(19)国家知识产权局

(12)发明专利申请

(10)申请公布号 CN 120106201 A
(43)申请公布日 2025.06.06

(21)申请号 202510586685.0

(22)申请日 2025.05.08

(71)申请人 杭州聪宝科技有限公司
地址 310000 浙江省杭州市西湖区文二路391号西湖国际科技大厦5号楼5层536室

(72)发明人 顾高生

(74)专利代理机构 北京超凡宏宇知识产权代理有限公司 11463
专利代理师 张红

(51)Int.Cl.
G06N 5/022(2023.01)
G06N 5/04(2023.01)
G06F 18/213(2023.01)
G06F 18/25(2023.01)

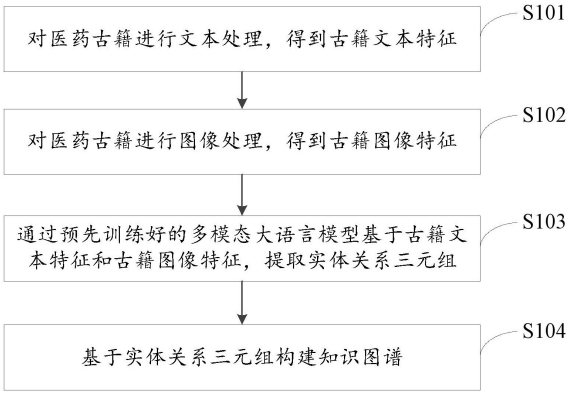
权利要求书2页 说明书14页 附图3页

(54)发明名称

一种医药古籍知识图谱构建方法及系统

(57)摘要

本发明提供了一种医药古籍知识图谱构建方法及系统,涉及知识图谱技术领域,该方法对医药古籍进行文本处理,得到古籍文本特征;对医药古籍进行图像处理,得到古籍图像特征;通过预先训练好的多模态大语言模型基于古籍文本特征和古籍图像特征,提取实体关系三元组;基于实体关系三元组构建知识图谱。本发明能够有效提升实体及其关系提取的准确性和全面性,更好的构建知识图谱,提高知识图谱的准确性和构建效率。



CN 120106201 A

1.一种医药古籍知识图谱构建方法,其特征在于,所述方法包括:

对医药古籍进行文本处理,得到古籍文本特征;

对所述医药古籍进行图像处理,得到古籍图像特征;

通过预先训练好的多模态大语言模型基于所述古籍文本特征和所述古籍图像特征,提取实体关系三元组;其中,所述实体关系三元组通过所述多模态大语言模型对所述古籍文本特征和所述古籍图像特征进行实体识别以及关系抽取后得到;

基于所述实体关系三元组构建知识图谱。

2.根据权利要求1所述的方法,其特征在于,所述对医药古籍进行文本处理,得到古籍文本特征,包括:

采集医药古籍文本中的古籍文本;

对所述古籍文本进行文本处理,得到古籍文本特征;其中,所述文本处理包括:去噪、分词和词性标注。

3.根据权利要求1所述的方法,其特征在于,所述对所述医药古籍进行图像处理,得到古籍图像特征,包括:

获取所述医药古籍对应的古籍图像;

对所述古籍图像进行图像识别,得到图像元素;

提取所述古籍图像中所述图像元素的元素特征;

根据所述元素特征对所述古籍图像进行分类,得到图像类别;

将所述元素特征和所述图像类别确定为古籍图像特征。

4.根据权利要求1所述的方法,其特征在于,所述通过预先训练好的多模态大语言模型基于所述古籍文本特征和所述古籍图像特征,提取实体关系三元组,包括:

基于预先训练的中医术语嵌入矩阵,对所述古籍文本特征的序列进行文本嵌入,得到文本特征向量;

对所述古籍图像特征的序列进行图像嵌入,得到图像特征向量;

对所述文本特征向量和所述图像特征向量进行融合,得到多模态特征向量;

基于所述多模态特征向量进行实体和关系的提取,得到实体关系三元组。

5.根据权利要求4所述的方法,其特征在于,所述基于所述多模态特征向量进行实体和关系的提取,得到实体关系三元组,包括:

根据命名实体识别技术和预设的实体集合,从所述多模态特征向量中提取上下文信息;

根据所述上下文信息确定所述医药古籍的文本和/或图表中的实体和所述实体的位置和类型;

根据预设的关系集合抽取所述医药古籍的文本和/或图表中每两个所述实体之间的关系,得到实体关系三元组。

6.根据权利要求1所述的方法,其特征在于,所述方法还包括:

采用图数据库,对所述知识图谱中的实体作为节点进行存储,对所述知识图谱中的关系作为边进行存储;

其中,所述节点对应有节点标签,所述节点标签包括:实体类型、实体属性和节点连接关系。

7.根据权利要求1所述的方法,其特征在于,所述方法还包括:

获取所述知识图谱中由多个具有连接关系的实体组成的目标路径;

根据所述知识图谱中的所述实体关系三元组确定在第一实体下第一关系的出现概率;其中,所述第一实体和所述第一关系对应于同一实体关系三元组,且所述第一实体是组成所述目标路径的多个实体中的任一实体;

根据所述目标路径上多个实体对应的出现概率,确定所述目标路径的可信度。

8.根据权利要求1所述的方法,其特征在于,所述方法还包括:

获取用户输入的待查询实体;

在所述知识图谱中查找与所述待查询实体相关联的多个候选实体和多个候选关系;

确定各所述候选实体的重要性度量值和各所述候选关系的权重;

根据各所述候选实体的重要性度量值和各所述候选关系的权重,确定各所述候选实体的推荐得分;

根据所述推荐得分在多个所述候选实体中推荐所述待查询实体对应的目标实体。

9.根据权利要求1所述的方法,其特征在于,所述多模态大语言模型的训练过程包括:

使用样本文本序列和预设的第一损失函数对所述多模态大语言模型进行训练,得到第一损失函数值;

使用样本图像特征和预设的第二损失函数对所述多模态大语言模型进行训练,得到第二损失函数值;

使用所述样本文本序列、所述样本图像特征和预设的第三损失函数对所述多模态大语言模型进行训练,得到第三损失函数值;

根据预设的超参数和所述第一损失函数值、第二损失函数值、第三损失函数值确定目标损失函数值;

根据所述目标损失函数值对所述多模态大语言模型进行训练,直至所述目标损失函数值收敛至预设值时,训练结束。

10.一种医药古籍知识图谱构建系统,其特征在于,所述系统包括以下模块:

文本处理模块:用于对医药古籍进行文本处理,得到古籍文本特征;

图像处理模块:用于对所述医药古籍进行图像处理,得到古籍图像特征;

三元组提取模块:用于通过预先训练好的多模态大语言模型基于所述古籍文本特征和所述古籍图像特征,提取实体关系三元组;其中,所述实体关系三元组通过所述多模态大语言模型对所述古籍文本特征和所述古籍图像特征进行实体识别以及关系抽取后得到;

知识图谱构建模块:用于基于所述实体关系三元组构建知识图谱。

一种医药古籍知识图谱构建方法及系统

技术领域

[0001] 本发明涉及数据处理技术领域,尤其是涉及一种医药古籍知识图谱构建方法及系统。

背景技术

[0002] 医药古籍是文化遗产的重要组成部分,随着科技的发展,知识图谱逐渐应用到医药古籍领域中,如语义搜索、智能问答、辅助决策等方面。然而,在构造知识图谱的过程中,医药古籍图文并存的表现形式以及通常存在语言晦涩、语法不规范、字词变异等问题,给从医药古籍中抽取实体关系带来了巨大的挑战,进而导致不能很好地在知识图谱中建立其中的属性、实体、关系等特性。因此,针对医学古籍亟需一种能够高效且准确地构建知识图谱的方法。

