(19)国家知识产权局

(12)发明专利申请

(10)申请公布号 CN 118820403 A
(43)申请公布日 2024.10.22

(21)申请号 202411000004.X

(22)申请日 2024.07.24

(71)申请人 江苏风云科技服务有限公司
地址 215000 江苏省苏州市工业园区金鸡湖大道1355号国际科技园科技广场四楼

(72)发明人 董爱平 严世振 夏晓东 龚祖明 徐雪阳

(74)专利代理机构 苏州创智高诺知识产权代理有限公司 32843
专利代理师 王敏

(51)Int.Cl.
G06F 16/33(2019.01)
G06F 16/35(2019.01)
G06F 16/36(2019.01)
G06N 5/02(2023.01)
G06N 5/025(2023.01)

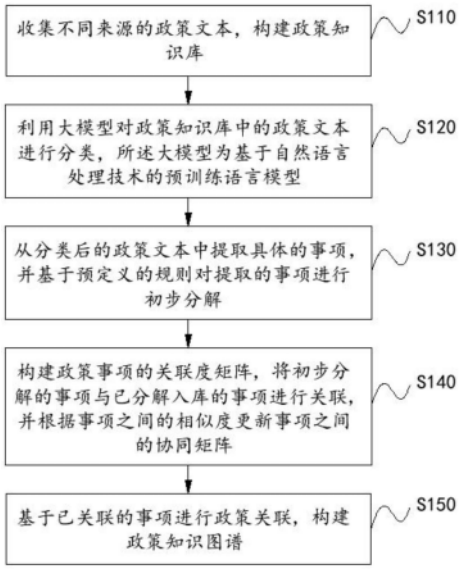
权利要求书2页 说明书8页 附图4页

(54)发明名称

基于大模型的政策文本去噪与关联事项提取方法及系统

(57)摘要

本申请公开了一种基于大模型的政策文本去噪与关联事项提取方法及系统,涉及数据处理技术领域。该方法包括:收集不同来源的政策文本,构建政策知识库;利用大模型对政策知识库中的政策文本进行分类;从分类后的政策文本中提取具体的事项,并基于预定义的规则对提取的事项进行初步分解;构建政策事项的关联度矩阵,将初步分解的事项与已分解入库的事项进行关联,并根据事项之间的相似度更新事项之间的协同矩阵;基于已关联的事项进行政策关联,构建政策知识图谱。本方案利用大模型和自然语言处理技术,通过对政策文本分类、去噪、提取事项和构建关联度,增强了政策信息的系统化管理和关联关系分析能力,提高了政策文本的处理和分析效率。



CN 118820403 A

1.基于大模型的政策文本去噪与关联事项提取方法,其特征在于,包括以下步骤:

收集不同来源的政策文本,构建政策知识库;

利用大模型对所述政策知识库中的政策文本进行分类,所述大模型为基于自然语言处理技术的预训练语言模型;

从分类后的政策文本中提取具体的事项,并基于预定义的规则对提取的事项进行初步分解;

构建政策事项的关联度矩阵,所述关联度矩阵用于表示不同事项之间的关联度,将初步分解的事项与已分解入库的事项进行关联,并根据事项之间的相似度更新事项之间的协同矩阵;

基于已关联的事项进行政策关联,构建政策知识图谱,所述政策知识图谱用于表示政策文本中事项之间的关系。

2.根据权利要求1所述的方法,其特征在于,还包括以下步骤:

复核事项之间的关联关系,复核过程包括专家评审和/或自动化检测,并将复核后的事项录入到标准化事项库中;

基于所述标准化事项库对所述大模型进行优化,所述优化过程包括调整模型参数和更新训练数据。

3.根据权利要求2所述的方法,其特征在于,所述收集不同来源的政策文本,构建政策知识库,包括:

收集政策文本,政策文本的来源包括政策文档、申报通知、学术文章及实施细则;

对收集的政策文本进行数据清洗处理,包括文本去重、去除无关信息及格式统一;

将清洗后的政策文本存储在政策知识库中,所述政策知识库为可查询数据库。

4.根据权利要求3所述的方法,其特征在于,所述利用大模型对所述政策知识库中的政策文本进行分类,包括:

识别所述政策知识库中每个政策文本的关键特征,为不同类别的政策文本建立分类知识库;

将人工分类标注数据作为训练集对大模型进行训练,并利用训练后的大模型对政策文本进行预分类;

对大模型的预分类结果进行人工抽样核验,并对大模型评估准确率低的政策文本数据进行人工标注;

将经过人工标注后的政策文本数据作为训练集继续对大模型进行训练,以优化大模型的分类性能;

利用优化后的大模型对所述政策知识库中的政策文本进行自动分类。

5.根据权利要求4所述的方法,其特征在于,所述从分类后的政策文本中提取具体的事项,并基于预定义的规则对提取的事项进行初步分解,包括:

确定用于识别和提取事项的关键词;

利用大模型对分类后的政策文本进行细粒度分析,识别出包含具体事项的段落或句子,并根据预定义的规则,从识别出的段落或句子中提取具体的事项,所述预定义规则包括识别事项的句型结构和术语;

通过自然语言处理NLP方法对提取的事项进行初步分解,得到政策事项文本。

6.根据权利要求5所述的方法,其特征在于,所述构建政策事项的关联度矩阵,将初步分解的事项与已分解入库的事项进行关联,并根据事项之间的相似度更新事项之间的协同矩阵,包括:

将所述政策事项文本转换为特征向量;

使用聚类算法对所述特征向量进行聚类,得到不同类别的政策事项;

对每个聚类内部的政策事项,使用LSH方法计算两两政策事项之间的相似度,构建每个聚类内部的关联度矩阵;

