(19)中华人民共和国国家知识产权局

(12)发明专利

(10)授权公告号 CN 112632226 B
(45)授权公告日 2021.10.26

(21)申请号 202011597968.9

(22)申请日 2020.12.29

(65)同一申请的已公布的文献号
申请公布号 CN 112632226 A

(43)申请公布日 2021.04.09

(73)专利权人 天津汇智星源信息技术有限公司
地址 300384 天津市滨海新区华苑产业区开华道22号5号楼西塔2001-2008室

(72)发明人 朵思惟 余梓飞 于锋杰 薛晨云

(74)专利代理机构 北京风雅颂专利代理有限公司 11403
代理人 孙晓凤

(51)Int.Cl.
G06F 16/33(2019.01)
G06F 16/35(2019.01)
G06F 16/36(2019.01)
G06F 40/211(2020.01)
G06F 40/216(2020.01)
G06F 40/295(2020.01)
G06F 40/30(2020.01)
G06K 9/62(2006.01)
G06N 3/04(2006.01)
G06N 3/08(2006.01)

(56)对比文件
CN 111813916 A,2020.10.23
CN 111414465 A,2020.07.14
CN 108052619 A,2018.05.18
CN 103488724 A,2014.01.01
US 2018276284 A1,2018.09.27
US 2020364619 A1,2020.11.19
CN 111813916 A,2020.10.23
陈金菊.“基于道路法规知识图谱的多轮自动问答研究”.《现代情报》.2020,
L. Ma 等.“Answer Graph-based Interactive Attention Network for Question Answering over Knowledge Base”.《2020 IEEE Intl Conf on Parallel & Distributed Processing with Applications》.2020,

审查员 李小敏

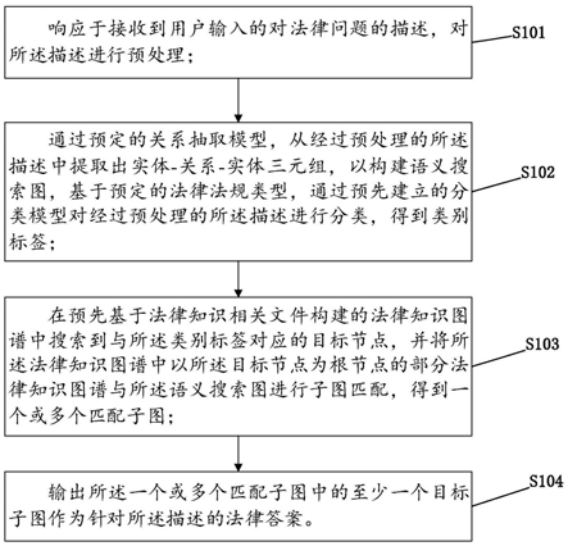
权利要求书2页 说明书10页 附图5页

(54)发明名称

基于法律知识图谱的语义搜索方法、装置和电子设备

(57)摘要

本说明书一个或多个实施例提供一种基于法律知识图谱的语义搜索方法、装置和电子设备。响应于接收到用户输入的法律问题的描述，对所述描述进行预处理；对经过所述预处理的所述描述进行要素提取，所述要素提取包括实体-关系-实体三元组的抽取，根据所述实体-关系-实体三元组构建语义搜索图，对所述语义搜索图基于法律法规类型建立类别标签，输出带有标签的语义搜索图；将所述语义搜索图与法律知识图谱进行子图匹配，将匹配度高的子图作为法律答案数据。本发明通过对用户问题建立语义网，并结合法律知识图谱进行匹配和推理，能够准确捕捉用户的搜索意图，从而直接给出满足用户搜索意图的答案，实际解决用户的法律问题。



CN 112632226 B

1.一种基于法律知识图谱的语义搜索方法,其特征在于,包括:

响应于接收到用户输入的对法律问题的描述,对所述描述进行预处理;

通过预定的关系抽取模型,从经过预处理的所述描述中提取出实体-关系-实体三元组,以构建语义搜索图;

基于预定的法律法规类型,通过预先建立的分类模型对经过预处理的所述描述进行分类,得到类别标签;

在预先基于法律知识相关文件构建的法律知识图谱中搜索到与所述类别标签对应的目标节点,并将所述法律知识图谱中以所述目标节点为根节点的部分法律知识图谱与所述语义搜索图进行子图匹配,得到一个或多个匹配子图,在所述语义搜索图中给定一个节点v,在所述部分法律知识图谱中找到对应的节点u,计算节点v和节点u的相似度,

在所述语义搜索图中给定一个关系rel,在所述部分法律知识图谱中找到对应的关系L,计算所述关系rel与所述关系L的相似度,

通过所述节点相似度和所述关系相似度计算所述语义搜索图和所述部分法律知识图谱的相似度得分,根据所述相似度得分输出匹配子图;

输出所述一个或多个匹配子图中的至少一个目标子图作为针对所述描述的法律答案。

2.根据权利要求1所述的语义搜索方法,其特征在于,所述关系抽取模型包括CASREL模型。

3.根据权利要求1或2所述的语义搜索方法,其特征在于,所述响应于接收到用户输入的对法律问题的描述,对所述描述进行预处理,包括下列中至少一个:

通过正则表达式去除所述描述的标点符号和/或特殊符号;

通过中文分词算法WMSeg对所述描述进行分词;

将所述描述中的繁简字体进行统一化;

通过标准表达方式对所述描述进行同义词归一化;

通过Soft-Masked BERT模型对所述描述进行文本纠错处理。

4.根据权利要求1或2所述的语义搜索方法,其特征在于,所述通过预定的关系抽取模型,从经过预处理的所述描述中提取出实体-关系-实体三元组,以构建语义搜索图,包括:

基于完整的训练模型BERT对所述描述进行编码,获取所述描述中每个词的特征表示;

对所述特征表示进行解码,构建分类器预测所述实体的 位置,识别出所述实体对应的主语;

根据所述主语提取所有可能与所述主语对应的关系,并根据所述关系识别出相应的宾语,得到实体-关系-实体三元组。

5.根据权利要求1或2所述的语义搜索方法,其特征在于,所述基于预定的法律法规类型,通过预先建立的分类模型对经过预处理的所述描述进行分类,得到类别标签,包括:

基于预训练模型BERT对所述描述进行特征向量表示得到特征向量,将所述特征向量输入到Softmax回归模型基于法律法规类型进行分类,得到所述描述的类别标签。

6.根据权利要求1所述的语义搜索方法,其特征在于,根据所述相似度得分对所述匹配子图进行排序,选取得分最高的预定数量个所述匹配子图作为目标子图。

7.一种基于知识图谱的语义搜索装置,其特征在于,包括:

预处理模块,被配置为响应于接收到用户输入的对法律问题的描述,对所述描述进行

预处理；

语义搜索图生成模块，被配置为通过预定的关系抽取模型，从经过预处理的所述描述中提取出实体-关系-实体三元组，以构建语义搜索图，基于预定的法律法规类型，通过预先建立的分类模型对经过预处理的所述描述进行分类，得到类别标签；