发明内容

[0003] 有鉴于此,本发明的目的在于提供一种医药古籍知识图谱构建方法及系统,该方法能够有效提升实体及其关系提取的准确性和全面性,更好的构建知识图谱,提高知识图谱的准确性和构建效率。

[0004] 第一方面,本发明实施例提供了一种医药古籍知识图谱构建方法,所述方法包括:

对医药古籍进行文本处理,得到古籍文本特征;

对所述医药古籍进行图像处理,得到古籍图像特征;

通过预先训练好的多模态大语言模型基于所述古籍文本特征和所述古籍图像特征,提取实体关系三元组;其中,实体关系三元组通过多模态大语言模型对古籍文本特征和古籍图像特征进行实体识别以及关系抽取后得到;

基于所述实体关系三元组构建知识图谱。

[0005] 在一些实施方式中,所述对医药古籍进行文本处理,得到古籍文本特征,包括:

采集医药古籍文本中的古籍文本;

对所述古籍文本进行文本处理,得到古籍文本特征;其中,所述文本处理包括:去噪、分词和词性标注。

[0006] 在一些实施方式中,所述对所述医药古籍进行图像处理,得到古籍图像特征,包括:

获取所述医药古籍对应的古籍图像;

对所述古籍图像进行图像识别,得到图像元素;

提取所述古籍图像中所述图像元素的元素特征;

根据所述元素特征对所述古籍图像进行分类,得到图像类别;

将所述元素特征和所述图像类别确定为古籍图像特征。

[0007] 在一些实施方式中,所述通过预先训练好的多模态大语言模型基于所述古籍文本特征和所述古籍图像特征,提取实体关系三元组,包括:

基于预先训练的中医术语嵌入矩阵,对所述古籍文本特征的序列进行文本嵌入,得到文本特征向量;

对所述古籍图像特征的序列进行图像嵌入,得到图像特征向量;

对所述文本特征向量和所述图像特征向量进行融合,得到多模态特征向量;

基于所述多模态特征向量进行实体和关系的提取,得到实体关系三元组。

[0008] 在一些实施方式中,所述基于所述多模态特征向量进行实体和关系的提取,得到实体关系三元组,包括:

根据命名实体识别技术和预设的实体集合,从所述多模态特征向量中提取上下文信息;

根据所述上下文信息确定所述医药古籍的文本和/或图表中的实体和所述实体的位置和类型;

根据预设的关系集合抽取所述医药古籍的文本和/或图表中每两个所述实体之间的关系,得到实体关系三元组。

[0009] 在一些实施方式中,所述方法还包括:

采用图数据库,对所述知识图谱中的实体作为节点进行存储,对所述知识图谱中的关系作为边进行存储;

其中,所述节点对应有节点标签,所述节点标签包括:实体类型、实体属性和节点连接关系。

[0010] 在一些实施方式中,所述方法还包括:

获取所述知识图谱中由多个具有连接关系的实体组成的目标路径;

根据所述知识图谱中的所述实体关系三元组确定在第一实体下第一关系的出现概率;其中,所述第一实体和所述第一关系对应于同一实体关系三元组,且所述第一实体是组成所述目标路径的多个实体中的任一实体;

根据所述目标路径上多个实体对应的出现概率,确定所述目标路径的可信度。

[0011] 在一些实施方式中,所述方法还包括:

获取用户输入的待查询实体;

在所述知识图谱中查找与所述待查询实体相关联的多个候选实体和多个候选关系;

确定各所述候选实体的重要性度量值和各所述候选关系的权重;

根据各所述候选实体的重要性度量值和各所述候选关系的权重,确定各所述候选实体的推荐得分;

根据所述推荐得分在多个所述候选实体中推荐所述待查询实体对应的目标实体。

[0012] 在一些实施方式中,所述多模态大语言模型的训练过程包括:

使用样本文本序列和预设的第一损失函数对所述多模态大语言模型进行训练,得到第一损失函数值;

使用样本图像特征和预设的第二损失函数对所述多模态大语言模型进行训练,得到第二损失函数值;

使用所述样本文本序列、所述样本图像特征和预设的第三损失函数对所述多模态大语言模型进行训练,得到第三损失函数值;

根据预设的超参数和所述第一损失函数值、第二损失函数值、第三损失函数值确定目标损失函数值;

根据所述目标损失函数值对所述多模态大语言模型进行训练,直至所述目标损失函数值收敛至预设值时,训练结束。

[0013] 第二方面,本发明实施例提供了一种医药古籍知识图谱构建系统,所述系统包括以下模块:

文本处理模块:用于对医药古籍进行文本处理,得到古籍文本特征;

图像处理模块:用于对所述医药古籍进行图像处理,得到古籍图像特征;

三元组提取模块:用于通过预先训练好的多模态大语言模型基于所述古籍文本特征和所述古籍图像特征,提取实体关系三元组;其中,实体关系三元组通过多模态大语言模型对古籍文本特征和古籍图像特征进行实体识别以及关系抽取后得到;

知识图谱构建模块:用于基于所述实体关系三元组构建知识图谱。

[0014] 第三方面,发明实施例还提供一种电子设备,包括存储器、处理器,存储器中存储有可在处理器上运行的计算机程序,其中,处理器执行计算机程序时实现上述第一方面提到的医药古籍知识图谱构建方法的步骤。

[0015] 第四方面,本发明实施例还提供一种可读存储介质,该可读存储介质上存储有计算机程序,其中,计算机程序被处理器运行时实现上述第一方面提到的医药古籍知识图谱构建方法的步骤。

[0016] 本发明实施例带来了至少以下有益效果:

本发明提供了一种医药古籍知识图谱构建方法及系统,该方案包括:对医药古籍进行文本处理,得到古籍文本特征;对医药古籍进行图像处理,得到古籍图像特征;通过预先训练好的多模态大语言模型基于古籍文本特征和古籍图像特征,提取实体关系三元组;基于实体关系三元组构建知识图谱。

[0017] 上述技术方案通过提取古籍文本特征和古籍图像特征,为实现多模态信息融合提供了数据基础;进而,通过多模态大语言模型基于古籍文本特征和古籍图像特征提取实体关系三元组,在此过程中,综合考虑文本和图像的特征,文本与图像中往往包含互补的信息,从而能够有效提升实体及其关系提取的准确性和全面性,通过多模态特征融合,可以更全面地理解古籍内容,实现对古籍文本中实体及其关系的精准识别和提取,提高实体关系三元组在文本和图像这些维度上的信息交互能力,有效处理了古籍中文本的复杂问题;同时,多模态大语言模型能够保证对关系提取的精确度,并提高提取效率。在此基础上,基于实体关系三元组能够更好的构建知识图谱,提高知识图谱的准确性和构建效率。

[0018] 本发明的其他特征和优点将在随后的说明书中阐述,或者,部分特征和优点可以从说明书推知或毫无疑义的确定,或者通过实施本发明的上述技术即可得知。

[0019] 为使本发明的上述目的、特征和优点能更明显易懂,下文特举较佳实施方式,并配合所附附图,作详细说明如下。

附图说明

[0020] 为了更清楚地说明本发明具体实施方式或现有技术中的技术方案,下面将对具体实施方式或现有技术描述中所需要使用的附图作简单地介绍,显而易见地,下面描述中的

附图是本发明的一些实施方式,对于本领域普通技术人员来讲,在不付出创造性劳动的前提下,还可以根据这些附图获得其他的附图。

[0021] 图1为本发明实施例提供的一种医药古籍知识图谱构建方法的流程图;

图2为本发明实施例提供的一种医药古籍知识图谱构建方法中智能解读与知识抽取步骤的流程图;

图3为本发明实施例提供的一种医药古籍知识图谱构建方法中知识图谱的推理步骤的流程图;

图4为本发明实施例提供的一种医药古籍知识图谱构建方法中知识图谱的智能推荐步骤的流程图;

图5为本发明实施例提供的一种医药古籍知识图谱构建系统的结构示意图;