对所述初步分解的事项,使用LSH方法基于其特征向量寻找相关聚类,计算其与所述相关聚类内已分解入库的事项的相似度,并根据所计算的相似度将所述初步分解的事项与已分解入库的事项进行关联;

根据所述初步分解的事项与其相关聚类内事项的相似度,采用哈希表动态更新事项之间的协同矩阵。

7.根据权利要求6所述的方法,其特征在于,所述基于已关联的事项进行政策关联,构建政策知识图谱,包括:

在已建立的协同矩阵中提取事项之间的关系,包括相似度和/或关联度;

为每个政策事项创建一个节点,所述节点的属性包括事项的名称、内容和/或分类;

根据提取事项之间的关系在所述节点之间创建边,所述边的属性包括关系类型、相似度和/或关联度;

基于所述节点和所述边构建知识图谱,并将构建的知识图谱存储到图数据库中。

8.根据权利要求2-7任一项所述的方法,其特征在于,所述复核事项之间的关联关系,包括:

如果对提取的事项存在分解或者关联错误,则通过人工修改事项分解或者关联,或通过大模型重新进行事项分解或者关联。

9.基于大模型的政策文本去噪与关联事项提取系统,其特征在于,所述系统包括:

政策文本收集模块,用于收集不同来源的政策文本,构建政策知识库;

政策文本分类模块,用于利用大模型对所述政策知识库中的政策文本进行分类,所述大模型为基于自然语言处理技术的预训练语言模型;

事项提取分解模块,用于从分类后的政策文本中提取具体的事项,并基于预定义的规则对提取的事项进行初步分解;

矩阵构建更新模块,用于构建政策事项的关联度矩阵,所述关联度矩阵用于表示不同事项之间的关联度,将初步分解的事项与已分解入库的事项进行关联,并根据事项之间的相似度更新事项之间的协同矩阵;

知识图谱构建模块,用于基于已关联的事项进行政策关联,构建政策知识图谱,所述政策知识图谱用于表示政策文本中事项之间的关系。

10.一种电子设备,包括存储器、处理器以及存储在所述存储器上并可在所述处理器上运行的计算机程序,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8中任一项所述的基于大模型的政策文本去噪与关联事项提取方法。

基于大模型的政策文本去噪与关联事项提取方法及系统

技术领域

[0001] 本申请涉及数据处理技术领域,具体来说涉及一种基于大模型的政策文本去噪与关联事项提取方法及系统。

背景技术

[0002] 政策文件的整理和分析是政府部门、研究机构以及企业在政策制定、研究和决策过程中至关重要的环节。传统的政策文件的整理和分析通常依靠领域专家进行人工处理,专家通过阅读、理解和归纳政策文本来提取关键信息和关联事项。这种方法虽然可以获得高质量的结果,但其效率低下,成本高昂,难以满足大规模政策文本处理的需求。此外,基于规则的政策文件处理方法可以依赖于预定义的规则和模板进行文本处理。该方法在处理格式化和结构化文本时效果较好,但在面对非结构化和复杂文本时,其效果显著下降。

[0003] 随着机器学习技术的发展,基于机器学习的自然语言处理(Natural LanguageProcessing,NLP)技术逐渐应用于政策文本处理。该方法通过训练模型来自动化识别和提取文本信息,极大地提高了处理效率。然而,传统的NLP和早期深度学习模型在语义理解和关联提取方面能力不足,难以处理复杂的政策文本关系。此外,由于政策发文来源于不同的地区和部门,不同级别、部门的政策侧重不同,现有技术难以消除噪音,且难以将不同级别的政策进行关联以提取合并信息并保持文本一致性。

发明内容

[0004] 本发明的目的在于提供一种基于大模型的政策文本去噪与关联事项提取方法及系统,利用大模型和自然语言处理技术对政策文本进行分类和去噪,提高政策文本信息提取的准确性,减少了噪音干扰。

[0005] 为实现上述目的,本发明公开了如下技术方案:

[0006] 本发明一方面提供了一种基于大模型的政策文本去噪与关联事项提取方法,该方法包括以下步骤:

[0007] 收集不同来源的政策文本,构建政策知识库;

[0008] 利用大模型对所述政策知识库中的政策文本进行分类,所述大模型为基于自然语言处理技术的预训练语言模型;

[0009] 从分类后的政策文本中提取具体的事项,并基于预定义的规则对提取的事项进行初步分解;

[0010] 构建政策事项的关联度矩阵,所述关联度矩阵用于表示不同事项之间的关联度,将初步分解的事项与已分解入库的事项进行关联,并根据事项之间的相似度更新事项之间的协同矩阵;

[0011] 基于已关联的事项进行政策关联,构建政策知识图谱,所述政策知识图谱用于表示政策文本中事项之间的关系。

[0012] 可选地,上述的政策文本去噪与关联事项提取方法,还包括以下步骤:

[0013] 复核事项之间的关联关系,复核过程包括专家评审和/或自动化检测,并将复核后的事项录入到标准化事项库中;

[0014] 基于所述标准化事项库对所述大模型进行优化,所述优化过程包括调整模型参数和更新训练数据。

[0015] 优选地,上述的方法中,所述收集不同来源的政策文本,构建政策知识库,包括:

[0016] 收集政策文本,政策文本的来源包括政策文档、申报通知、学术文章及实施细则;

[0017] 对收集的政策文本进行数据清洗处理,包括文本去重、去除无关信息及格式统一;

[0018] 将清洗后的政策文本存储在政策知识库中,所述政策知识库为可查询数据库。

[0019] 进一步的,上述的方法中,所述利用大模型对所述政策知识库中的政策文本进行分类,包括:

[0020] 识别所述政策知识库中每个政策文本的关键特征,为不同类别的政策文本建立分类知识库;

[0021] 将人工分类标注数据作为训练集对大模型进行训练,并利用训练后的大模型对政策文本进行预分类;

[0022] 对大模型的预分类结果进行人工抽样核验,并对大模型评估准确率低的政策文本数据进行人工标注;

[0023] 将经过人工标注后的政策文本数据作为训练集继续对大模型进行训练,以优化大模型的分类性能;

[0024] 利用优化后的大模型对所述政策知识库中的政策文本进行自动分类。

[0025] 进一步的,上述的方法中,所述从分类后的政策文本中提取具体的事项,并基于预定义的规则对提取的事项进行初步分解,包括:

[0026] 确定用于识别和提取事项的关键词;

[0027] 利用大模型对分类后的政策文本进行细粒度分析,识别出包含具体事项的段落或句子,并根据预定义的规则,从识别出的段落或句子中提取具体的事项,所述预定义规则包括识别事项的句型结构和术语;

[0028] 通过自然语言处理NLP方法对提取的事项进行初步分解,得到政策事项文本。

[0029] 进一步的,上述的方法中,所述构建政策事项的关联度矩阵,将初步分解的事项与已分解入库的事项进行关联,并根据事项之间的相似度更新事项之间的协同矩阵,包括:

[0030] 将所述政策事项文本转换为特征向量;

[0031] 使用聚类算法对所述特征向量进行聚类,得到不同类别的政策事项;

[0032] 对每个聚类内部的政策事项,使用LSH方法计算两两政策事项之间的相似度,构建每个聚类内部的关联度矩阵;

[0033] 对所述初步分解的事项,使用LSH方法基于其特征向量寻找相关聚类,计算其与所述相关聚类内已分解入库的事项的相似度,并根据所计算的相似度将所述初步分解的事项与已分解入库的事项进行关联;

[0034] 根据所述初步分解的事项与其相关聚类内事项的相似度,采用哈希表动态更新事项之间的协同矩阵。

[0035] 进一步的,上述的方法中,所述基于已关联的事项进行政策关联,构建政策知识图谱,包括:

[0036] 在已建立的协同矩阵中提取事项之间的关系,包括相似度和/或关联度;

[0037] 为每个政策事项创建一个节点,所述节点的属性包括事项的名称、内容和/或分类;

[0038] 根据提取事项之间的关系在所述节点之间创建边,所述边的属性包括关系类型、相似度和/或关联度;

[0039] 基于所述节点和所述边构建知识图谱,并将构建的知识图谱存储到图数据库中。

[0040] 进一步的,上述的方法中,所述复核事项之间的关联关系,包括:

[0041] 如果对提取的事项存在分解或者关联错误,则通过人工修改事项分解或者关联,或通过大模型重新进行事项分解或者关联。

[0042] 本发明另一方面提供了一种基于大模型的政策文本去噪与关联事项提取系统,所述系统包括:

[0043] 政策文本收集模块,用于收集不同来源的政策文本,构建政策知识库;

[0044] 政策文本分类模块,用于利用大模型对所述政策知识库中的政策文本进行分类,所述大模型为基于自然语言处理技术的预训练语言模型;

[0045] 事项提取分解模块,用于从分类后的政策文本中提取具体的事项,并基于预定义的规则对提取的事项进行初步分解;

[0046] 矩阵构建更新模块,用于构建政策事项的关联度矩阵,所述关联度矩阵用于表示不同事项之间的关联度,将初步分解的事项与已分解入库的事项进行关联,并根据事项之间的相似度更新事项之间的协同矩阵;

[0047] 知识图谱构建模块,用于基于已关联的事项进行政策关联,构建政策知识图谱,所述政策知识图谱用于表示政策文本中事项之间的关系。

[0048] 本发明还提供了一种电子设备,包括存储器、处理器以及存储在存储器上并可在所述处理器上运行的计算机程序,所述处理器执行所述计算机程序时实现如上所述的基于大模型的政策文本去噪与关联事项提取方法。

[0049] 发明内容中提供的效果仅仅是实施例的效果,而不是发明所有的全部效果,上述技术方案中的一个技术方案具有如下优点或有益效果:

[0050] 本公开实施例提供的技术方案,首先通过构建政策知识库,将来自不同来源的政策文本进行集中管理;然后利用预训练的自然语言处理模型对政策文本进行分类和去噪,自动化识别和过滤掉无关或冗余的信息;接下来构建政策事项的关联度矩阵和协同矩阵,定量分析不同政策事项之间的关联度和相似度;然后基于已关联的事项构建政策知识图谱,直观地展示政策文本中事项之间的关系,帮助用户全面理解政策内容及其内在关联。本方案利用大模型和自然语言处理技术,通过对政策文本分类、去噪、提取事项和构建关联度,增强了政策信息的系统化管理和关联关系分析能力,提高了政策文本的处理和分析效率。

附图说明

[0051] 此处的附图被并入说明书中并构成说明书的一部分,示出了符合本申请的实施例,并与说明书一起用于解释本申请的原理。

[0052] 图1为本申请一实施例提供的基于大模型的政策文本去噪与关联事项提取方法流

程示意图;

[0053] 图2为本申请另一实施例提供的基于大模型的政策文本去噪与关联事项提取方法流程示意图;

[0054] 图3为图2中的基于大模型的政策文本去噪与关联事项提取方法工作流程示意图;

[0055] 图4为本申请一实施例提供的基于大模型的政策文本去噪与关联事项提取系统结构示意图;