子图匹配模块，被配置为在预先基于法律知识相关文件构建的法律知识图谱中搜索到与所述类别标签对应的目标节点，并将所述法律知识图谱中以所述目标节点为根节点的部分法律知识图谱与所述语义搜索图进行子图匹配，得到一个或多个匹配子图，在所述语义搜索图中给定一个节点v，在所述部分法律知识图谱中找到对应的节点u，计算节点v和节点u的相似度，

在所述语义搜索图中给定一个关系rel，在所述部分法律知识图谱中找到对应的关系L，计算所述关系rel与所述关系L的相似度，

通过所述节点相似度和所述关系相似度计算所述语义搜索图和所述部分法律知识图谱的相似度得分，根据所述相似度得分输出匹配子图；

输出模块，被配置为输出所述一个或多个匹配子图中的至少一个目标子图作为针对所述描述的法律答案。

8.一种电子设备，包括存储器、处理器及存储在所述存储器上并可由所述处理器执行的计算机程序，其特征在于，所述处理器执行所述计算机程序时实现如权利要求1至6中任意一项所述的方法。

9.一种非暂态计算机可读存储介质，其特征在于，所述非暂态计算机可读存储介质存储计算机指令，所述计算机指令在被计算机执行时，使所述计算机实现根据权利要求1至6中任意一项所述的方法。

基于法律知识图谱的语义搜索方法、装置和电子设备

技术领域

[0001] 本说明书一个或多个实施例涉及知识图谱技术领域，尤其涉及一种基于法律知识图谱的语义搜索方法、装置和电子设备。

背景技术

[0002] 随着科技的发展，自动化的法律咨询服务在缓解人工法律服务资源不足的问题上发挥着越来越重要的作用。在民众进行法律咨询的过程中，一个高效精准的法律搜索系统可以为民众提供精准、全面的一站式解决方案。

[0003] 传统的法律搜索系统大多基于检索提问式关键词匹配技术和排序算法，返回的结果主要依据素材中是否存在关键词，无法获知用户搜索语句的真正含义。这往往与用户对结果精准、即得的需求相矛盾。比如当搜索“民事案件类型有哪些?”时，传统搜索系统呈现的是包含关键词“民事”、“案件”等的信息，而用户想要得到的答案实际是“劳动纠纷、人格权纠纷”等信息。

[0004] 基于此，需要一种能够准确捕捉用户的搜索意图，从而直接给出满足用户搜索意图答案的语义搜索方案。

发明内容

[0005] 有鉴于此，本说明书一个或多个实施例的目的在于提出一种基于法律知识图谱的语义搜索方法、装置和电子设备，以解决无法准确捕捉用户法律搜索意图的问题。

[0006] 基于上述目的，本说明书一个或多个实施例提供了一种基于法律知识图谱的语义搜索方法，包括：

[0007] 响应于接收到用户输入的对法律问题的描述，对所述描述进行预处理；

[0008] 通过预定的关系抽取模型，从经过预处理的所述描述中提取出实体-关系-实体三元组，以构建语义搜索图；

[0009] 基于预定的法律法规类型，通过预先建立的分类模型对经过预处理的所述描述进行分类，得到类别标签；

[0010] 在预先基于法律知识相关文件构建的法律知识图谱中搜索到与所述类别标签对应的目标节点，并将所述法律知识图谱中以所述目标节点为根节点的部分法律知识图谱与所述语义搜索图进行子图匹配，得到一个或多个匹配子图；

[0011] 输出所述一个或多个匹配子图中的至少一个目标子图作为针对所述描述的法律答案。

[0012] 进一步的，所述关系抽取模型包括CASREL模型。

[0013] 进一步的，所述响应于接收到用户输入的对法律问题的描述，对所述描述进行预处理，包括下列中至少一个：

[0014] 通过正则表达式去除所述描述的标点符号和/或特殊符号；

[0015] 通过中文分词算法WMSeg对所述描述进行分词；

[0016] 将所述描述中的繁简字体进行统一化；

[0017] 通过标准表达方式对所述描述进行同义词归一化；

[0018] 通过Soft-Masked BERT模型对所述描述进行文本纠错处理。

[0019] 进一步的，所述通过预定的关系抽取模型，从经过预处理的所述描述中提取出实体-关系-实体三元组，以构建语义搜索图，包括：

[0020] 基于完整的训练模型BERT对所述描述进行编码，获取所述描述中每个词的特征表示；

[0021] 对所述特征表示进行解码，构建分类器预测所述实体位置，识别出所述实体对应的主语；

[0022] 根据所述主语提取所有可能与所述主语对应的关系，并根据所述关系识别出相应的宾语，得到实体-关系-实体三元组。

[0023] 进一步的，所述基于预定的法律法规类型，通过预先建立的分类模型对经过预处理的所述描述进行分类，得到类别标签，包括：

[0024] 基于预训练模型BERT对所述描述进行特征向量表示得到特征向量，将所述特征向量输入到Softmax回归模型基于法律法规类型进行分类，得到所述描述的类别标签。

[0025] 进一步的，所述在预先基于法律知识相关文件构建的法律知识图谱中搜索到与所述类别标签对应的目标节点，并将所述法律知识图谱中以所述目标节点为根节点的部分法律知识图谱与所述语义搜索图进行子图匹配，得到一个或多个匹配子图，包括：

[0026] 在所述语义搜索图中给定一个节点v，在所述部分法律知识图谱中找到对应的节点u，计算节点v和节点u的相似度；

[0027] 在所述语义搜索图中给定一个关系rel，在所述部分法律知识图谱中找到对应的关系L，计算所述关系rel与所述关系L的相似度；

[0028] 通过所述节点相似度和所述关系相似度计算所述语义搜索图和所述部分法律知识图谱的相似度得分，根据所述相似度得分输出匹配子图。

[0029] 进一步的，根据所述相似度得分对所述匹配子图进行排序，选取得分最高的预定数量个所述匹配子图作为目标子图。

[0030] 基于同一发明构思，本说明书一个或多个实施例提供了一种基于知识图谱的语义搜索装置，包括：

[0031] 预处理模块，被配置为响应于接收到用户输入的对法律问题的描述，对所述描述进行预处理；

[0032] 语义搜索图生成模块，被配置为通过预定的关系抽取模型，从经过预处理的所述描述中提取出实体-关系-实体三元组，以构建语义搜索图，基于预定的法律法规类型，通过预先建立的分类模型对经过预处理的所述描述进行分类，得到类别标签；

[0033] 子图匹配模块，被配置为在预先基于法律知识相关文件构建的法律知识图谱中搜索到与所述类别标签对应的目标节点，并将所述法律知识图谱中以所述目标节点为根节点的部分法律知识图谱与所述语义搜索图进行子图匹配，得到一个或多个匹配子图；

[0034] 输出模块，被配置为输出所述一个或多个匹配子图中的至少一个目标子图作为针对所述描述的法律答案。

[0035] 基于同一发明构思，本说明书一个或多个实施例还提供了一种电子设备，包括存

储器、处理器及存储在所述存储器上并可由所述处理器执行的计算机程序，所述处理器执行所述计算机程序时实现如上任一所述的方法。

[0036] 基于同一发明构思，本说明书一个或多个实施例还提供了一种非暂态计算机可读存储介质，所述非暂态计算机可读存储介质存储计算机指令，所述计算机指令在被计算机执行时，使所述计算机实现如上任一所述的方法。