图6为本发明实施例提供的一种电子设备的结构示意图。

[0022] 图标:

510-文本处理模块;520-图像处理模块;530-三元组提取模块;540-知识图谱构建模块;

101-处理器;102-存储器;103-总线;104-通信接口。

具体实施方式

[0023] 为使本发明实施例的目的、技术方案和优点更加清楚,下面将结合附图对本发明的技术方案进行清楚、完整地描述,显然,所描述的实施例是本发明一部分实施例,而不是全部的实施例。基于本发明中的实施例,本领域普通技术人员在没有做出创造性劳动前提下所获得的所有其他实施例,都属于本发明保护的范围。

[0024] 目前在煤炭产运销储管理中,数据的采集和填报存在人为篡改数据的可能,从而影响数据真实性,导致煤炭监管的准确性和效率都比较差。

[0025] 为了能够高效且准确地构建针对医学古籍的知识图谱,本发明实施例提供的一种医药古籍知识图谱构建方法及系统,该方法通过多模态大语言模型基于古籍文本特征和古籍图像特征提取实体关系三元组,并基于此构建知识图谱,能够有效提升实体及其关系提取的准确性和全面性,更好的构建知识图谱,提高知识图谱的准确性和构建效率。

[0026] 为便于对本实施例进行理解,首先对本发明实施例所公开的一种医药古籍知识图谱构建方法进行详细介绍,如图1所示,可以包括以下步骤。

[0027] 文本数据采集与处理步骤S101:对医药古籍进行文本处理,得到古籍文本特征。

[0028] 本实施例包括:采集医药古籍文本中的古籍文本;对古籍文本进行文本处理,得到古籍文本特征;其中,文本处理包括:去噪、分词和词性标注等。

[0029] 在具体实施例中,可以使用诸如OCR(Optical Character Recognition,光学字符识别)等技术,将医药古籍扫描件或电子版中的文字转化为计算机可识别的文本格式,得到古籍文本;在此假设古籍文本表示为$T=\{t_1,t_2,\cdots\cdots,t_n\}$,$t_n$表示古籍文本中的第n个字符。

[0030] 对提取出的古籍文本T先进行去噪处理,去噪函数$D(t_i)$可以采用基于统计的噪声滤除方法(其中,t_i表示n个字词中的任一字词),例如计算字符出现的频率,若某字符出现频率远低于正常文本字符频率且不符合中医药术语特征,则判定为噪声字符并去除,去噪后的文本可以表示为$T'=\{D(t_1),D(t_2),\cdots\cdots,D(t_n)\}$。

[0031] 对去噪后的文本T'进行分词操作,分词函数$S(t_i')$例如可以基于词典匹配与统计机器学习相结合的方式。设词典集合为Dic,对于文本t_i',若其中存在子串s在Dic中,则将s作为一个词分割出来;同时,利用隐马尔可夫模型(Hidden Markov Model,HMM)等统计模型对未在词典集合Dic中匹配到的部分字词进行分词预测。分词后的文本可以表示为T''={S(D(t_1)),S(D(t_2)),……,S(D(t_n))}。

[0032] 分词后的文本T''进行词性标注,词性标注函数P(t''$_k$)可采用条件随机场(conditional random field,CRF)模型,参照以下公式所示:

$$P(t_k'') = \arg\max_i \sum_i \lambda f(t_k'', i, l)$$

[0033] 具体而言,设特征函数为$f(t_k'', i, l)$,其中i表示词在文本中的位置,*l*表示词性标签,通过大量已标注的中医药文本数据训练CRF模型,得到模型参数λ,则标注后的文本可以表示为T'''={P(S(D(t_1))),P(S(D(t_2))),……,P(S(D(t_n)))}。

[0034] 接下来,可以通过词向量表示标注后的文本T''',得到古籍文本特征。假设古籍文本中第j个词的词向量为$\vec{w}j$,则字词t_i可表示为$\vec{t_i} = \sum_{j=1}^{k} \alpha_j \vec{w}_j$,其中$\alpha_j$为该字词$t_i$在古籍文本中的权重,可通过TF-IDF(词频-逆文档频率)等算法确定,即$\alpha_j = \frac{tf_{ij} \times \log(N / d_f j)}{\sqrt{\sum_{l=1}^{k} (tf_{il} \times \log(N / d_f l)^2}}$,其中$tf_{ij}$是字词j在文本$t_i$中的词频,$d_f j$是包含字词j的文本数量,N是文本总数。

[0035] 图像数据采集与处理S102:对医药古籍进行图像处理,得到古籍图像特征。

[0036] 在本实施例中,可以包括:首先,获取医药古籍对应的古籍图像;对古籍图像进行图像识别,得到图像元素。

[0037] 本实施例可以将古籍图像表示为I={i_1,i_2,……,i_m},对于图像识别,可以运用深度学习图像识别算法(表示为R( )),例如采用卷积神经网络。示例性的,卷积神经网络的卷积层为Conv,池化层为Pool,全连接层为FC,则对古籍图像进行图像识别的过程可表示为R(I)=FC(Pool(Conv(I)));通过图像识别可以识别出古籍图像中的插图、药材图谱和表格等图像元素。

[0038] 而后,提取古籍图像中图像元素的元素特征。具体的,可以通过卷积神经网络使用预设的图像特征提取函数(表示为F( )),对古籍图像中的图像元素进行特征提取,提取出的元素特征可以表示为X_f=F(R(I)),其中,X_f例如可以是卷积神经网络中某一中间层的特征图作为图像的元素特征的表示。

[0039] 接着,根据元素特征对古籍图像进行分类,得到图像类别;进而,将元素特征和图像类别确定为古籍图像特征。

[0040] 具体的,可以通过卷积神经网络使用预设的图像分类函数(表示为C( )),根据元素特征对古籍图像进行分类,以将古籍图像分类为不同的图像类别,图像类别可以表示为Xc=C(R(I))。

[0041] 上述分类过程可以通过构建一个多分类的神经网络模型来实现,如采用Softmax

函数作为输出层激活函数,计算图像属于各个类别的概率$p(y|I)=Softmax(W \cdot R(I)+b)$,其中,W为权重矩阵,b为偏置向量,y为类别标签,取概率最大的类别作为图像分类结果。例如,若古籍图像为药材图谱,通过特征提取可得到药材的形状、颜色等元素特征;利用分类函数根据元素特征对古籍图像进行分类,确定其所属药材类别,将该药材类别作为图像类别。

[0042] 元素特征和图像类别可以共同作为古籍图像特征,且古籍图像特征与前述古籍文本特征是共同构建多模态数据的基础,以便后续多模态大语言模型进行深度融合与分析处理,从而为中医药古籍的智能解读与知识抽取提供全面的数据支持。

[0043] 智能解读与知识抽取步骤S103:通过预先训练好的多模态大语言模型基于古籍文本特征和古籍图像特征,提取实体关系三元组;其中,实体关系三元组通过多模态大语言模型对古籍文本特征和古籍图像特征进行实体识别以及关系抽取后得到。

[0044] 知识图谱构建步骤S104:基于实体关系三元组构建知识图谱。

[0045] 在上述步骤S103中,为了使多模态大语言模型可以直接应用于智能解读与知识抽取,需要事先训练该多模态大语言模型,多模态大语言模型的参数需要经过训练得到,对多模态大语言模型进行训练的目的,是最终确定可满足要求的参数。基于此,本实施例对步骤S103展开描述之前,先给出一种多模态大语言模型的训练方法,参照以下步骤(1)-(5)所示。

