

[0021] (3)在计算模型损失时采用L2正则化，在损失函数后添加正则化项$\lambda\theta_i^2/2$对网络权值进行约束，λ为L2正则化参数；

[0022] (4)根据超参数的重要性水平，对模型的学习率、隐层节点数和小批(mini batch)大小进行调参，得到训练后的深度学习模型；

[0023] 步骤5、将需要匹配的地址直接输入到训练后的深度学习模型，输出匹配后的结果。

[0024] 接上述技术方案，增强序列推理模型(ESIM)调参后采用的超参数设置如表2所示：

[0025] 表2增强序列推理模型(ESIM)超参数描述及实施例设置

[0026]

| 超参数 | 描述 | 实验设置 |
| --- | --- | --- |
| 学习率 | 每个训练回合中参数更新的幅度，范围在 0-1 之间 | 0.0001 |
| 隐层节点数 | 神经网络隐层神经元数 | 50 |
| 小批（mini batch）大小 | 每个训练回合中使用的训练数据集的大小，取值在 10-100 范围内 | 50 |
| 随机失活（dropout）的保留概率 | 每个 dropout 过程中神经网络节点被保留下来的概率，范围在 0-1 之间 | 0.5 |
| L2 正则化参数 | L2 正则化项的值 | 0.01 |
| 训练回合数 | 训练集的迭代次数 | 20 |

[0027] 接上述技术方案，采用结巴(jieba)中文分词库对语料库中的地址进行分词，分词模式为精确模式；先用统计词典中所有可能成词的词语构建前缀词典，再根据该前缀词典得到输入文本的所有可能切分方式，并基于所有可能切分形成该输入文本的一个有向无环图，最后采用动态规划算法从后向前计算概率最大的文本切分形式。

[0028] 接上述技术方案，在分词过程中加载搜狗输入法词库和清华大学开放中文词库的地名词典作为分词的自定义词典。

[0029] 接上述技术方案，步骤3中，采用主题模型工具包(gensim)中的词向量(Word2Vec)模型对分词后的语料库进行词向量训练，训练过程中采用的模型为连续词袋(Continuous Bag-of-Words，CBOW)模型，训练方法为负采样；训练时过滤词频小于5的词，并设置窗口大小为10、随机梯度下降的最大迭代次数为10，其余参数均设为默认值，最后生成该语料库的词表及其对应的256维词向量。

[0030] 本发明还提供了一种基于深度学习模型的地址匹配系统，包括：

[0031] 语料库预处理模块，对地址语料库进行数据预处理，包括去除语料库中的重复地址、空格及特殊符号，以及校正错别字改；所述地址语料库为标准地址库，其数据结构如下表1所示，其中，每一个待查询地址address_a分别对应1个正样本和1个负样本，采用UTF-8编码，正样本为匹配的address_b，负样本为不匹配的address_b；

[0032] 表1标准地址库数据结构

[0033]

| 元素 | 描述 |
| --- | --- |
| address_a | 待查询地址 |
| address_b | 标准地址库地址 |
| label | 匹配标记，1为匹配，0为不匹配 |

[0034] 语料库分词模块，用于对预处理后的地址语料库进行中文分词，将地址文本中的词语与词语之间加上标记；

[0035] 词向量训练模块，用于对经过中文分词后的地址进行词向量训练，生成词表及其对应的多维词向量；

[0036] 实验数据集地址文本语义匹配模块，包括实验数据集预处理子模块和增强序列推理模型(ESIM)训练子模块，其中，

[0037] 实验数据集预处理子模块，用于实验数据集进行一系列预处理，以满足模型输入的要求，具体的数据预处理操作如下：

[0038] (1)将实验数据集中的地址文本进行分词；

[0039] (2)载入步骤3中生成的词表，将分词后的地址文本转化为词表ID序列；

[0040] (3)将label转化为独热(One-Hot)编码，设置正样本索引为1，负样本索引为2，则label为1时独热(One-Hot)编码为[1,0]，label为0时独热(One-Hot)编码为[0,1]；