[0037] 从上面所述可以看出，本说明书一个或多个实施例提供的一种基于法律知识图谱的语义搜索方法、装置和电子设备，通过对用户问题建立语义网，并结合法律知识图谱进行匹配和推理，能够准确捕捉用户的搜索意图，从而直接给出满足用户搜索意图的答案，实际解决用户的法律问题。

附图说明

[0038] 为了更清楚地说明本说明书一个或多个实施例或现有技术中的技术方案，下面将对实施例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，下面描述中的附图仅仅是本说明书一个或多个实施例，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。

[0039] 图1为本说明书一个或多个实施例的语义搜索方法的流程示意图；

[0040] 图2为本说明书一个或多个实施例的预处理操作的流程示意图

[0041] 图3为本说明书一个或多个实施例的要素提取操作的示意图；

[0042] 图4为本说明书一个或多个实施例的子图匹配操作的示意图

[0043] 图5为本说明书一个或多个实施例的语义搜索装置的模块示意图；

[0044] 图6为本说明书一个或多个实施例的电子设备的硬件结构示意图。

具体实施方式

[0045] 为使本公开的目的、技术方案和优点更加清楚明白，以下结合具体实施例，并参照附图，对本公开进一步详细说明。

[0046] 需要说明的是，除非另外定义，本说明书一个或多个实施例使用的技术术语或者科学术语应当为本公开所属领域内具有一般技能的人士所理解的通常意义。“包括”或者“包含”等类似的词语意指出现该词前面的元件或者物件涵盖出现在该词后面列举的元件或者物件及其等同，而不排除其他元件或者物件。“连接”或者“相连”等类似的词语并非限定于物理的或者机械的连接，而是可以包括电性的连接，不管是直接的还是间接的。

[0047] 如背景技术所述，当前法律搜索系统大多是基于关键词匹配技术，不能准确理解用户搜索语句的真正含义，难以解决用户法律搜索方面问题的需求，从而无法给出基于用户问题语义的精准回答。

[0048] 有鉴于此，本公开一个或多个实施例提供了一种基于法律知识图谱的语义搜索方法，对用户输入的法律问题首先进行预处理，包括去除特殊符号、分词、同义词归一化和语法纠错等。然后对所述法律问题的描述进行要素提取，提取出实体和关系，构建语义搜索图，对用户语义做初步的理解。基于法律法规类型对所述描述进行分类，输出带有法律法规类别标签的语义搜索图。将所述类别标签与预先基于法律知识相关文件构建的法律知识图谱的节点相对应，将所述语义搜索图和基于所述节点为根节点的部分法律知识图谱进行匹

配，得到一个或多个匹配子图。输出所述一个或多个匹配子图中的至少一个目标子图作为针对所述描述的法律答案。

[0049] 可见，本公开的一个或多个实施例的基于法律知识图谱的语义搜索方法通过对用户问题建立语义网，并结合法律知识图谱进行匹配和推理，能够准确捕捉用户的搜索意图，从而直接给出满足用户搜索意图的答案，而不是传统搜索系统给出的仅包含关键词的相关信息。

[0050] 以下，通过具体的实施例来详细说明本公开的一个或多个实施例的技术方案。

[0051] 参考图1，本公开的一个实施例的基于法律知识图谱的语义搜索方法，包括以下步骤：

[0052] 步骤S101、响应于接收到用户输入的对法律问题的描述，对所述描述进行预处理。

[0053] 在本步骤中，参考图2，所述预处理包括以下步骤：

[0054] 步骤S201、去除标点符号和特殊符号，由于标点符号仅对阅读理解有作用，而对语义理解没有太大作用，我们通过正则表达式的方式删除标点符号及特殊符号。

[0055] 步骤S202、分词，应用简单高效的中文分词算法WMSeg对所述描述进行分词。

[0056] 步骤S203、繁简体统一化，由于中文文字在历史上经历了多次改革，很多字存在多种书写形式，如“车”，“車”，因此对同一个字需要进行繁简体的统一化。

[0057] 步骤S204、同义词归一化，将缩写、别称和具有多种表述方式的词汇用标准表达方式进行统一，这样可以减少计算机处理不同信息的数量，提高计算效率和准确度。

[0058] 步骤S205、文本纠错，常见的文本错误主要包括字形引起的错误和拼音相似引起的错误。其他错误还包括方言、口语化和重复输入等。随着近两年预训练模型的流行，BERT类模型被迁移应用到文本纠错任务中，并取得了很好的效果。本实施例中应用文本纠错的最优模型Soft-Masked BERT对所述描述进行文本纠错。

[0059] 基于上述步骤S201至步骤S205完成对所述描述的预处理。

[0060] 步骤S102、通过预定的关系抽取模型，从经过预处理的所述描述中提取出实体-关系-实体三元组，以构建语义搜索图，基于预定的法律法规类型，通过预先建立的分类模型对经过预处理的所述描述进行分类，得到类别标签。

[0061] 本实施例中，所述要素提取具体包括：

[0062] (1)基于完整的预训练模型BERT对所述描述进行编码，获取所述描述中的每个词的特征表示，输出词向量，其中可以采用预训练模型BERT的任意一层进行编码。

[0063] (2)识别出所述描述中的主语。本步骤的主要作用是对预训练模型BERT编码获取到的词的特征表示进行解码，构建两个二分类分类器预测实体对应的主语的“起始”和“终止”索引位置。对每一个词，计算其作为“起始”或“终止”的概率，然后根据给定的阈值，大于阈值的则标记为1，小于阈值的标记为0，具体公式如下

[0064] $$p_i^{start\_s} = \sigma(W_{start}x_i + b_{start})$$

[0065] $$p_i^{end\_s} = \sigma(W_{end}x_i + b_{end})$$

[0066] 其中$p_i^{start_s}$为起始概率，$p_i^{end_s}$为终止概率，W_{start}^r和W_{end}^r为权重矩阵，b_{start}^r和b_{end}^r为偏置向量，σ(·)为sigmoid激活函数。形如$\sigma(W_{start}x_i+b_{start})$的运算为神经

网络中一个常规偏置神经元的基本运算。通过给定阈值判定所述实体作为“起始”或是“终止”，从而识别出所述实体对应的主语。如图3所示，在主语识别过程中，“李”被标记为“起始”，“，”既不是“起始”也不是“终止”，“刚”被标记为“终止”，在这里采用了最近匹配的原则，即与“李”最近的一个“终止”词为“刚”，所以“李刚”被识别为一个主语。