[0046] (1)使用样本文本序列和预设的第一损失函数对多模态大语言模型进行训练,得到第一损失函数值。

[0047] 本实施例的目标是预测文本序列中的下一个词元,也即,使用样本文本序列和预设的第一损失函数,对多模态大语言模型的词元预测能力进行训练。

[0048] 具体的,样本文本序列可以表示为X_t,通过待训练的多模态大语言模型对样本文本序列进行预测,输出的下一个(也即第k+1个)词元的预测概率分布为:

$$P(x_{t(k+1)}|x_{t1},x_{t2},\ldots,x_{tk})=softmax(head_q \cdot W_{qm} \cdot head_k+head_v \cdot W_{vm})$$

其中,$head_q$,$head_k$,$head_v$是多头注意力机制中的查询、键、值向量,W_{qm},W_{vm}是相应的权重矩阵。

[0049] 示例性的,采用如下公式所示的交叉熵损失作为第一损失函数:

$$L_{LM}=-\sum_{k=1}^{n-1}\log P(x_{t(k+1)}|x_{t1},x_{t2},\ldots\ldots,x_{tk})$$

[0050] 通过最小化交叉熵损失函数计算第一损失函数值L_{LM}。

[0051] (2)使用样本图像特征和预设的第二损失函数对多模态大语言模型进行训练,得到第二损失函数值。

[0052] 本实施例的目标是对于样本图像特征,通过多模态大语言模型生成相应的文本描述Y_t;也即,也即,使用样本图像特征和预设的第二损失函数,对多模态大语言模型的图像描述生成任务的能力进行训练。

[0053] 具体的,样本图像特征可以表示为X_i(注意:这里的i表示图像,image),通过待训练的多模态大语言模型根据样本图像特征生成文本描述Y_t的概率分布为:

$$P(y_{tj}|y_{t1},y_{t2},\ldots,y_{t(j-1)},X_i)=softmax(head_q' \cdot W_{qdm} \cdot head_k'+head_v' \cdot W_{vdm})$$

其中,$head_q'$,$head_k'$,$head_v'$是用于图像描述生成任务的多头注意力机制中的向量,W_{qdm},W_{vdm}是相应权重矩阵。

[0054] 示例性的,采用如下公式所示的负对数似然损失函数作为第二损失函数,并计算第二损失函数值L_{IC}:

$$L_{IC}=-\sum_{j=1}^{m}\log P(y_{tj}\mid y_{t1},y_{t2},\ldots\ldots,y_{t(j-1)},X_i)$$

[0055] (3)使用样本文本序列、样本图像特征和预设的第三损失函数对多模态大语言模型进行训练,得到第三损失函数值。

[0056] 本实施例的目标是实现跨模态匹配,也即,使用样本文本序列、样本图像特征和预设的第三损失函数,对多模态大语言模型的跨模态匹配的判断能力进行训练。

[0057] 具体的,通过待训练的多模态大语言模型,判断样本文本序列X_t和样本图像特征X_i是否匹配。在判断过程中,通过一个二元分类器,输出匹配概率p_m:

$p_m=\mathrm{sigmoid}(\mathrm{MLP}([F(X_t,X_i)]))$

其中,MLP是多层感知机。

[0058] 示例性的,采用如下公式所示的二元交叉熵作为第三损失函数,并计算第三损失函数值L_{CM}:

$L_{CM}=-[y_m\log p_m+(1-y_m)\log(1-p_m)]$

其中,y_m是真实的匹配标签。

[0059] (4)根据预设的超参数和第一损失函数值、第二损失函数值、第三损失函数值确定目标损失函数值。

[0060] 具体的,在模型训练过程中,根据预设的超参数和第一损失函数值、第二损失函数值、第三损失函数值确定总的损失函数为:

$L=\alpha L_{LM}+\beta L_{IC}+\gamma L_{CM}$

其中α,β,γ是平衡不同预训练任务的超参数,L为目标损失函数值。

[0061] (5)根据目标损失函数值对多模态大语言模型进行训练,直至目标损失函数值收敛至预设值时,训练结束。

[0062] 至此,完成多模态大语言模型的训练过程。

[0063] 在以上实施例的基础上,本实施例对上述智能解读与知识抽取步骤S103展开描述,如图2所示,可以包括以下步骤。

[0064] 步骤S201,基于预先训练的中医术语嵌入矩阵,对古籍文本特征的序列进行文本嵌入,得到文本特征向量。

[0065] 步骤S202,对古籍图像特征的序列进行图像嵌入,得到图像特征向量。

[0066] 在具体实施例中,多模态大语言模型例如可以是基于Transformer的编码器-解码器结构。在编码器部分,输入古籍文本特征的序列$Yt=[y_{t1},y_{t2},\ldots,y_{tn}]$(其中,$y_{ti}$表示文本中的第i个词元)和古籍图像特征的序列$Y_i=[y_{i1},y_{i2},\ldots,y_{im}]$(其中,$y_{ij}$表示图像的第j个特征向量,这里的i表示图像,image)。

[0067] 在编码器中,首先通过特定的输入嵌入层对古籍文本特征的序列进行文本嵌入,以及对古籍图像特征的序列进行图像嵌入。

[0068] 其中,文本嵌入可以包括:$Et(y_{ti})=W_t\cdot y_{ti}+bt$。其中,$W_t$是文本嵌入矩阵,$b_t$是偏置项,$Et(y_{ti})$是文本特征向量。

[0069] 图像嵌入可以包括:$Ei(y\{ij\})=W_i\cdot y_{ij}+bi$。其中,$W_i$是图像嵌入矩阵,$b_i$是偏置项,

Ei(y{ij})是图像特征向量。

[0070] 在一些实施例中,还可以在多模态大语言模型的嵌入层或中间层引入中医药领域知识。例如,对于中医术语T=[t_1,t_2,...,t_p],可以预先训练一个中医术语嵌入矩阵W_{ct},在此情况下,则可以参照以下公式对古籍文本特征的序列进行文本嵌入:

$$E_t(y_{ti})=W_t \cdot y_{ti}+b_t+W_{ct} \cdot e(t_i)$$

其中,$e(t_i)$是中医术语t_i的特定嵌入表示(可以是独热编码或其他预训练的向量表示),$E_t(y_{ti})$是文本特征向量。该文本嵌入方式可以使多模态大语言模型在学习过程中更好地理解中医药领域的专业术语和概念,提高多模态大语言模型在中医药古籍处理中的准确性和专业性。

[0071] 步骤S203,对文本特征向量和图像特征向量进行融合,得到多模态特征向量。

[0072] 在本实施例中,通过融合层将嵌入后的文本特征向量和图像特征向量进行融合。一种示例融合方式可以为:将文本特征向量和图像特征向量进行拼接,再对拼接后向量进行线性变换,例如:

$$F(Y_t,Y_i)=W_f \cdot [E_t(Y_t);E_i(Y_i)]+b_f$$

其中,$F(Y_t,Y_i)$是多模态特征向量,$E_t(Y_t)$是文本特征向量,$E_i(Y_i)$是图像特征向量,W_f是融合层的权重矩阵,b_f是偏置项,[;]表示拼接操作。

[0073] 另一种示例融合方式可以为:利用注意力机制进行跨模态融合,在此过程中,假设文本特征向量对图像特征向量的注意力权重为β_{ij},图像特征向量对文本特征向量的注意力权重为γ_{ij},则融合后的多模态特征向量$\vec{m}_{ij}$可表示为:

$$\vec{m}_{ij}=\beta_{ij}\vec{t}_i+\gamma_{ij}\vec{f}_j$$

[0074] 其中,$\vec{t}_i$表示字词t_i的文本特征向量,$\vec{f}_j$表示古籍图像的第j个图像特征向量,$\beta_{ij}=\frac{\exp(e_{ij})}{\sum_{s=1}^{n}\exp(e_{is})}$,$ei_j=a(\vec{t}_i,\vec{f}_j)$($a$为注意力计算函数,例如可采用点积注意力函数$a(\vec{x},\vec{y})=\vec{x}^T\vec{y}$)。

[0075] 步骤S204,基于多模态特征向量进行实体和关系的提取,得到实体关系三元组。本实施例可以包括:

根据命名实体识别技术和预设的实体集合,从多模态特征向量中提取上下文信息;根据上下文信息确定医药古籍的文本和/或图表中的实体和实体的位置和类型;根据预设的关系集合抽取医药古籍的文本和/或图表中每两个实体之间的关系,得到实体关系三元组。

[0076] 具体而言,对于知识抽取,假设中医术语、病症、药材等实体集合表示为E={e_1,e_2,……,e_q},关系集合为R={r_1,r_2,……,r_s}。通过命名实体识别技术确定医药古籍的文本和/或图表中实体出现的位置和类型,构建实体识别函数。其中,对于文本而言,实体识别函数可以表示为$\varphi(t_i)$,对于图表而言,实体识别函数可以表示为$\varphi(i_j)$。以$\varphi(t_i)$为例,使用实体识别函数$\varphi(t_i)$进行实体识别,若文本和/或图表中存在实体e_k,则$\varphi(t_i)(e_k)=1$,否则为0。

[0077] 对于关系抽取,本实施例可以基于规则或深度学习模型,假设关系抽取函数为

$\psi(t_i,e_{k1},e_{k2})$，若在文本t_i中实体e_{k1}和e_{k2}存在关系r_1，则$\psi(t_i,e_{k1},e_{k2})=r_l$。

[0078] 按照以上实施例抽取医药古籍的文本和/或图表中的实体以及每两个实体之间的关系，由此得到实体关系三元组。

[0079] 进而，针对步骤S104，在基于实体关系三元组构建知识图谱时，可以以实体为节点、以关系为边，构建知识图谱，知识图谱可表示为G=(E,R)，其中，节点的属性可包含实体的相关信息(如药材的产地、性味归经等)。

[0080] 在知识图谱中，设实体集合为E，关系集合为R。对于一个特定的实体$e_i\in E$(例如某一中药材实体“人参”)，可以用向量e_i来表示它的语义特征。假设实体向量维度为d，则$e_i=[e_{i1},e_{i2},\cdots\cdots,e_{id}]$，其中$e_{ij}$表示该实体在第$_j$个维度上的特征值。

[0081] 对于关系$r_k\in R$(如“具有功效”关系)，可以用向量r_k表示，$r_k=[r_{k1},r_{k2},\cdots\cdots,r_{kd}]$。

[0082] 当存在一个三元组(e_i,r_k,e_j)表示实体e_i通过关系r_k与实体e_j相连时(例如“人参，具有功效，大补元气”)，在知识图谱的表示学习中，常使用基于能量函数的模型，如TransE模型，其基本公式为：$e_i+e_k\approx e_j$，即通过最小化能量函数$f(e_i,r_k,e_j)=||e_i+e_k-e_j||$来学习实体和关系的向量表示，其中$||\cdot||$表示向量的范数(如L2范数)。

[0083] 本实施例通过多模态的知识图谱将单一文字形式的节点扩展为包含数字图像、文本等多模态数据的方式对知识进行呈现，多模态数据通过与文本的对应实体、关系及属性进行融合，完成多模态知识图谱的构建。

[0084] 本实施例通过构建知识图谱，能够结构化表示实现对中医药古籍知识的有效存储与管理，以便后续进行高效的检索与智能推荐服务，在语义搜索、知识问答和临床决策支持等智慧医疗领域做出更好地发展。具体例如，基于知识图谱的查询可表示为Q(E,R,condition)，其中condition为查询条件，如查询某种病症相关的药材治疗方法等。

[0085] 在构建知识图谱后，本实施例提供的方法还可以进一步包括知识图谱的存储与查询、基于知识图谱的推理与智能推荐等内容，参照以下实施例。

[0086] 在本实施例中，知识图谱的存储方法可以包括：采用图数据库，对知识图谱中的实体作为节点进行存储，对知识图谱中的关系作为边进行存储；其中，节点对应有节点标签，节点标签包括：实体类型、实体属性和节点连接关系。

[0087] 以采用图数据库(如Neo4j)存储知识图谱作为示例，可以将实体作为节点，关系作为边进行存储。具体的，图数据库中的节点集合为N，边集合为L。对于一个节点$n_i\in N$对应一个实体e_i，其存储结构可能包含节点标签，该节点标签包括诸如包括：实体类型(如“药材”“病症”等)，实体属性(如实体的名称、产地等)以及与其他节点之间的节点连接关系。

[0088] 在本实施例中，在知识图谱的查询方面，例如查询具有某种功效的药材，可表示为：MATCH(e_1:药材)-[r:具有功效]->(e_2:功效名称:"特定功效")RETURNe_1。其中，MATCH子句用于指定查询的图模式，即寻找从类型为“药材”的节点e_1通过“具有功效”关系r连接到名称为“特定功效”的类型为“功效”的节点e_2，RETURN子句指定返回符合条件的药材节点e1。

[0089] 参照图3，在本实施例中，知识图谱的推理方法可以包括：

步骤S301：获取知识图谱中由多个具有连接关系的实体组成的目标路径。

[0090] 在知识图谱的推理方面，本实施例可以利用知识图谱中的路径推理。假设$P(e_i,e_j)$

表示从实体e_i到实体e_j的一条目标路径,目标路径由一系列关系连接的实体组成,如$P(e_i, e_j)=(e_i, r_1, e_{i+1}, r_2, \cdots, e_j)$。

[0091] 步骤S302:根据知识图谱中的实体关系三元组确定在第一实体下第一关系的出现概率;其中,第一实体和第一关系对应于同一实体关系三元组,且第一实体是组成目标路径的多个实体中的任一实体。

[0092] 例如,其中,通过统计知识图谱中相关的实体关系三元组数量,来估计在第一实体e_m下第一关系r_k的出现概率,该出现概率可以表示为$p(r_k|e_m)$。

[0093] 步骤S303:根据目标路径上多个实体对应的出现概率,确定目标路径的可信度。

[0094] 推理的可信度可以通过目标路径上关系的组合来计算,具体如可以采用基于概率的方法,按照以下公式确定目标路径P的可信度:

$$p(P)=p(r_1|e_i)\times p(r_2|e_{i+1})\times\cdots\cdots$$

其中,p(P)表示目标路径P的可信度。

[0095] 参照图4,在本实施例中,知识图谱的智能推荐方法可以包括:

步骤S401:获取用户输入的待查询实体。

[0096] 步骤S402:在知识图谱中查找与待查询实体相关联的多个候选实体和多个候选关系。

[0097] 对于智能推荐,可以向用户推荐与某种病症相关的药材治疗方案。获取用户输入的待查询实体,例如用户输入的待查询实体为病症实体e_s。系统首先在知识图谱中找到与待查询实体e_s相关的所有候选关系和候选实体。

[0098] 步骤S403:确定各候选实体的重要性度量值和各候选关系的权重。

[0099] 步骤S404:根据各候选实体的重要性度量值和各候选关系的权重,确定各候选实体的推荐得分。

[0100] 其中,候选实体e_i的推荐得分$S(e_i)$可以表示为:

$$S(e_i)=\sum_{r_k,e_j} w_{rk}\times f(e_s,r_k,e_i)\times g(e_i)$$

[0101] 其中,w_{rk}是关系r_k的权重,$f(e_s, r_k, e_i)$是基于前面提到的实体与关系表示计算的关联度,例如使用TransE模型中的能量函数值的倒数或其他相似性度量,$g(e_i)$表示候选实体e_i的重要性度量值(具体如连接度、引用频率等的归一化值)。

[0102] 步骤S405:根据推荐得分在多个候选实体中推荐待查询实体对应的目标实体。

[0103] 根据推荐得分$S(e_i)$由高到低对候选实体进行排序,将排列在首位或前N位的候选实体作为目标实体推荐给用户。

[0104] 综上所述,从上述实施例中提到的医药古籍知识图谱构建方法可知,该方法包括:对医药古籍进行文本处理,得到古籍文本特征;对医药古籍进行图像处理,得到古籍图像特征;通过预先训练好的多模态大语言模型基于古籍文本特征和古籍图像特征,提取实体关系三元组;基于实体关系三元组构建知识图谱。