[0041] 增强序列推理模型(ESIM)训练子模块，具体用于：

[0042] (1)采用小批处理(mini batch)进行训练，并添加随机失活(dropout)层，使每一个小批(mini batch)都使用随机丢弃了一部分神经网络节点的深度神经网络进行训练；

[0043] (2)在词嵌入层中采用动态词向量，将载入的预训练词向量设置为可训练模式，模型在训练过程中根据输入文本对预训练词向量进行更新，加速收敛；

[0044] (3)在计算模型损失时采用L2正则化，在损失函数后添加正则化项$\lambda\theta_i^2/2$对网络权值进行约束，λ为L2正则化参数；

[0045] (4)根据超参数的重要性水平，对模型的学习率、隐层节点数和小批(mini batch)大小进行调参，得到训练后的深度学习模型；

[0046] 匹配模块，用于将需要匹配的地址直接输入到训练后的深度学习模型，输出匹配后的结果。

[0047] 接上述技术方案，增强序列推理模型(ESIM)调参后采用的超参数设置如表2所示：

[0048] 表2增强序列推理模型(ESIM)超参数描述及实施例设置

[0049]

| 超参数 | 描述 | 实验设置 |
|---|---|---|
| 学习率 | 每个训练回合中参数更新的幅度，范围在 0-1 之间 | 0.0001 |
| 隐层节点数 | 神经网络隐层神经元数 | 50 |
| 小批（mini batch）大小 | 每个训练回合中使用的训练数据集的大小，取值在 10-100 范围内 | 50 |
| 随机失活（dropout）的保留概率 | 每个 dropout 过程中神经网络节点被保留下来的概率，范围在 0-1 之间 | 0.5 |
| L2 正则化参数 | L2 正则化项的值 | 0.01 |
| 训练回合数 | 训练集的迭代次数 | 20 |

。

[0050]

[0051] 接上述技术方案，语料库分词模块具体采用结巴(jieba)中文分词库对语料库中的地址进行分词，分词模式为精确模式；先用统计词典中所有可能成词的词语构建前缀词典，再根据该前缀词典得到输入文本的所有可能切分方式，并基于所有可能切分形成该输入文本的一个有向无环图，最后采用动态规划算法从后向前计算概率最大的文本切分形式。

[0052] 接上述技术方案，词向量训练模块具体采用主题模型工具包(gensim)中的词向量(Word2Vec)模型对分词后的语料库进行词向量训练，训练过程中采用的模型为连续词袋

(Continuous Bag-of-Words,CBOW)模型,训练方法为负采样;训练时过滤词频小于5的词,并设置窗口大小为10、随机梯度下降的最大迭代次数为10,其余参数均设为默认值,最后生成该语料库的词表及其对应的256维词向量。

[0053] 本发明还提供了一种计算机存储介质,其内存储有可被处理器执行的计算机程序,该计算机程序执行如上述技术方案所述的基于深度学习模型的地址匹配算法。

[0054] 本发明产生的有益效果是:利用结巴(jieba)中文分词库对语料库中的地址进行分词;然后利用词向量(Word2vec)模型,进行地址词向量训练;最后利用增强序列推理模型(ESIM)进行地址文本语义相似度计算,并输出匹配结果。该方法不同于传统的地址匹配算法侧重于利用匹配地址的字面重叠直接进行相似度计算与文本匹配,该算法侧重研究地址文本在语义上的相似程度,并以此为基础完成匹配任务,可很好地解决地址数据量庞大、地址标准化率较低等现象造成的地址匹配精度差这一问题。

## 附图说明

[0055] 下面将结合附图及实施例对本发明作进一步说明,附图中:

[0056] 图1是本发明的基于深度学习模型的地址匹配算法流程图;

[0057] 图2是本发明的地址词向量训练连续词袋(CBOW)模型原理图;

[0058] 图3是本发明的地址增强序列推理模型(ESIM)结构图;