[0067] (3)根据上一步识别出的主语，找出所有与所述主语可能的关系，并根据所述关系识别出相应的宾语。在本步骤中同时识别出和主语相关的关系和对应的宾语。在这步解码的时候不仅考虑了BERT编码的特征向量，还考虑到识别出来的主语的特征，能够根据主语的特征更精准判定相关宾语，如下列表达式：$h_N + \mathrm{v}_{sub}^k$，其中v_{sub}^k代表主语的特征向量，若存在多个词，将其取向量平均，h_N代表BERT编码向量。对于识别出来的每一个主语，对应的每一种关系会解码出其宾语的“开始”和“结束”的索引位置，与标记主语位置类似，公式如下：

[0068] $$p_i^{start\_o} = \sigma\left(W_{start}^r \left(\mathrm{x}_i + \mathrm{v}_{sub}^k\right) + b_{start}^r\right),$$

[0069] $$p_i^{end\_o} = \sigma\left(W_{end}^r \left(\mathrm{x}_i + \mathrm{v}_{sub}^k\right) + b_{end}^r\right)$$

[0070] 通过给定阈值判定所述实体作为“起始”或是“终止”，从而识别出所述实体对应的宾语。如图3所示，展示了第一个主语的生成过程，即“李刚”，对于这个主语，在关系“出生地”中识别出了两个宾语，即“贵州安顺”和“贵州省省长”，而在其他的关系中未曾识别出相应的宾语。以上我们便可以抽取到两个三元组，如下：(李刚，出生地，贵州安顺)，(李刚，职位，贵州省省长)。

[0071] 通过预训练模型BERT对所述描述进行特征向量表示，基于法律法规类型将所述描述的所述特征向量表示输入到Softmax回归模型中进行分类，最后输出带有法律法规类别标签的语义搜索图。其中分类的类型来源于法律领域专家划分的法律法规类型，这些类型同时也是知识图谱中的部分节点，具体地：

[0072] (1)输入用户提问的文本训练集：$\mathrm{T}=\{(\mathrm{x}_1,\mathrm{y}_1),\cdots,(\mathrm{x}_N,\mathrm{y}_N)\}$，$\mathrm{i}=1,2,\cdots,\mathrm{N}$，其中$\mathrm{x}_i$为每条经过预处理的的文本，$\mathrm{y}_i$为每条文本所属的类别。

[0073] (2)基于BERT的中文短文本分类模型在训练集T上进行微调，得到训练集句子级别的特征表示，得到训练集句子级别的特征表示$\mathrm{V}=\{\mathrm{v}_1,\cdots,\mathrm{v}_N\}$，$\mathrm{i}=1,2,\cdots,\mathrm{N}$，其中$\mathrm{v}_i$表示每个文本$\mathrm{x}_i$所对应的句子级别的特征表示。

[0074] (3)将第2步得到的句子级别的特征表示$\mathrm{V}=\{\mathrm{v}_1,\cdots,\mathrm{v}_N\}$输入Softmax回归模型进行训练，计算给定样本$\mathrm{x}_i$属于第j个类别的概率：

[0075] $$p\left(y^i = j \middle| x^i; \theta\right) = \frac{e^{\theta_j^T x^i}}{\sum_{l=1}^k e^{\theta_l^T x^i}}$$

[0076] 选出最大概率值对应的类别作为样本x_i所属的类别，上式中$\theta = [\theta_1^T, \theta_2^T, \ldots, \theta_k^T]^T$为模型参数。

[0077] (4)输出文本分类训练模型，再将用户输入的所述描述的文本带入到分类训练模型中，输出所述语义搜索图的相应法律法规类别标签。

[0078] 步骤S103、在预先基于法律知识相关文件构建的法律知识图谱中搜索到与所述类

别标签对应的目标节点，并将所述法律知识图谱中以所述目标节点为根节点的部分法律知识图谱与所述语义搜索图进行子图匹配，得到一个或多个匹配子图。

[0079] 根据步骤S102中得到的法律法规类别标签在法律知识图谱中找到对应的目标节点。以下只考虑以这一个节点为根节点的部分法律知识图谱的子图匹配问题。本实施例中，子图匹配主要分别以下几步：

[0080] (1)在语义搜索图中给定一个节点v，如果v是一个实体短语或是类型短语，我们就用实体链接算法从法律知识图谱中取得所有v对应的实体和类别，并且定义这个候选集为$C(v_i)$；如果v是一个疑问词，我们就假设这个候选集$C(v_i)$由部分法律知识图谱中的所有节点组成。我们用arg_v定义语义搜索图中节点v对应词语的向量表示，将arg_v映射到部分法律知识图谱上的节点u，并用arg_u表示部分法律知识图谱中节点u对应词语的向量表示，节点v和节点u的相似度$\delta(arg_v,arg_u)$计算公式如下；

[0081] $$\delta(arg_v,arg_u)=\frac{|arg_v\cdot arg_u|}{\|arg_v\|\cdot\|arg_u\|}$$

[0082] (2)类似地，对于语义搜索图中的一个给定边$\overrightarrow{v_iv_j}$，我们在部分法律知识图谱中找到相应的边，并定义这个候选集为$C(\overrightarrow{v_iv_j})$。语义搜索图中的每一条边都对应着一个“关系”，我们需要计算这个“关系”rel_i和部分法律知识图谱中的“关系”L的相似度，为此先做如下的准备工作：对于给定的关系(relation mention)rel_i，对任意在rel_i的支撑集$Sup(rel_i)=\{(v_i^1,v_i'^1),\dots(v_i^m,v_i'^m)\}$的节点对$(v_i^j,v_i'^j)$，将$v_i^j$和$v_i'^j$之间所有的简单路径的集合记为$Paths(v_i^j,v_i'^j)$，定义$PS(rel_i)=\{Paths(v_i^j,v_i'^j)|1\le j\le m\}$，语义搜索图中的“关系”$rel_i$和部分法律知识图谱中的“关系”L的相似度计算如下：

[0083] $\delta(rel_i,L)=tf(L,PS(rel_i))\times idf(L,T)$

[0084] 这里我们借鉴了文字挖掘中常用的tf-idf(term frequency-inverse document frequency)统计思想，用tf-idf测度评估法律知识图谱中候选“关系”L与语义搜索图中给定的“关系”rel_i的相似程度。tf-idf的主要思想是：如果某个词或短语在一篇文章中出现的频数(tf)高，并且在所有文章中出现频率(idf)很低，则认为该词能够很好的代表这篇文章的某个特征，具有很好的区分能力。这里我们将tf-idf统计思想应用到关系的相似度计算中，将“关系”L类比为“某个词或短语”，将$PS(rel_i)$类比为“一篇文章”，那么“关系”L在$PS(rel_i)$中出现的频数tf为：

[0085] $$tf\left(L,PS(rel_i)\right)=|\{Paths(v_i^j,v_i'^j)|L\in Paths\left(v_i^j,v_i'^j\right),Paths\left(v_i^j,v_i'^j\right)\in PS(rel_i)\}|$$

[0086] 我们继续将所有关系的集合$T=\{rel_1,\cdots,rel_n\}$中每个rel_i所生成的$PS(rel_i)$的总和类比为“所有文章”的集合，那么“关系”L在所有这些$PS(rel_i)$中出现的频率的倒数取对数为