[0105] 上述技术方案通过提取古籍文本特征和古籍图像特征,为实现多模态信息融合提供了数据基础;进而,通过多模态大语言模型基于古籍文本特征和古籍图像特征提取实体关系三元组,在此过程中,综合考虑文本和图像的特征,文本与图像中往往包含互补的信息,从而能够有效提升实体及其关系提取的准确性和全面性,通过多模态特征融合,可以更

全面地理解古籍内容,实现对古籍文本中实体及其关系的精准识别和提取,提高实体关系三元组在文本和图像这些维度上的信息交互能力,有效处理了古籍中文本的复杂问题;同时,多模态大语言模型能够保证对关系提取的精确度,并提高提取效率。在此基础上,基于实体关系三元组能够更好的构建知识图谱,提高知识图谱的准确性和构建效率。

[0106] 此外,在实际应用中,本发明在一定程度上有助于推动智慧医学建设和中医药文化的传承。

[0107] 对应于上述方法实施例,本发明实施例提供了一种医药古籍知识图谱构建系统,如图5所示,该系统包括以下模块:

文本处理模块510:用于对医药古籍进行文本处理,得到古籍文本特征;

图像处理模块520:用于对所述医药古籍进行图像处理,得到古籍图像特征;

三元组提取模块530:用于通过预先训练好的多模态大语言模型基于所述古籍文本特征和所述古籍图像特征,提取实体关系三元组;其中,实体关系三元组通过多模态大语言模型对古籍文本特征和古籍图像特征进行实体识别以及关系抽取后得到;

知识图谱构建模块540:用于基于所述实体关系三元组构建知识图谱。

[0108] 在一些实施方式中,文本处理模块510,用于:

采集医药古籍文本中的古籍文本;

对所述古籍文本进行文本处理,得到古籍文本特征;其中,所述文本处理包括:去噪、分词和词性标注。

[0109] 在一些实施方式中,图像处理模块520,用于:

获取所述医药古籍对应的古籍图像;

对所述古籍图像进行图像识别,得到图像元素;

提取所述古籍图像中所述图像元素的元素特征;

根据所述元素特征对所述古籍图像进行分类,得到图像类别;

将所述元素特征和所述图像类别确定为古籍图像特征。

[0110] 在一些实施方式中,三元组提取模块530,用于:

基于预先训练的中医术语嵌入矩阵,对所述古籍文本特征的序列进行文本嵌入,得到文本特征向量;

对所述古籍图像特征的序列进行图像嵌入,得到图像特征向量;

对所述文本特征向量和所述图像特征向量进行融合,得到多模态特征向量;

基于所述多模态特征向量进行实体和关系的提取,得到实体关系三元组。

[0111] 在一些实施方式中,三元组提取模块530,用于:

根据命名实体识别技术和预设的实体集合,从所述多模态特征向量中提取上下文信息;

根据所述上下文信息确定所述医药古籍的文本和/或图表中的实体和所述实体的位置和类型;

根据预设的关系集合抽取所述医药古籍的文本和/或图表中每两个所述实体之间的关系,得到实体关系三元组。

[0112] 在一些实施方式中,医药古籍知识图谱构建系统还包括知识图谱存储模块,其用于:

采用图数据库,对所述知识图谱中的实体作为节点进行存储,对所述知识图谱中的关系作为边进行存储;

其中,所述节点对应有节点标签,所述节点标签包括:实体类型、实体属性和节点连接关系。

[0113] 在一些实施方式中,医药古籍知识图谱构建系统还包括路径推理模块,其用于:

获取所述知识图谱中由多个具有连接关系的实体组成的目标路径;

根据所述知识图谱中的所述实体关系三元组确定在第一实体下第一关系的出现概率;其中,所述第一实体和所述第一关系对应于同一实体关系三元组,且所述第一实体是组成所述目标路径的多个实体中的任一实体;

根据所述目标路径上多个实体对应的出现概率,确定所述目标路径的可信度。

[0114] 在一些实施方式中,医药古籍知识图谱构建系统还包括智能推荐模块,其用于:

获取用户输入的待查询实体;

在所述知识图谱中查找与所述待查询实体相关联的多个候选实体和多个候选关系;

确定各所述候选实体的重要性度量值和各所述候选关系的权重;

根据各所述候选实体的重要性度量值和各所述候选关系的权重,确定各所述候选实体的推荐得分;

根据所述推荐得分在多个所述候选实体中推荐所述待查询实体对应的目标实体。

[0115] 在一些实施方式中,医药古籍知识图谱构建系统还包括模型训练模块,其用于:

使用样本文本序列和预设的第一损失函数对所述多模态大语言模型进行训练,得到第一损失函数值;

使用样本图像特征和预设的第二损失函数对所述多模态大语言模型进行训练,得到第二损失函数值;

使用所述样本文本序列、所述样本图像特征和预设的第三损失函数对所述多模态大语言模型进行训练,得到第三损失函数值;

根据预设的超参数和所述第一损失函数值、第二损失函数值、第三损失函数值确定目标损失函数值;

根据所述目标损失函数值对所述多模态大语言模型进行训练,直至所述目标损失函数值收敛至预设值时,训练结束。

[0116] 通过上述实施例中提到的医药古籍知识图谱构建系统可知,该系统通过多模态大语言模型基于古籍文本特征和古籍图像特征提取实体关系三元组,在此过程中,综合考虑文本和图像的特征,文本与图像中往往包含互补的信息,从而能够有效提升实体及其关系提取的准确性和全面性,通过多模态特征融合,可以更全面地理解古籍内容,实现对古籍文本中实体及其关系的精准识别和提取,提高实体关系三元组在文本和图像这些维度上的信息交互能力,有效处理了古籍中文本的复杂问题;同时,多模态大语言模型能够保证对关系提取的精确度,并提高提取效率。在此基础上,基于实体关系三元组能够更好的构建知识图谱,提高知识图谱的准确性和构建效率。

[0117] 本实施例提供的医药古籍知识图谱构建系统,与上述实施例提供的医药古籍知识图谱构建方法具有相同的技术特征,所以也能解决相同的技术问题,达到相同的技术效果。

为简要描述,实施例部分未提及之处,可参考前述医药古籍知识图谱构建方法实施例中相应内容。

[0118] 本实施例还提供一种电子设备,该电子设备的结构示意图如图6所示,该设备包括处理器101和存储器102;其中,存储器102用于存储一条或多条计算机指令,一条或多条计算机指令被处理器执行,以实现上述医药古籍知识图谱构建方法。

[0119] 图6所示的电子设备还包括总线103和通信接口104,处理器101、通信接口104和存储器102通过总线103连接。

[0120] 其中,存储器102可能包含高速随机存取存储器(RAM,Random Access Memory),也可能还包括非不稳定的存储器(non-volatile memory),例如至少一个磁盘存储器。总线103可以是ISA总线、PCI总线或EISA总线等。所述总线可以分为地址总线、数据总线、控制总线等。为便于表示,图6中仅用一个双向箭头表示,但并不表示仅有一根总线或一种类型的总线。