[0056] 图5为本申请一实施例提供的电子设备的结构示意图;

[0057] 附图标记:

[0058] 410-政策文本收集模块,420-政策文本分类模块,430-事项提取分解模块,440-矩阵构建更新模块,450-知识图谱构建模块;

[0059] 510-输入单元,520-存储器,530-处理器,540-输出单元。

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仅用以解释本发明,并不用于限定本发明。

[0061] 需要说明的,本说明书中针对“一个实施例”、“实施例”、“示例实施例”等的引用,指的是描述的该实施例可包括特定的特征、结构或特性,但是,不是每个实施例必须包含这些特定特征、结构或特性。此外,这样的表述并非指的是同一个实施例。进一步,在结合实施例描述特定的特征、结构或特性时,不管有没有明确的描述,已经表明将这样的特征、结构或特性结合到其它实施例中是在本领域技术人员的知识范围内的。

[0062] 此外,在说明书及后续的权利要求当中使用了某些词汇来指称特定组件或部件,所属领域中具有通常知识者应可理解,制造商可以用不同的名词或术语来称呼同一个组件或部件。本说明书及后续的权利要求并不以名称的差异来作为区分组件或部件的方式,而是以组件或部件在功能上的差异来作为区分的准则。在通篇说明书及后续的权利要求书中所提及的“包括”和“包含”为一开放式的用语,故应解释成“包含但不限定于”。另外,“连接”一词在此系包含任何直接及间接的电性连接手段。间接的电性连接手段包括通过其它装置进行连接。

[0063] 参考图1,图1示出了本申请一实施例提供的基于大模型的政策文本去噪与关联事项提取方法流程示意图,该方法包括以下步骤:

[0064] S110、收集不同来源的政策文本,构建政策知识库。

[0065] 在一些实施例中,政策文本的来源包括政策文档、申报通知、学术文章及实施细则等。

[0066] S120、利用大模型对政策知识库中的政策文本进行分类,所述大模型为基于自然语言处理技术的预训练语言模型。

[0067] 在一些实施例中,可使用BERT或RoBERTa作为预训练语言模型,对政策知识库中的政策文本进行分类。

[0068] S130、从分类后的政策文本中提取具体的事项,并基于预定义的规则对提取的事项进行初步分解。

[0069] 在一些实施例中,可通过自然语言处理NLP方法对提取的事项进行初步分解。

[0070] S140、构建政策事项的关联度矩阵,所述关联度矩阵用于表示不同事项之间的关联度,将初步分解的事项与已分解入库的事项进行关联,并根据事项之间的相似度更新事项之间的协同矩阵。

[0071] S150、基于已关联的事项进行政策关联,构建政策知识图谱,所述政策知识图谱用于展示政策文本中事项之间的关系,以帮助识别政策之间的关联和影响,便于政策分析和决策。

[0072] 在本实施例中,通过构建政策知识库,将来自不同来源的政策文本进行集中管理,使得信息更加系统化和有序化。利用预训练的自然语言处理模型对政策文本进行分类和去噪,能够自动化识别和过滤无关或冗余的信息。构建关联度矩阵和协同矩阵,能够定量分析不同政策事项之间的关联度和相似度。基于已关联的事项构建政策知识图谱,可以直观地展示政策文本中事项之间的关系。本方案通过自动化、系统化、结构化、关联化和可视化的手段,显著提升了政策文本的处理、分析和利用的效率和效果。

[0073] 为了更详细的描述本发明的技术方案,下面结合图2、图3对本申请实施例的基于大模型的政策文本去噪与关联事项提取过程进行具体说明。

[0074] 参考图2、图3,图2示出了本申请另一实施例提供的基于大模型的政策文本去噪与关联事项提取方法流程示意图,图3为图2中方法的工作流程示意图。如图2、图3所示,该方法包括以下步骤:

[0075] S210、收集不同来源的政策文本,构建政策知识库。

[0076] 在一些实施例中,步骤S210具体包括:

[0077] 收集政策文本,政策文本的来源包括政府网站、政策文档、实施细则、申报通知、法律数据库及学术文章等;

[0078] 对收集的政策文本进行数据清洗处理,包括文本去重、去除无关信息及格式统一;

[0079] 将清洗后的政策文本存储在政策知识库中,所述政策知识库为可查询数据库,如MongoDB、Elasticsearch。

[0080] S220、利用大模型对政策知识库中的政策文本进行分类,所述大模型为基于自然语言处理技术的预训练语言模型。

[0081] 在一些实施例中,步骤S220可通过以下步骤实现:

[0082] 识别政策知识库中每个政策文本的关键特征(如行业、标签等),根据预定义的规则为不同类别的政策文本建立分类知识库,预定义的规则包括分类规则;

[0083] 例如,根据行业知识和经验,对于包含“实施细则”的政策文本,可以通过关键词如“实施细则”、“具体措施”等或者语义上的相似性进行区分。

[0084] 将已积累的人工分类标注数据作为训练集,使用预训练的大模型(如BERT、RoBERTa等)作为基础模型进行训练,以使大模型适应特定的政策文本分类任务,并利用训练后的大模型对政策文本进行预分类;

[0085] 对大模型的预分类结果进行人工抽样核验,并对大模型评估准确率低的政策文本数据进行人工标注;

[0086] 将经过人工标注后的政策文本数据作为训练集继续对大模型进行训练,以优化大模型的分类性能;