[0087] $$idf(L,T)=log\frac{|T|}{\left|\{rel_i\in T|L\in PS(rel_i)\}\right|+1},$$

[0088] 至此，我们通过tf-idf测度，计算得到了语义搜索图中的关系rel_i和部分法律知

识图谱中的“关系”L的相似度。如图4所示，(a)为用户输入法律问题描述，(b)为语义搜索图，(c)为候选节点和边的相似度得分，(d)为法律知识图谱中对应于语义搜索图的候选节点组成的子图集合。比如，节点V_2(“戴某”)对应于知识图谱中的候选节点有<“他人”>，<“国家工作人员”>和<“金融机构工作人员”>，和这些候选节点的相似度得分分别是0.7，0.3和0.2，所以判定“戴某”对应节点<“他人”>，其他节点和关系的判定类似。

[0089] (3)对于一个有n个节点$\{v_1,\cdots,v_n\}$的语义搜索图Q^S，部分法律知识图谱中有n个节点$\{u_1,\cdots,u_n\}$的子图M与之匹配的得分计算公式如下：

[0090]
$$Score(M)=\alpha\sum_{v_i\in V(Q^S)}\log\left(\delta\left(arg_{v_i},arg_{u_i}\right)\right)+(1-\alpha)\sum_{\overline{v_iv_j}\in E(Q^S)}\log\left(\delta\left(rel_{\overline{v_iv_j}},P_{ij}\right)\right)$$

[0091] 其中$\delta\left(arg_{v_i},arg_{u_i}\right)$和$\delta\left(rel_{\overline{v_iv_j}},P_{ij}\right)$分别由上述第2，3步计算可得，α是一个权重系数，一般取0.5。Score(M)越高，说明子图M与语义搜索图Q^S越匹配，根据所述得分Score(M)输出法律候选答案数据集。

[0092] 步骤S104、输出所述一个或多个匹配子图中的至少一个目标子图作为针对所述描述的法律答案。

[0093] 根据所述得分Score(M)对所述匹配子图进行排序，选取得分最高的k个所述匹配子图作为输出的法律答案，其中k为预定的大于1的整数。

[0094] 当用户输入对法律问题的描述后，通过上述步骤S101至S104完成语义搜索，最终为用户输出相关法律答案。

[0095] 可见，在本实施例中，基于法律知识图谱，通过对用户问题进行要素提取，构建基于用户提问的语义搜索图，并结合法律知识图谱对用户提问的分类将用户语义搜索图和法律知识图谱做子图匹配，充分利用法律知识图谱中的关联信息，最终精准地理解用户的搜索意图，并给出准确答案，实际解决用户的法律问题。

[0096] 需要说明的是，本说明书一个或多个实施例的方法可以由单个设备执行，例如一台计算机或服务器等。本实施例的方法也可以应用于分布式场景下，由多台设备相互配合来完成。在这种分布式场景的情况下，这多台设备中的一台设备可以只执行本说明书一个或多个实施例的方法中的某一个或多个步骤，这多台设备相互之间会进行交互以完成所述的方法。

[0097] 需要说明的是，上述对本说明书特定实施例进行了描述。其它实施例在所附权利要求书的范围内。在一些情况下，在权利要求书中记载的动作或步骤可以按照不同于实施例中的顺序来执行并且仍然可以实现期望的结果。另外，在附图中描绘的过程不一定要求示出的特定顺序或者连续顺序才能实现期望的结果。在某些实施方式中，多任务处理和并行处理也是可以的或者可能是有利的。

[0098] 基于同一发明构思，与上述任意实施例方法相对应的，本说明书一个或多个实施例还提供了一种基于法律知识图谱的语义搜索装置。

[0099] 参考图5，所述基于知识图谱的语义搜索装置，包括：

[0100] 预处理模块501，被配置为响应于接收到用户输入的对法律问题的描述，对所述描述进行预处理；

[0101] 语义搜索图生成模块502，被配置为通过预定的关系抽取模型，从经过预处理的所述描述中提取出实体-关系-实体三元组，以构建语义搜索图，基于预定的法律法规类型，通过预先建立的分类模型对经过预处理的所述描述进行分类，得到类别标签；

[0102] 子图匹配模块503，被配置为在预先基于法律知识相关文件构建的法律知识图谱中搜索到与所述类别标签对应的目标节点，并将所述法律知识图谱中以所述目标节点为根节点的部分法律知识图谱与所述语义搜索图进行子图匹配，得到一个或多个匹配子图；

[0103] 输出模块504，被配置为输出所述一个或多个匹配子图中的至少一个目标子图作为针对所述描述的法律答案。

[0104] 作为一个可选的实施例，所述关系抽取模型包括CASREL模型。

[0105] 作为一个可选的实施例，所述预处理模块501，具体被配置为，包括：

[0106] 通过正则表达式去除所述描述的标点符号和/或特殊符号；

[0107] 通过中文分词算法WMSeg对所述描述进行分词；

[0108] 对所述描述进行繁简字体统一化；

[0109] 对所述描述采用标准表达方式进行同义词归一化；

[0110] 通过Soft-Masked BERT模型对所述描述进行文本纠错处理。

[0111] 作为一个可选的实施例，所述语义搜索图生成模块502，具体被配置为，包括：

[0112] 基于完整的训练模型BERT对所述描述进行编码，获取所述描述中每个词的特征表示；

[0113] 对所述特征表示进行解码，构建分类器预测所述实体位置，识别出所述实体对应的主语；

[0114] 根据所述主语提取所有可能与所述主语对应的关系，并根据所述关系识别出相应的宾语，得到实体-关系-实体三元组；

[0115] 基于预训练模型BERT对所述描述进行特征向量表示得到特征向量，将所述特征向量输入到Softmax回归模型基于法律法规类型进行分类，得到所述描述的类别标签。

[0116] 作为一个可选的实施例，所述子图匹配模块503，具体被配置为，包括：

[0117] 在所述语义搜索图中给定一个节点v，在所述部分法律知识图谱中找到对应的节点u，计算节点v和节点u的相似度；

[0118] 在所述语义搜索图中给定一个关系rel，在所述部分法律知识图谱中找到对应的关系L，计算所述关系rel与所述关系L的相似度；

[0119] 通过所述节点相似度和所述关系相似度计算所述语义搜索图和所述部分法律知识图谱的相似度得分，根据所述相似度得分输出匹配子图。

[0120] 作为一个可选的实施例，根据所述相似度得分对所述匹配子图进行排序，选取得分最高的预定数量个所述匹配子图作为目标子图。

[0121] 为了描述的方便，描述以上装置时以功能分为各种模块分别描述。当然，在实施本说明书一个或多个实施例时可以把各模块的功能在同一个或多个软件和/或硬件中实现。

[0122] 上述实施例的装置用于实现前述任一实施例中相应的基于法律知识图谱的语义搜索方法，并且具有相应的方法实施例的有益效果，在此不再赘述。

[0123] 基于同一发明构思，与上述任意实施例方法相对应的，本说明书一个或多个实施例还提供了一种电子设备，包括存储器、处理器及存储在存储器上并可在处理器上运行的计算机程序，所述处理器执行所述程序时实现上任意一实施例所述的基于法律知识图谱的语义搜索方法。