[0059] 图4是本发明的双向长短期记忆(BiLSTM)模型结构图。

## 具体实施方式

[0060] 为了使本发明的目的、技术方案及优点更加清楚明白,以下结合附图及实施例,对本发明进行进一步详细说明。应当理解,此处所描述的具体实施例仅用以解释本发明,并不用于限定本发明。

[0061] 本发明实施例基于深度学习模型的地址匹配算法,如图1示,包括以下步骤:

[0062] 步骤1.数据预处理。对语料库进行去除语料库中的重复地址、去除空格及特殊符号、对语料库中的错别字进行修改等预处理工作。

[0063] 本发明实施例中采用的语料库为标准地址库,用于地址文本语义匹配的数据集共含有84,474对标记地址数据,其数据结构如表1所示。其中,每一个待查询地址address_a分别对应1个正样本(匹配的address_b)和1个负样本(不匹配的address_b)。该数据采用UTF-8编码。

[0064] 表1实施例数据结构

[0065]

| 元素 | 描述 |
|---|---|
| address_a | 待查询地址 |
| address_b | 标准地址库地址 |
| label | 是否匹配标记,1为匹配,0为不匹配 |

[0066] 步骤2.中文分词。对于自然语言处理来说,词是最小的有意义的研究单位。在拉丁文语系中,词与词之间有明显分隔符的存在,而这一点是中文所不具备的,中文字符之间没有空格等分隔符。因此对中文文本的分析,需要将文本转换为最小语义单位“词”才能进行,即将一个汉字序列切分成一个一个单独的词。由于中文词语缺少形式上的分隔符,故在进

行词向量训练之前,应先通过中文分词将地址文本中的词语与词语之间加上标记,即分词。本发明采用结巴(jieba)中文分词库对语料库中的地址进行分词,分词模式为精确模式。结巴(jieba)分词的原理为:先用统计词典中所有可能成词的词语构建前缀词典,再根据该前缀词典得到输入文本的所有可能切分方式,并基于所有可能切分形成该输入文本的一个有向无环图,最后采用动态规划算法从后向前计算概率最大的文本切分形式。

[0067] 为使分词结果更加准确,在分词过程中可加载搜狗输入法词库和清华大学开放中文词库(THU Open Chinese Lexicon,THUOCL)的地名词典作为分词的自定义词典。

[0068] 本发明采用机械分词和统计分词相结合的分词方法,首先使用已有的地名分词词典来进行字符串匹配机械分词,同时使用统计模型识别词典不包含的新词。该方法结合了机械分词和统计分词的优点,不仅切分速度快、效率高,而且可以结合上下文共现频率识别生词、消除歧义。

[0069] 步骤3.地址词向量训练。本发明采用主题模型工具包(gensim)中的词向量(Word2Vec)模型对分词后的语料库进行词向量训练。训练过程中采用的模型为连续词袋(Continuous Bag-of-Words,CBOW)模型,训练方法为负采样;训练时过滤词频小于5的词,并设置窗口大小为10、随机梯度下降的最大迭代次数为10,其余参数均设为默认值。最后生成该语料库的词表及其对应的256维词向量。

[0070] 步骤4.地址文本语义匹配,包括以下子步骤:

[0071] 步骤4.1.数据预处理。在进行模型训练之前,首先对实验数据集进行一系列预处理,以满足模型输入的要求。具体的数据预处理操作如下:

[0072] (1)中文分词。将地址文本进行分词,词与词之间以空格分隔。

[0073] (2)地址文本转化为词表ID序列。载入步骤3中生成的词表,用词语在词表中的ID(即行号)表示上一步分词后的文本。

[0074] (3)label转化为独热(One-Hot)编码。设置正样本索引为1,负样本索引为2,则label为1时独热(One-Hot)编码为[1,0],label为0时独热(One-Hot)编码为[0,1]。

[0075] 步骤4.2.增强序列推理模型(Enhanced Sequential Inference Model,ESIM)训练,本发明在增强序列推理模型(ESIM)训练中采取了以下策略:

[0076] (1)为防止过拟合,提高模型的泛化能力,采用小批处理(mini batch)进行训练,并添加随机失活(dropout)层,使每一个小批(mini batch)都使用随机丢弃了一部分神经网络节点的深度神经网络进行训练。

[0077] (2)在词嵌入层中采用动态词向量,即将载入的预训练词向量设置为可训练模式(trainable=True)。这样,模型在训练过程中可根据输入文本对预训练词向量进行更新,从而加速收敛。

[0078] (3)在计算模型损失时采用了L2正则化,即在损失函数后添加正则化项$\lambda\theta_i^2/2$(λ为L2正则化参数)对网络权值进行约束,从而避免模型过于复杂,降低过拟合的风险。

[0079] (4)根据超参数的重要性水平,主要对模型的学习率、隐层节点数和小批(mini batch)大小进行调参。

[0080] 本发明在增强序列推理模型(ESIM)调参后采用的超参数设置如表2所示。

[0081] 表2增强序列推理模型(ESIM)超参数描述及实施例设置

[0082]

| 超参数 | 描述 | 实验设置 |
| --- | --- | --- |
| 学习率 | 每个训练回合中参数更新的幅度。范围在 0-1 之间。 | 0.0001 |

[0083]

| | | |
| --- | --- | --- |
| | 学习率过小或使模型的收敛速度过慢；学习率过大则或导致模型的训练损失因越过最优值而无法收敛。 | |
| 隐层节点数 | 神经网络隐层神经元数。隐层节点数过少或使模型无法有效进行学习，导致模型精度降低；隐层节点数过多则或使模型复杂度增大，从而增加过拟合的风险。 | 50 |
| 小批（mini batch）大小 | 每个训练回合中使用的训练数据集的大小。一般取值在 10-100 范围内。Mini batch 过大或使模型陷入局部最小，从而导致过拟合；mini batch 过小则或使模型的收敛速度过慢。 | 50 |
| 随机失活（dropout）的保留概率 | 每个 dropout 过程中神经网络节点被保留下来的概率，范围在 0-1 之间。 | 0.5 |
| L2 正则化参数 | L2 正则化项的值。 | 0.01 |
| 训练回合数 | 训练集的迭代次数。 | 20 |

[0084] 步骤2中,中文地址分词的步骤原理如下：

[0085] (1)基于前缀词典进行高效的词图扫描,生成一个句子中汉字字符所有可能成词序列情况所构成的有向无环图(Directed Acyclic Graph,DAG)；

[0086] (2)进行动态规划找到最大概率的路径,从而找出基于词频的最大概率切分组合；

[0087] (3)采用基于中文成词能力的隐马尔科夫模型和维特比(Viterbi)算法处理未登录词。

[0088] 进一步地,步骤3中,本发明采用连续词袋(Continuous Bag-of-Words,CBOW)模型(如图2所示)在地址语料库上进行词向量训练,最终生成的词向量将作为地址文本表征用作地址语义匹配模型的输入。本发明使用的词向量(Word2vec)模型运用负采样(Negative sampling)算法对模型的训练过程进行了优化。一般情况下,模型在训练过程中每使用一个训练样本就将更新神经网络的所有权重;对于大语料库而言,这种训练方法将使模型的计算效率将大大降低。为降低模型的计算负担,因此采用负采样使得每使用一个训练样本仅对一部分网络权重进行更新。

[0089] 地址词向量训练的步骤原理如下：

[0090] (1)模型初始化。扫描语料库U,生成词表V,并对词表中的每一个词随机生成一个长度为l的词向量w,则语料库U可看作由词向量序列($w_1$,$w_2$,…,$w_N$)组成;随机初始化所有模

型参数；

[0091] (2)负采样。假设词表V对应一条长度为1的线段，按照词频的大小可得到该词表内每个词对应的长度：

[0092] $$\mathrm{len}(w_i)=\frac{count(w_i)^{\frac{3}{4}}}{\sum_V count(w_i)^{\frac{3}{4}}}$$