[0087] 利用优化后的大模型对政策知识库中的政策文本进行自动分类。

[0088] 本步骤中,将大模型识别处理与人工标注相结合,利用大模型的强大语义理解能力和人工标注数据的准确性,有效提升政策文本分类的精度和效率。通过对大模型的分类结果进行人工抽样核验,帮助发现模型可能存在的分类错误或模糊边界情况,确保评估模型分类的准确性和一致性。针对模型评估准确率低的数据,进行详细的人工标注,有助于识别和修正模型在特定情况下的分类偏差或错误,确保大模型在实际应用中的可靠性。

[0089] S230、从分类后的政策文本中提取具体的事项,并基于预定义的规则对提取的事项进行初步分解。

[0090] 在一些实施例中,步骤S230具体包括:

[0091] 确定用于识别和提取事项的关键词(如“时间”、“规定”、“要求”、“措施”等);

[0092] 利用大模型对分类后的政策文本进行细粒度分析,识别出包含具体事项的段落或句子,并根据预定义的规则,从识别出的段落或句子中提取具体事项,其中,预定义规则包括识别事项的句型结构和术语;

[0093] 通过自然语言处理NLP方法对提取的事项进行初步分解,得到政策事项文本,如责任主体、执行时间、具体措施等。

[0094] S240、构建政策事项的关联度矩阵,所述关联度矩阵用于表示不同事项之间的关联度,将初步分解的事项与已分解入库的事项进行关联,并根据事项之间的相似度更新事项之间的协同矩阵。

[0095] 在一些实施例中,步骤S240可通过以下步骤实现:

[0096] 使用向量化方法将政策事项文本转换为特征向量,用于后续的计算和分析,向量化方法可以采用TF-IDF、Word2Vec或BERT;

[0097] 采用聚类算法对特征向量进行聚类,确保同类政策事项被归为同一个类别,从而得到不同类别的政策事项,以便于关联度分析,聚类算法可以为K-means、层次聚类或DBSCAN;

[0098] 对每个聚类内部的政策事项,使用局部敏感哈希LSH方法计算两两政策事项之间的相似度,如使用余弦相似度或欧氏距离方法,构建每个聚类内部的关联度矩阵;

[0099] 对初步分解的事项,使用局部敏感哈希LSH方法基于其特征向量寻找相关聚类,计算其与相关聚类内已分解入库的事项的相似度,并根据相似度将初步分解的事项与已分解入库的事项进行关联;

[0100] 根据初步分解的事项与其相关聚类内事项的相似度,采用哈希表动态更新事项之间的协同矩阵,使协同矩阵能够随着新事项的加入不断调整和优化。

[0101] 本步骤中,通过向量化方法准确表示政策事项的语义信息,并利用LSH方法快速计算相似度,显著降低了计算复杂度,实现了复杂度从$O(n^2)$到$O(n\log n)$的优化。通过聚类算法将同类政策事项归为同一个类别,聚类后在每个类别内部计算相似度,减少无关事项的比较,提高了相似度计算的精度。将LSH方法和哈希表结合使用,显著降低了算法的时间复杂度,使系统能够在较短时间内完成大规模数据的处理和分析,提升系统的响应速度和处理效率。即本步骤通过构建政策事项的关联度矩阵,结合LSH方法、聚类算法、哈希表和动态更新机制,有效提高了政策事项关联度分析的效率和精度,显著降低了计算复杂度,实现了高效、实时和可扩展的政策数据处理和分析。

[0102] S250、基于已关联的事项进行政策关联,构建政策知识图谱,所述政策知识图谱用于展示政策文本中事项之间的关系,以帮助识别政策之间的关联和影响,便于政策分析和决策。

[0103] 在一些实施例中,基于已关联的事项进行政策关联,构建政策知识图谱,包括:

[0104] 在已建立的协同矩阵中提取事项之间的关系,包括相似度、关联度等,相似度和关联度可通过文本分析或主题模型等方法计算获得;

[0105] 为每个政策事项创建一个节点,节点的属性包括事项的名称、内容、分类等;

[0106] 根据提取事项之间的关系在节点之间创建边,边的属性包括关系类型、相似度、关联度等,边的权重表示关系的强度;

[0107] 基于节点和边构建知识图谱,并将构建的知识图谱存储到图数据库中,图数据库可采用Neo4j或ArangoDB等支持图结构的数据库;

[0108] 利用图数据库的API或查询语言(如Cypher、Gremlin等)进行知识图谱管理,包括数据的插入、更新、删除、查询操作等。

[0109] S260、复核事项之间的关联关系,复核过程包括专家评审和/或自动化检测,并将复核后的事项录入到标准化事项库中。

[0110] 复核事项之间的关联关系,如果发现提取的事项存在分解错误,则通过人工和/或大模型来进行修改;如果发现提取的事项存在关联错误,则通过人工进行修改。

[0111] S270、基于标准化事项库对大模型进行优化,优化过程包括调整模型参数和更新训练数据。

[0112] 在本实施例中,通过利用大模型的强大语义理解能力,结合人工标注数据的准确性,实现了高效、准确的政策文本分类和事项提取。通过动态优化和更新机制,确保系统在处理大规模数据时的高效性和可扩展性。同时,基于知识图谱的可视化展示和复核机制,提高了政策分析和决策的科学性和可靠性。

[0113] 需要说明的是,对于方法实施例,为了简单描述,故将其都表述为一系列的动作组合,但是本领域技术人员应该知悉,本申请实施例并不受所描述的动作顺序的限制,因为依据本申请实施例,某些步骤可以采用其他顺序或者同时进行。其次,本领域技术人员也应该知悉,说明书中所描述的实施例均属于优选实施例,所涉及的动作并不一定是本申请实施例所必须的。