[0121] 通信接口104用于通过网络接口与至少一个用户终端及其它网络单元连接,将封装好的IPv4报文或IPv4报文通过网络接口发送至用户终端。

[0122] 处理器101可能是一种集成电路芯片,具有信号的处理能力。在实现过程中,上述方法的各步骤可以通过处理器101中的硬件的集成逻辑电路或者软件形式的指令完成。上述的处理器101可以是通用处理器,包括中央处理器(Central Processing Unit,简称CPU)、网络处理器(Network Processor,简称NP)等;还可以是数字信号处理器(Digital Signal Processor,简称DSP)、专用集成电路(Application Specific Integrated Circuit,简称ASIC)、现场可编程门阵列(Field-Programmable Gate Array,简称FPGA)或者其他可编程逻辑器件、分立门或者晶体管逻辑器件、分立硬件组件。可以实现或者执行本公开实施例中的公开的各方法、步骤及逻辑框图。通用处理器可以是微处理器或者该处理器也可以是任何常规的处理器等。结合本公开实施例所公开的方法的步骤可以直接体现为硬件译码处理器执行完成,或者用译码处理器中的硬件及软件模块组合执行完成。软件模块可以位于随机存储器,闪存、只读存储器,可编程只读存储器或者电可擦写可编程存储器、寄存器等本领域成熟的存储介质中。该存储介质位于存储器102,处理器101读取存储器102中的信息,结合其硬件完成前述实施例的方法的步骤。

[0123] 本发明实施例还提供了一种可读存储介质,该可读存储介质上存储有计算机程序,该计算机程序被处理器运行时执行前述实施例的医药古籍知识图谱构建方法的步骤。

[0124] 在本申请所提供的几个实施例中,应该理解到,所揭露的系统、设备和方法,可以通过其它的方式实现。以上所描述的装置实施例仅仅是示意性的,例如,所述单元的划分,仅仅为一种逻辑功能划分,实际实现时可以有另外的划分方式,又例如,多个单元或组件可以结合或者可以集成到另一个系统,或一些特征可以忽略,或不执行。另一点,所显示或讨论的相互之间的耦合或直接耦合或通信连接可以是通过一些通信接口,设备或单元的间接耦合或通信连接,可以是电性,机械或其它的形式。

[0125] 所述作为分离部件说明的单元可以是或者也可以不是物理上分开的,作为单元显示的部件可以是或者也可以不是物理单元,即可以位于一个地方,或者也可以分布到多个网络单元上。可以根据实际的需要选择其中的部分或者全部单元来实现本实施例方案的目的。

[0126] 另外,在本发明各个实施例中的各功能单元可以集成在一个处理单元中,也可以是各个单元单独物理存在,也可以两个或两个以上单元集成在一个单元中。

[0127] 所述功能如果以软件功能单元的形式实现并作为独立的产品销售或使用时,可以存储在一个处理器可执行的非易失的计算机可读取存储介质中。基于这样的理解,本发明的技术方案本质上或者说对现有技术做出贡献的部分或者该技术方案的部分可以用软件产品的形式体现出来,该计算机软件产品存储在一个存储介质中,包括若干指令用以使得一台计算机设备(可以是个人计算机,服务器,或者网络设备等)执行本发明各个实施例所述方法的全部或部分步骤。而前述的存储介质包括:U盘、移动硬盘、只读存储器(ROM,Read-Only Memory)、随机存取存储器(RAM,Random Access Memory)、磁碟或者光盘等各种可以存储程序代码的介质。

[0128] 最后应说明的是:以上所述实施例,仅为本发明的具体实施方式,用以说明本发明的技术方案,而非对其限制,本发明的保护范围并不局限于此,尽管参照前述实施例对本发明进行了详细的说明,本领域的普通技术人员应当理解:任何熟悉本技术领域的技术人员在本发明揭露的技术范围内,其依然可以对前述实施例所记载的技术方案进行修改或可轻易想到变化,或者对其中部分技术特征进行等同替换;而这些修改、变化或者替换,并不使相应技术方案的本质脱离本发明实施例技术方案的精神和范围,都应涵盖在本发明的保护范围之内。因此,本发明的保护范围应所述以权利要求的保护范围为准。

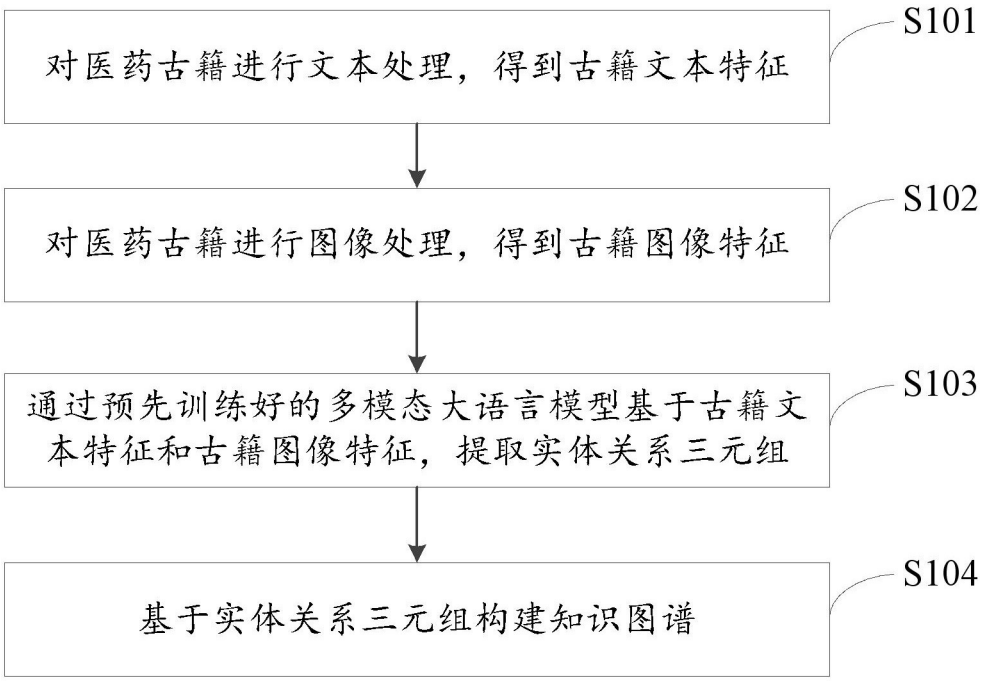

对医药古籍进行文本处理，得到古籍文本特征
S101
对医药古籍进行图像处理，得到古籍图像特征
S102
通过预先训练好的多模态大语言模型基于古籍文本特征和古籍图像特征，提取实体关系三元组
S103
基于实体关系三元组构建知识图谱
S104

图1

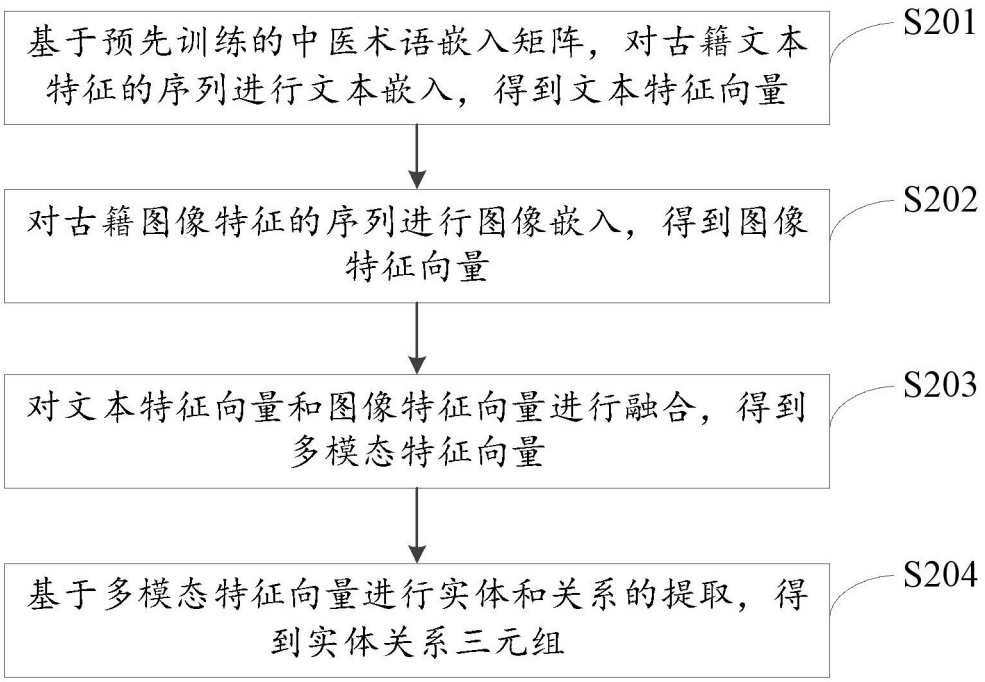

基于预先训练的中医术语嵌入矩阵，对古籍文本特征的序列进行文本嵌入，得到文本特征向量
S201
对古籍图像特征的序列进行图像嵌入，得到图像特征向量
S202
对文本特征向量和图像特征向量进行融合，得到多模态特征向量
S203
基于多模态特征向量进行实体和关系的提取，得到实体关系三元组
S204