[0093] 其中，count($w_i$)为词$w_i$的词频。

[0094] 将词表V对应的线段平均分为M(M＞＞V)等份，每份的长度为1/M。假设当前词为$w_k$，生成neg个0～M之间的整数，查找其对应位置上的词，即可生成neg个负例wt，t∈{1，…，neg}，记为Neg($w_k$)。

[0095] (3)随机梯度上升训练。对于当前词$w_k$，假设当前词在窗口d范围内的上下文为$w_{k+i}$；i∈{-d，…-1，1，…，d}，记为Context($w_k$)。训练的目标函数为最大化当前词被预测为正例的概率，即

[0096] $$g(w_k)=\prod_{u\epsilon w_k\cup Neg(w_k)} p(u|Context(w_k))$$

[0097] 最后采用随机梯度上升法进行训练。

[0098] 进一步地，步骤4中，经过中文分词、生成词表ID序列和独热(One-Hot)编码等数据预处理后，本发明采用增强序列推理模型(ESIM)作为地址文本语义匹配的基本模型，并通过小批处理(mini batch)、随机失活(dropout)层、L2正则化等参数设置优化模型匹配效果。增强序列推理模型(ESIM)基本思路为：先提取待匹配的两个文本在词级别上的表示信息，再将提取文本间对应位置的交互信息，并构建文本的匹配矩阵，最后抽取该矩阵更高层次的匹配特征，并输出结果(如图3所示)。地址文本语义匹配的步骤原理如下：

[0099] (1)词嵌入层。在词嵌入层(Embedding layer)中，首先载入预训练地址词向量矩阵，并输入待匹配地址的词ID序列$w_a=(w_a^1,w_a^2,\cdots,w_a^{l_a})$与标准地址库地址的词ID序列$w_b=(w_b^1,w_b^2,\cdots,w_b^{l_b})$。然后，根据词id查找矩阵相应位置上的词向量，并拼接为该地址文本的向量表示，即两个词嵌入矩阵$a=(a_1,a_2,\cdots,a_{l_a})$和$b=(b_1,b_2,\cdots,b_{l_b})$。

[0100] (2)输入编码层。输入编码层(Input encoding layer)运用双向长短期记忆模型(Bi-directional long short-term memory，BILSTM)对输入的地址词嵌入矩阵a和b进行进一步编码。双向长短期记忆模型(BILSTM)编码原理为：分别用一个前向长短期记忆模型(LSTM)和一个后向长短期记忆模型(LSTM)从左往右作用于词嵌入矩阵，再将两个长短期记忆模型(LSTM)的输出进行拼接作为该词嵌入矩阵的编码(如图4所示)。输入编码层的工作可用下列公式进行表示：

[0101] $$\bar{a}_i=BiLSTM(a,i),\forall i\in[1,\ldots,l_a]$$

[0102] $$\bar{b}_j=BiLSTM(b,j),\forall j\in[1,\ldots,l_b]$$

[0103] (3)局部推断层。局部推断层(Local inference modeling layer)主要利用改进的可分解注意力(Decomposable attention)机制的对两个文本编码间的相似性进行局部推断，其在实现上主要分为三个部分：

[0104] ①权重矩阵的生成。根据软注意力(Soft attention)原理,计算$\bar{a}_i$和$\bar{b}_j$的点积作为其局部相似性的表示,并以此生成两编码序列的注意力权重矩阵。注意力权重的表达式为:

[0105] $$e_{ij} = \bar{a}_i^T \bar{b}_j$$

[0106] ②序列的局部推断。对编码序列$\bar{a}_i$,利用softmax函数求出其对应的注意力权重的概率分布,并将该结果与$\bar{b}_j$进行点乘;对$\bar{b}_j$也采用类似的方法。该步骤可得到两个编码序列之间的交互表示,其对应的数学表达式为:

[0107] $$\tilde{a}_i = \sum_{j=1}^{l_b} \frac{\exp(e_{ij})}{\sum_{t=1}^{l_b} exp(e_{it})} \bar{b}_j, \forall i \in [1, \ldots, l_a]$$