[0114] 参考图4,图4示出了本申请一实施例提供的基于大模型的政策文本去噪与关联事项提取系统结构示意图,利用该系统可以实现政策文本去噪与关联事项提取,下文描述的该系统可以与上文描述的政策文本去噪与关联事项提取方法相互对应参照。具体的,该系统包括:

[0115] 政策文本收集模块410,用于收集不同来源的政策文本,构建政策知识库;

[0116] 政策文本分类模块420,用于利用大模型对所述政策知识库中的政策文本进行分类,所述大模型为基于自然语言处理技术的预训练语言模型;

[0117] 事项提取分解模块430,用于从分类后的政策文本中提取具体的事项,并基于预定义的规则对提取的事项进行初步分解;

[0118] 矩阵构建更新模块440,用于构建政策事项的关联度矩阵,所述关联度矩阵用于表示不同事项之间的关联度,将初步分解的事项与已分解入库的事项进行关联,并根据事项

之间的相似度更新事项之间的协同矩阵；

[0119] 知识图谱构建模块450，用于基于已关联的事项进行政策关联，构建政策知识图谱，所述政策知识图谱用于表示政策文本中事项之间的关系。

[0120] 关于本实施例中基于大模型的政策文本去噪与关联事项提取系统未详述的过程，可参照上述基于大模型的政策文本去噪与关联事项提取方法实施例中的相关部分，在此不再赘述。

[0121] 本申请实施例还提供一种电子设备，如图5所示，电子设备可包括输入单元510、存储器520、处理器530及输出单元540。其中，存储器520存储有可在处理器530上运行的程序指令，处理器530调用程序指令能够执行基于前述多个实施例中的方法和/或技术方案。该电子设备可以为手机、电脑等移动终端设备。关于本实施例中电子设备未详述的过程，可参照上述政策文本去噪与关联事项提取方法实施例中的相关部分，在此不再赘述。

[0122] 以上对本申请实施例提供的基于大模型的政策文本去噪与关联事项提取方法及系统进行了详细介绍。说明书中各个实施例采用递进的方式描述，每个实施例重点说明的都是与其他实施例的不同之处，各个实施例之间相同相似部分互相参见即可。对于实施例公开的设备而言，由于其与实施例公开的方法相对应，所以描述的比较简单，相关之处参见方法部分说明即可。应当指出，对于本领域的普通技术人员来说，在不脱离本申请原理的前提下，还可以对本申请进行若干改进和修饰，这些改进和修饰也落入本申请权利要求的保护范围内。

[0123] 显然，本领域的技术人员应该明白，上述的本申请的各模块或各步骤可以用通用的计算装置来实现，它们可以集中在单个的计算装置上，或者分布在多个计算装置所组成的网络上，可选地，它们可以用计算装置可执行的程序代码来实现，从而，可以将它们存储在存储装置中由计算装置来执行，或者将它们分别制作成各个集成电路模块，或者将它们中的多个模块或步骤制作成单个集成电路模块来实现。这样，本申请不限制于任何特定的硬件和软件结合。

[0124] 虽然，上文中已经用一般性说明及具体实施例对本申请作了详尽的描述，但在本申请基础上，可以对之作一些修改或改进，这对本领域技术人员而言是显而易见的。因此，在不偏离本申请精神的基础上所做的这些修改或改进，均属于本申请要求保护的范围。

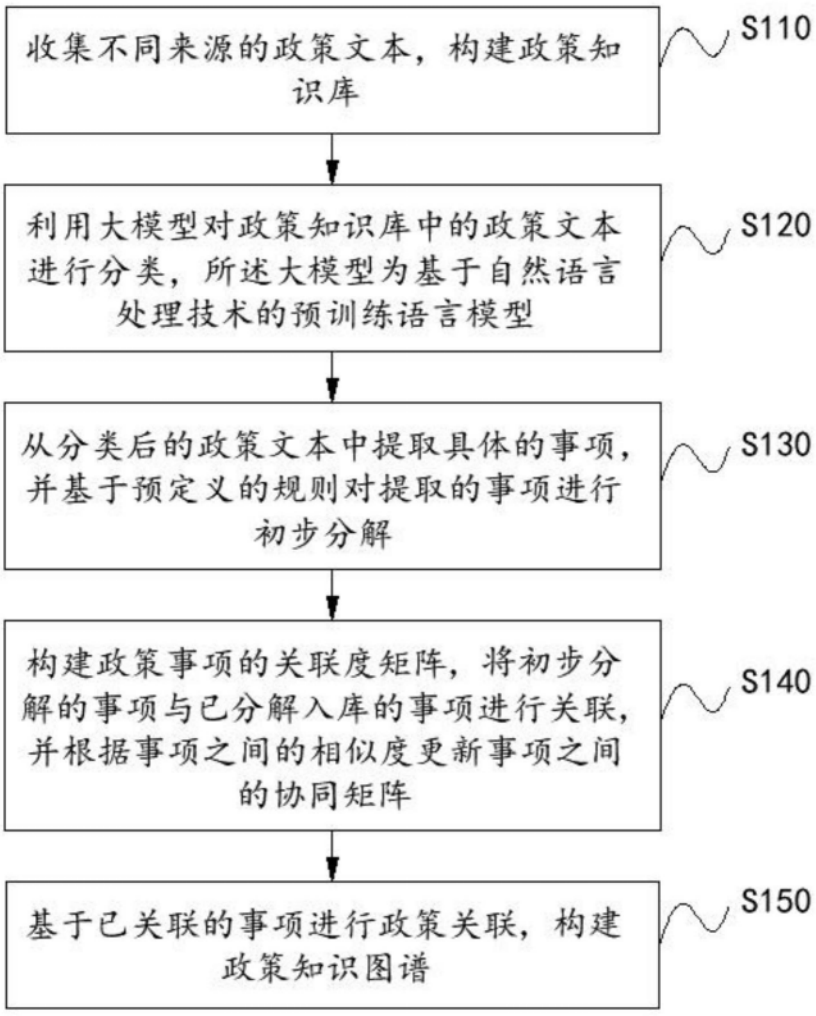

收集不同来源的政策文本，构建政策知识库
S110
利用大模型对政策知识库中的政策文本进行分类，所述大模型为基于自然语言处理技术的预训练语言模型
S120
从分类后的政策文本中提取具体的事项，并基于预定义的规则对提取的事项进行初步分解
S130
构建政策事项的关联度矩阵，将初步分解的事项与已分解入库的事项进行关联，并根据事项之间的相似度更新事项之间的协同矩阵
S140
基于已关联的事项进行政策关联，构建政策知识图谱
S150