图2

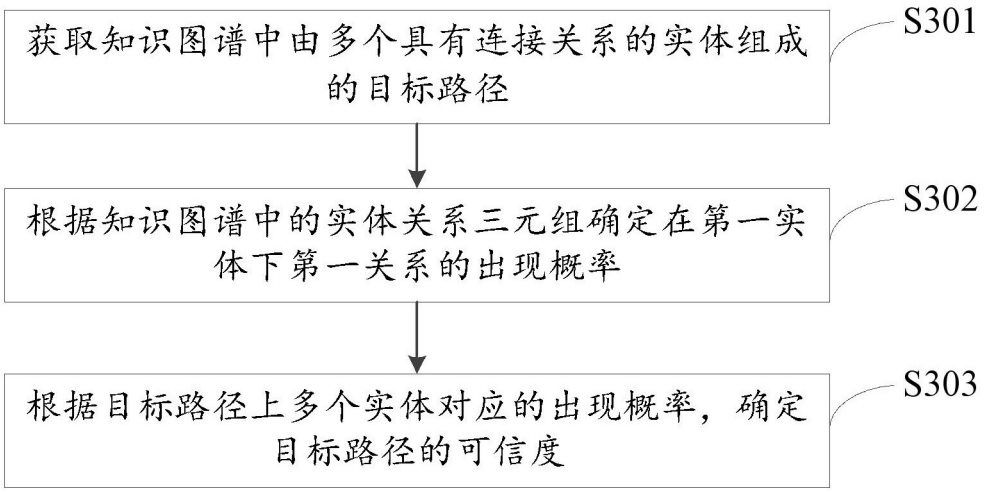

获取知识图谱中由多个具有连接关系的实体组成的目标路径
S301
根据知识图谱中的实体关系三元组确定在第一实体下第一关系的出现概率
S302
根据目标路径上多个实体对应的出现概率，确定目标路径的可信度
S303

图3

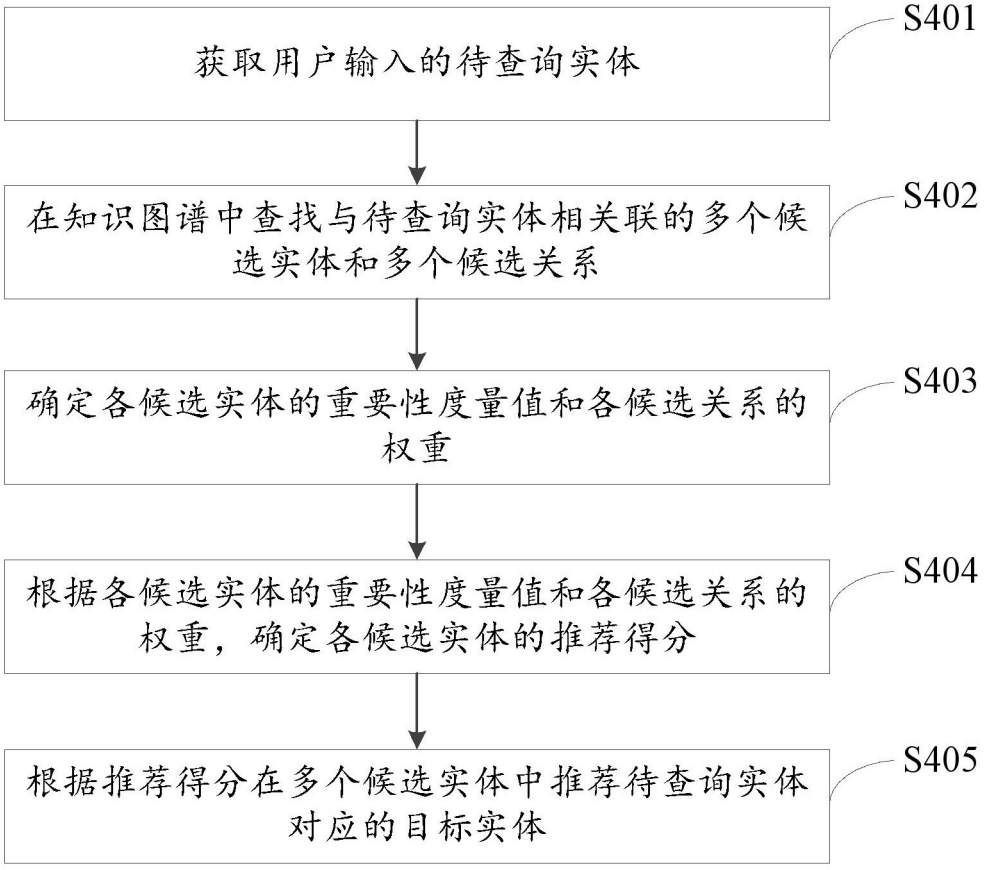

获取用户输入的待查询实体
S401
在知识图谱中查找与待查询实体相关联的多个候选实体和多个候选关系
S402
确定各候选实体的重要性度量值和各候选关系的权重
S403
根据各候选实体的重要性度量值和各候选关系的权重，确定各候选实体的推荐得分
S404
根据推荐得分在多个候选实体中推荐待查询实体对应的目标实体
S405

图4


文本处理模块
510
图像处理模块
520
三元组提取模块
530
知识图谱构建模块
540

图5

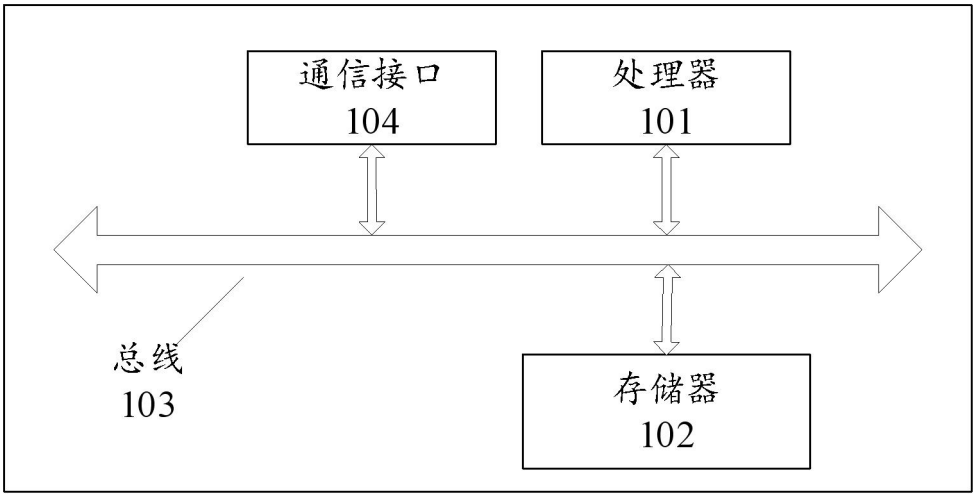

通信接口
104
处理器
101
总线
103
存储器
102

图6