[0108] $$\tilde{b}_j = \sum_{i=1}^{l_a} \frac{\exp(e_{ij})}{\sum_{t=1}^{l_a} exp(e_{tj})} \bar{a}_i, \forall j \in [1, \ldots, l_b]$$

[0109] ③局部推断信息的增强。对二元组$(\bar{a}, \tilde{a})$,将其中的两个元素与其差和元素乘积相连接,从而得到一组可表示序列$\bar{a}_i$基于$\bar{b}_j$的局部推断信息序列,从而增强局部推理信息;对二元组$(\bar{b}, \tilde{b})$也采用类似的方法。该步骤对应的数学表达式为:

[0110] $$m_a = [\bar{a}; \tilde{a}; \bar{a} - \tilde{a}; \bar{a} \odot \tilde{a}]$$

[0111] $$m_b = [\bar{b}; \tilde{b}; \bar{b} - \tilde{b}; \bar{b} \odot \tilde{b}]$$

[0112] (4)推断组合层。推断组合层(Inference composition layer)的主要作用是在文本编码相似性的局部推断的基础上进行全局推断,其在实现上主要分为两个部分:

[0113] ①局部推断信息的提取。该步骤使用输入编码层中的双向长短期记忆模型(BILSTM)进一步提取$m_a$和$m_b$中的局部推断信息,其表达式为:

[0114] $$v_{a,i} = BiLSTM(m_a, i), \forall i \in [1, \ldots, l_a]$$

[0115] $$v_{b,j} = BiLSTM(m_b, j), \forall j \in [1, \ldots, l_b]$$

[0116] ②池化。对上一步提取到的信息$v_{a,i}$和$v_{b,j}$,采用最大池化和平均池化的方法进行降维并保留重要特征,最终连接成一个固定长度的向量v。该步骤的表达式为:

[0117] $$v_{a,ave} = \sum_{i=a}^{l_a} \frac{v_{a,i}}{l_a}, v_{a,max} = \max_{l_a} v_{a,i}$$

[0118] $$v_{b,ave} = \sum_{j=b}^{l_b} \frac{v_{b,j}}{l_b}, v_{b,max} = \max_{l_b} v_{b,j}$$

[0119] $v = [v_{a,ave}; v_{a,max}; v_{b,ave}; v_{b,max}]$

[0120] (5)预测输出层。预测输出层(Prediction layer)利用多层感知机(MLP)对最终结果进行拟合并输出:将上一步输出的向量v作为输入层,最后输出两地址文本之间的匹配关系(0为不匹配,1为匹配)。这里使用的多层感知机(MLP)共含有三个全连接隐藏层,其激活函数分别为ReLU、tanh和softmax。其中,ReLU函数的表达式为:

[0121] $f(x) = \max(0, x)$

[0122] tanh为双曲正切函数,其表达式为:

[0123] $$f(x)=\frac{sinh(x)}{cosh(x)}=\frac{e^{x}-e^{-x}}{e^{x}+e^{-x}}$$

[0124] softmax函数用于计算各类别的预测概率。最终使用argmax函数输出预测类别。

[0125] 综上，本发明设计并实现了基于深度学习的地址匹配算法，不同于传统的地址匹配算法，该算法侧重研究地址文本在语义上的相似程度，并以此为基础完成匹配任务，有利于解决地址数据量庞大、地址标准化率较低等现象造成的地址匹配精度差这一问题。

[0126] 本发明是参照根据本发明实施例的方法、设备(系统)、和计算机程序产品的流程图和/或方框图来描述的。应理解可由计算机程序指令实现流程图和/或方框图中的每一流程和/或方框、以及流程图和/或方框图中的流程和/或方框的结合。可提供这些计算机程序指令到通用计算机、专用计算机、嵌入式处理机或其他可编程数据处理设备的处理器以产生一个机器，使得通过计算机或其他可编程数据处理设备的处理器执行的指令产生用于实现在流程图一个流程或多个流程和/或方框图一个方框或多个方框中指定的功能的装置。