图1

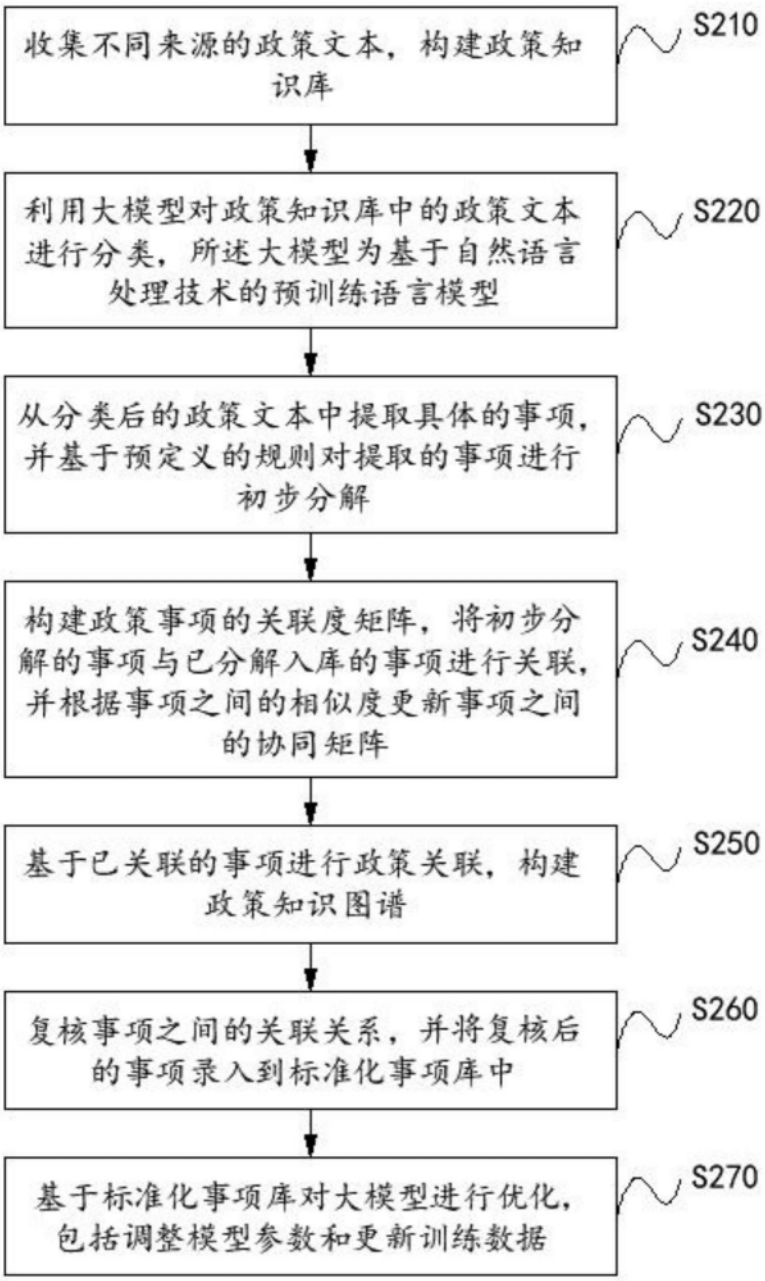

收集不同来源的政策文本，构建政策知识库
S210
利用大模型对政策知识库中的政策文本进行分类，所述大模型为基于自然语言处理技术的预训练语言模型
S220
从分类后的政策文本中提取具体的事项，并基于预定义的规则对提取的事项进行初步分解
S230
构建政策事项的关联度矩阵，将初步分解的事项与已分解入库的事项进行关联，并根据事项之间的相似度更新事项之间的协同矩阵
S240
基于已关联的事项进行政策关联，构建政策知识图谱
S250
复核事项之间的关联关系，并将复核后的事项录入到标准化事项库中
S260
基于标准化事项库对大模型进行优化，包括调整模型参数和更新训练数据
S270