[0124] 图6示出了本实施例所提供的一种更为具体的电子设备硬件结构示意图，该设备可以包括：处理器1010、存储器1020、输入/输出接口1030、通信接口1040和总线1050。其中处理器1010、存储器1020、输入/输出接口1030和通信接口1040通过总线1050实现彼此之间在设备内部的通信连接。

[0125] 处理器1010可以采用通用的CPU(Central Processing Unit，中央处理器)、微处理器、应用专用集成电路(Application Specific Integrated Circuit，ASIC)、或者一个或多个集成电路等方式实现，用于执行相关程序，以实现本说明书实施例所提供的技术方案。

[0126] 存储器1020可以采用ROM(Read Only Memory，只读存储器)、RAM(Random Access Memory，随机存取存储器)、静态存储设备，动态存储设备等形式实现。存储器1020可以存储操作系统和其他应用程序，在通过软件或者固件来实现本说明书实施例所提供的技术方案时，相关的程序代码保存在存储器1020中，并由处理器1010来调用执行。

[0127] 输入/输出接口1030用于连接输入/输出模块，以实现信息输入及输出。输入输出/模块可以作为组件配置在设备中(图中未示出)，也可以外接于设备以提供相应功能。其中输入设备可以包括键盘、鼠标、触摸屏、麦克风、各类传感器等，输出设备可以包括显示器、扬声器、振动器、指示灯等。

[0128] 通信接口1040用于连接通信模块(图中未示出)，以实现本设备与其他设备的通信交互。其中通信模块可以通过有线方式(例如USB、网线等)实现通信，也可以通过无线方式(例如移动网络、WIFI、蓝牙等)实现通信。

[0129] 总线1050包括一通路，在设备的各个组件(例如处理器1010、存储器1020、输入/输出接口1030和通信接口1040)之间传输信息。

[0130] 需要说明的是，尽管上述设备仅示出了处理器1010、存储器1020、输入/输出接口1030、通信接口1040以及总线1050，但是在具体实施过程中，该设备还可以包括实现正常运行所必需的其他组件。此外，本领域的技术人员可以理解的是，上述设备中也可以仅包含实现本说明书实施例方案所必需的组件，而不必包含图中所示的全部组件。

[0131] 上述实施例的电子设备用于实现前述任一实施例中相应的基于知识图谱的语义搜索方法，并且具有相应的方法实施例的有益效果，在此不再赘述。

[0132] 基于同一发明构思，与上述任意实施例方法相对应的，本说明书一个或多个实施例还提供了一种非暂态计算机可读存储介质，所述非暂态计算机可读存储介质存储计算机指令，所述计算机指令用于使所述计算机执行如上任一实施例所述的基于法律知识图谱的语义搜索方法。

[0133] 本实施例的计算机可读介质包括永久性和非永久性、可移动和非可移动媒体可以由任何方法或技术来实现信息存储。信息可以是计算机可读指令、数据结构、程序的模块或其他数据。计算机的存储介质的例子包括，但不限于相变内存(PRAM)、静态随机存取存储器(SRAM)、动态随机存取存储器(DRAM)、其他类型的随机存取存储器(RAM)、只读存储器

(ROM)、电可擦除可编程只读存储器(EEPROM)、快闪记忆体或其他内存技术、只读光盘只读存储器(CD-ROM)、数字多功能光盘(DVD)或其他光学存储、磁盒式磁带,磁带磁磁盘存储或其他磁性存储设备或任何其他非传输介质,可用于存储可以被计算设备访问的信息。

[0134] 上述实施例的存储介质存储的计算机指令用于使所述计算机执行如上任一实施例所述的基于知识图谱的语义搜索方法,并且具有相应的方法实施例的有益效果,在此不再赘述。

[0135] 所属领域的普通技术人员应当理解:以上任何实施例的讨论仅为示例性的,并非旨在暗示本公开的范围(包括权利要求)被限于这些例子;在本公开的思路下,以上实施例或者不同实施例中的技术特征之间也可以进行组合,步骤可以以任意顺序实现,并存在如上所述的本说明书一个或多个实施例的不同方面的许多其它变化,为了简明它们没有在细节中提供。

[0136] 另外,为简化说明和讨论,并且为了不会使本说明书一个或多个实施例难以理解,在所提供的附图中可以示出或可以不示出与集成电路(IC)芯片和其它部件的公知的电源/接地连接。此外,可以以框图的形式示出装置,以便避免使本说明书一个或多个实施例难以理解,并且这也考虑了以下事实,即关于这些框图装置的实施方式的细节是高度取决于将要实施本说明书一个或多个实施例的平台的(即,这些细节应当完全处于本领域技术人员的理解范围内)。在阐述了具体细节(例如,电路)以描述本公开的示例性实施例的情况下,对本领域技术人员来说显而易见的是,可以在没有这些具体细节的情况下或者这些具体细节有变化的情况下实施本说明书一个或多个实施例。因此,这些描述应被认为是说明性的而不是限制性的。

[0137] 尽管已经结合了本公开的具体实施例对本公开进行了描述,但是根据前面的描述,这些实施例的很多替换、修改和变型对本领域普通技术人员来说将是显而易见的。例如,其它存储器架构(例如,动态RAM(DRAM))可以使用所讨论的实施例。

[0138] 本说明书一个或多个实施例旨在涵盖落入所附权利要求的宽泛范围之内的所有这样的替换、修改和变型。因此,凡在本说明书一个或多个实施例的精神和原则之内,所做的任何省略、修改、等同替换、改进等,均应包含在本公开的保护范围之内。