[0127] 尽管已描述了本发明的优选实施例，但本领域内的技术人员一旦得知了基本创造性概念，则可对这些实施例做出另外的变更和修改。所以，所附权利要求意欲解释为包括优选实施例以及落入本发明范围的所有变更和修改。

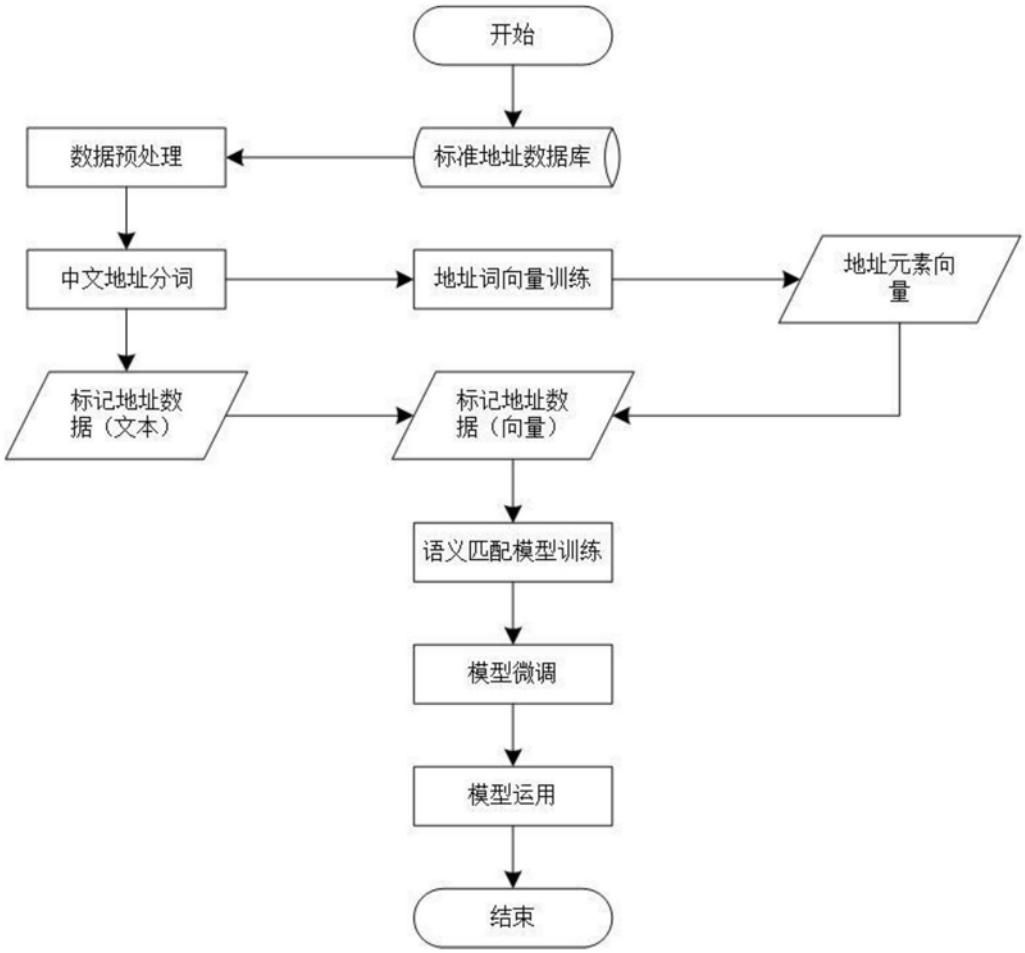

开始
标准地址数据库
数据预处理
中文地址分词
地址词向量训练
地址元素向量
标记地址数据（文本）
标记地址数据（向量）
语义匹配模型训练
模型微调
模型运用
结束