图2

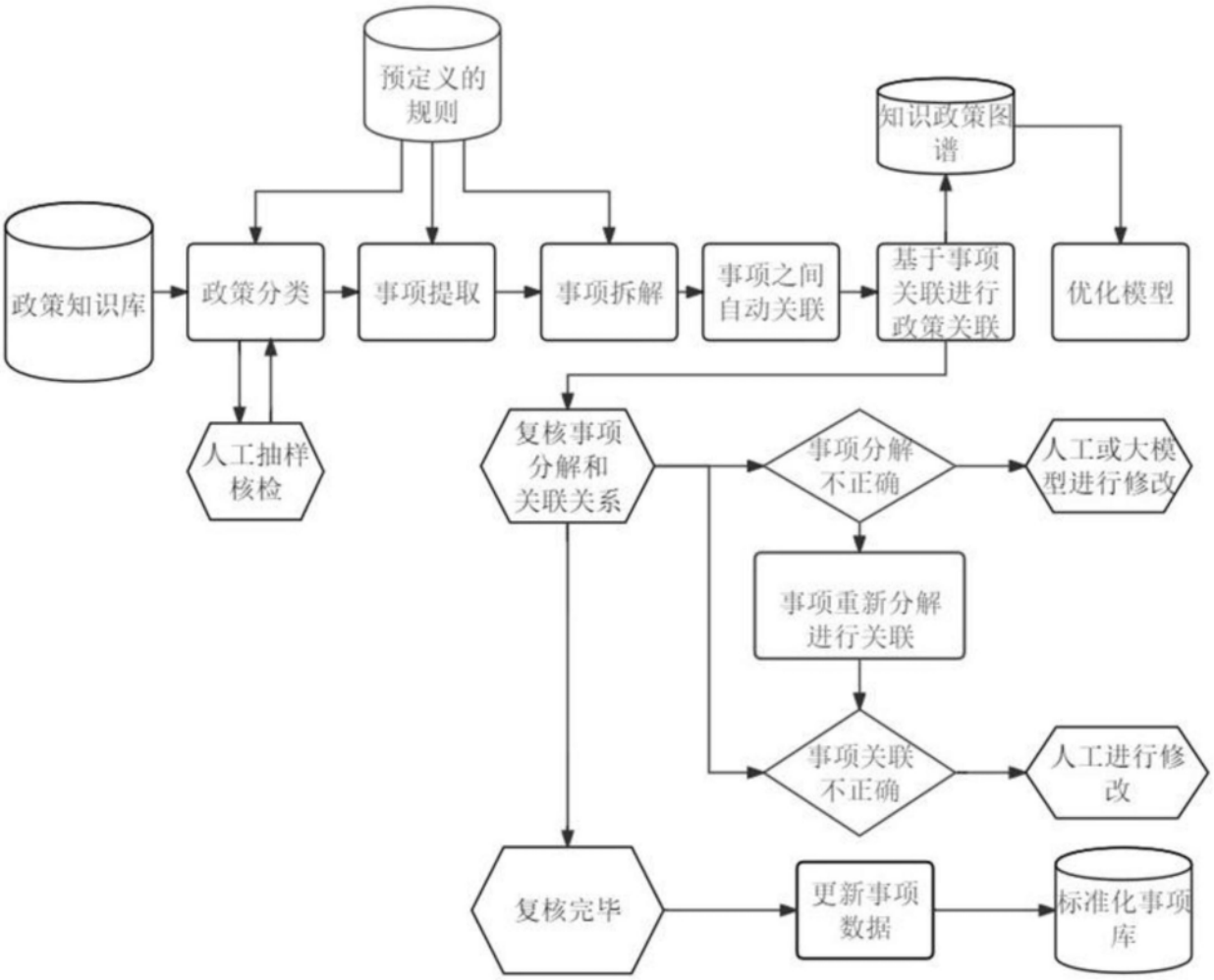

预定义的规则
知识政策图谱
政策知识库
政策分类
事项提取
事项拆解
事项之间自动关联
基于事项关联进行政策关联
优化模型
人工抽样核检
复核事项分解和关联关系
事项分解不正确
人工或大模型进行修改
事项重新分解进行关联
事项关联不正确
人工进行修改
复核完毕
更新事项数据
标准化事项库

图3

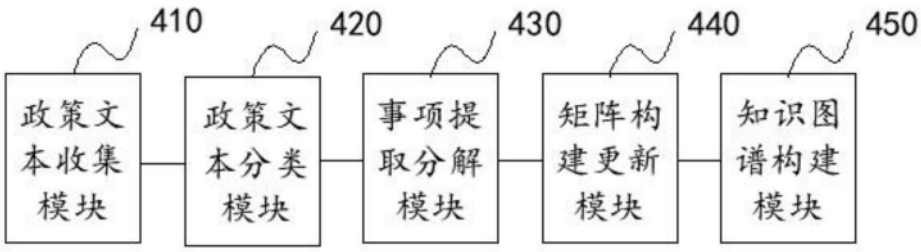

410
420
430
440
450
政策文本收集模块
政策文本分类模块
事项提取分解模块
矩阵构建更新模块
知识图谱构建模块

图4

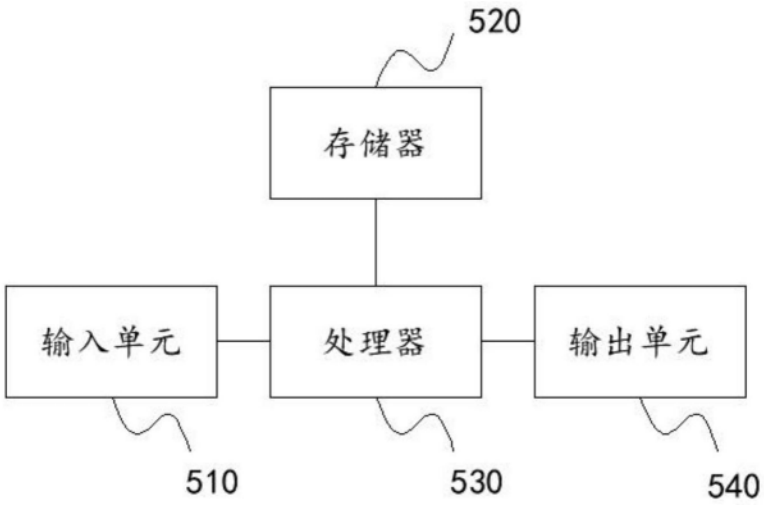


图5