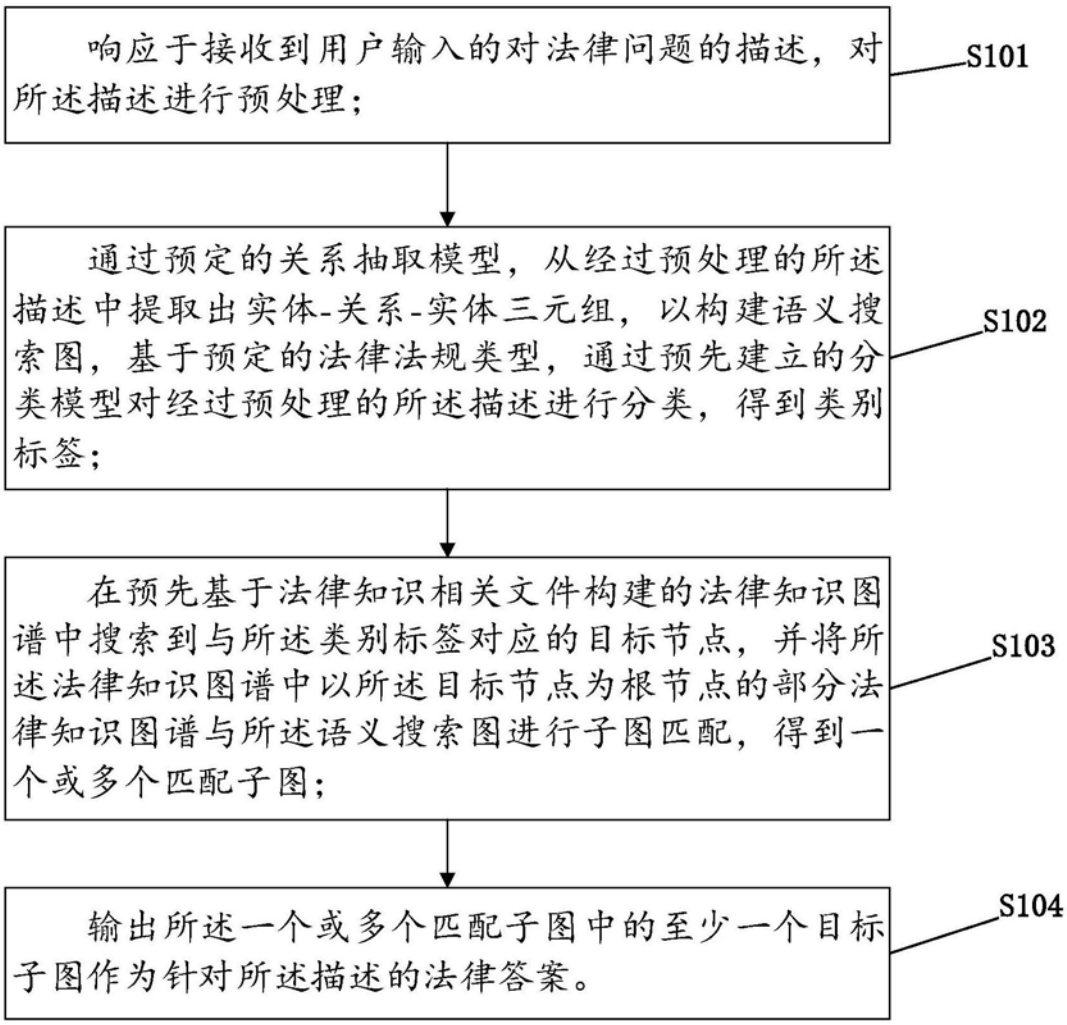

响应于接收到用户输入的对法律问题的描述，对所述描述进行预处理；
S101
通过预定的关系抽取模型，从经过预处理的所述描述中提取出实体-关系-实体三元组，以构建语义搜索图，基于预定的法律法规类型，通过预先建立的分类模型对经过预处理的所述描述进行分类，得到类别标签；
S102
在预先基于法律知识相关文件构建的法律知识图谱中搜索到与所述类别标签对应的目标节点，并将所述法律知识图谱中以所述目标节点为根节点的部分法律知识图谱与所述语义搜索图进行子图匹配，得到一个或多个匹配子图；
S103
输出所述一个或多个匹配子图中的至少一个目标子图作为针对所述描述的法律答案。
S104