图1

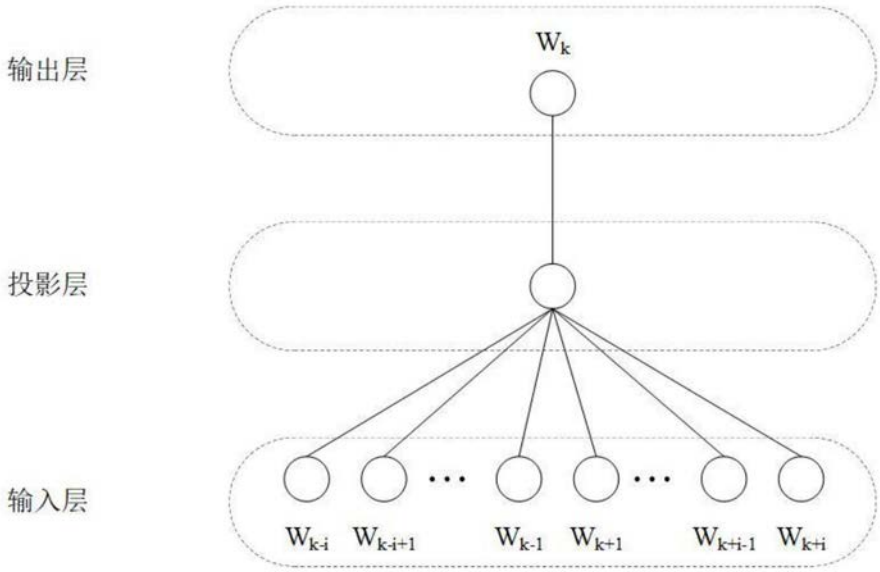

输出层
投影层
输入层
$W_k$
$W_{k-i}$
$W_{k-i+1}$
$W_{k-1}$
$W_{k+1}$
$W_{k+i-1}$
$W_{k+i}$

图2

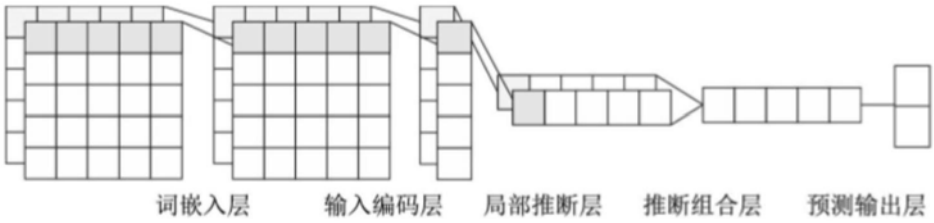

词嵌入层
输入编码层
局部推断层
推断组合层
预测输出层

图3

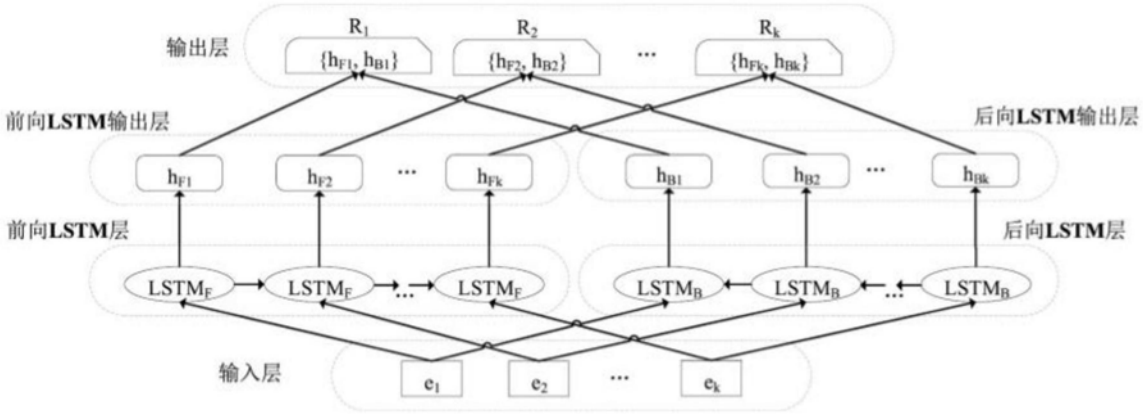

输出层
前向LSTM输出层
后向LSTM输出层
前向LSTM层
后向LSTM层
输入层

图4