图1

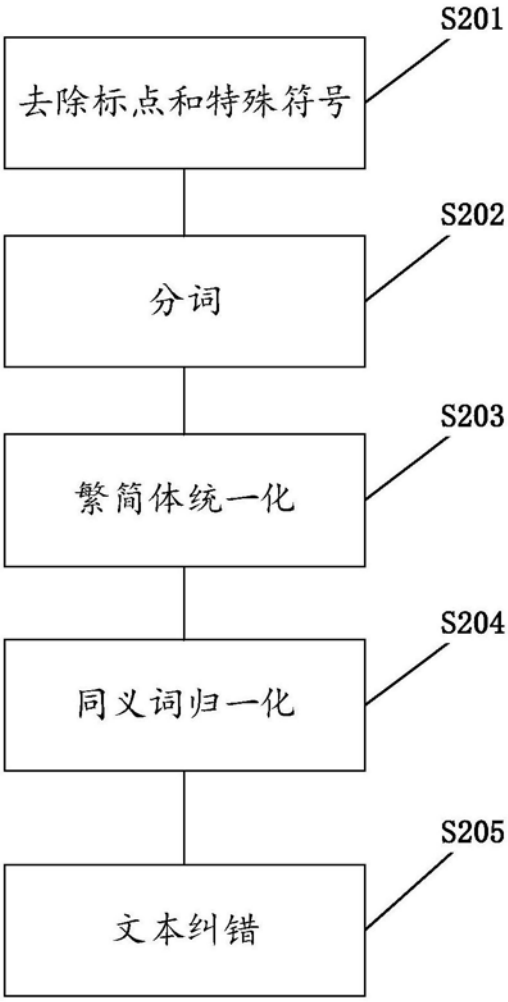


图2

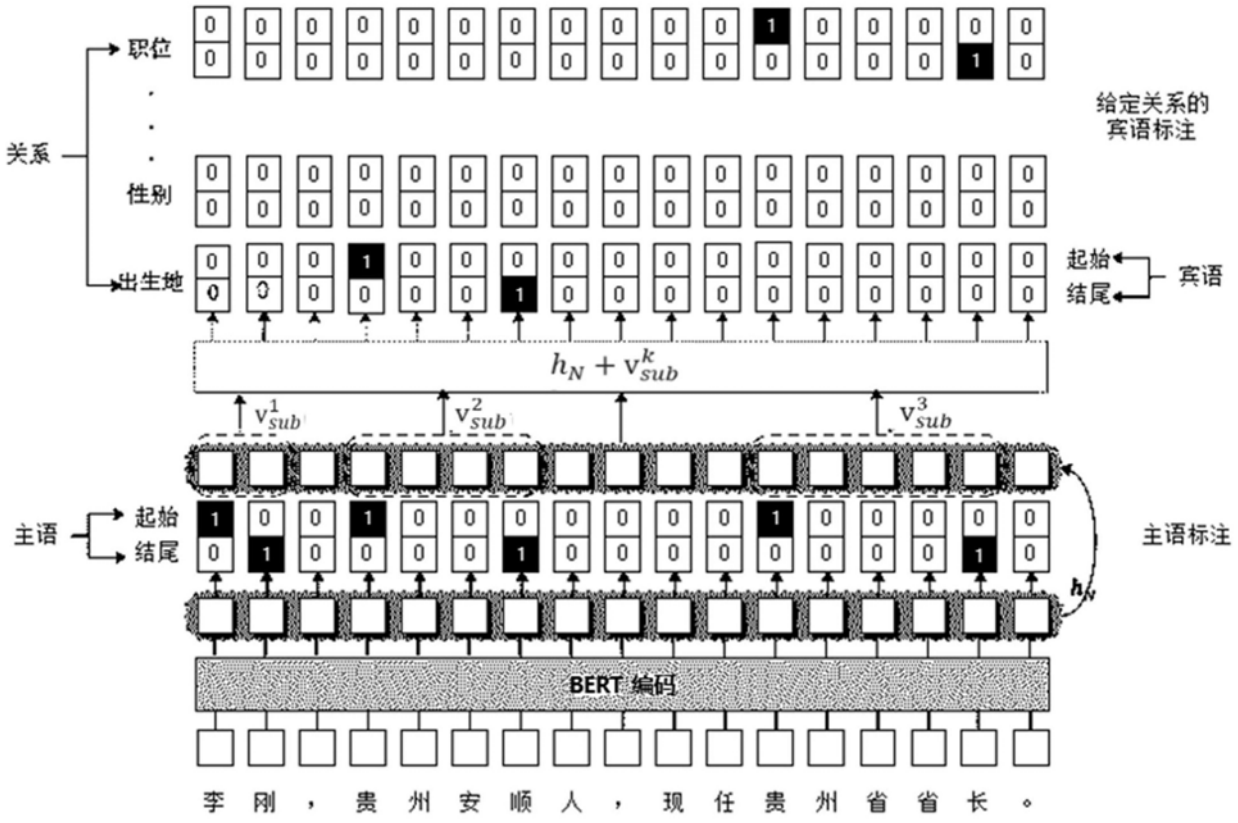

关系
职位
性别
出生地
给定关系的
宾语标注
起始
结尾
宾语
$h_N + v_{sub}^k$
v_{sub}^1
v_{sub}^2
v_{sub}^3
主语
起始
结尾
主语标注
BERT 编码
李 刚 ， 贵 州 安 顺 人 ， 现 任 贵 州 省 省 长 。

图3

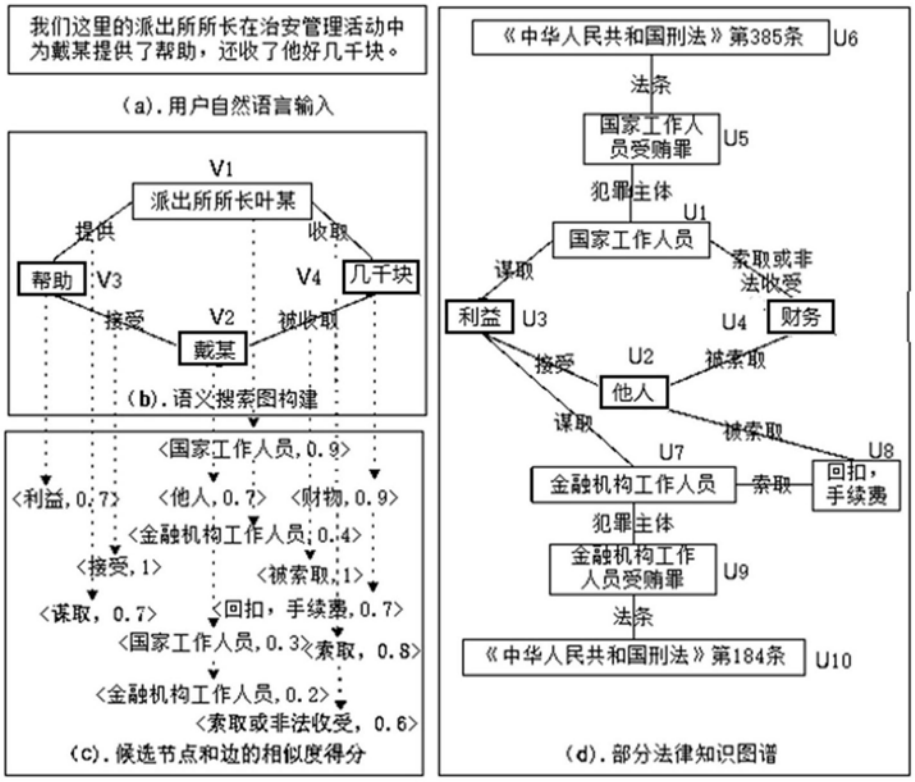

我们这里的派出所所长在治安管理活动中为戴某提供了帮助，还收了他好几千块。
(a).用户自然语言输入
V1
派出所所长叶某
提供
收取
帮助
V3
V4
几千块
接受
V2
被收取
戴某
(b).语义搜索图构建
<国家工作人员,0.9>
<利益,0.7>
<他人,0.7>
<财物,0.9>
<金融机构工作人员,0.4>
<接受,1>
<被索取,1>
<谋取,0.7>
<回扣,手续费,0.7>
<国家工作人员,0.3>
<索取,0.8>
<金融机构工作人员,0.2>
<索取或非法收受,0.6>
(c).候选节点和边的相似度得分
《中华人民共和国刑法》第385条 U6
法条
国家工作人员受贿罪 U5
犯罪主体
U1
国家工作人员
谋取
索取或非法收受
利益 U3
U4
财务
接受
U2
被索取
他人
谋取
被索取
U8
U7
金融机构工作人员
索取
回扣，手续费
犯罪主体
金融机构工作人员受贿罪 U9
法条
《中华人民共和国刑法》第184条 U10
(d).部分法律知识图谱

图4

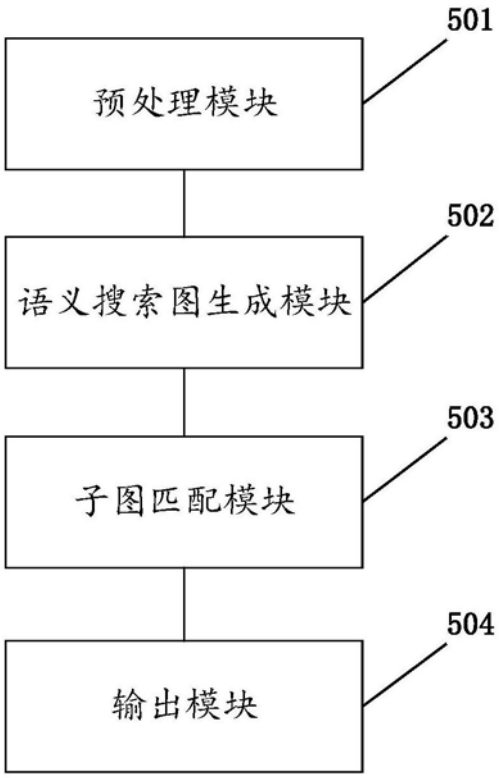

501
预处理模块
502
语义搜索图生成模块
503
子图匹配模块
504
输出模块

图5

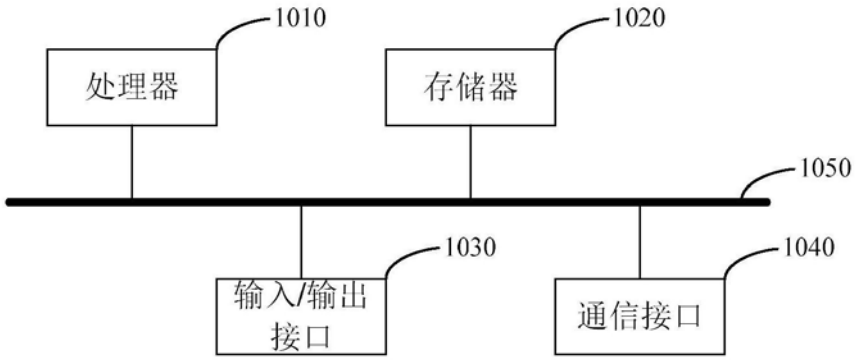


图6