(19)国家知识产权局

(12)发明专利

(10)授权公告号 CN 116703337 B
(45)授权公告日 2024.09.03

(21)申请号 202310986410.7

(22)申请日 2023.08.08

(65)同一申请的已公布的文献号
申请公布号 CN 116703337 A

(43)申请公布日 2023.09.05

(73)专利权人 金现代信息产业股份有限公司
地址 250000 山东省济南市高新区新泺大街1166号奥盛大厦2号楼21层东区2101

(72)发明人 黄学辉 刘栋 邵柄莱 陈磊 李长川 孙凯凯 李彤 刘滕飞

(74)专利代理机构 济南圣达知识产权代理有限公司 37221
专利代理师 张庆骞

(51)Int.Cl.
*G06Q 10/10*(2023.01)
*G06N 5/02*(2023.01)

(56)对比文件
CN 113704498 A,2021.11.26
CN 116303981 A,2023.06.23
CN 116468131 A,2023.07.21

审查员 张一欣

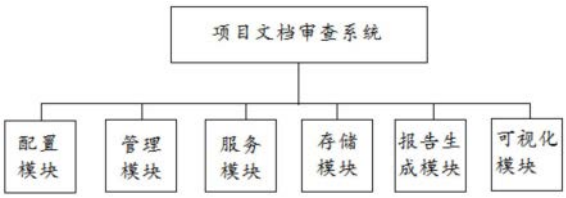
权利要求书2页 说明书10页 附图1页

(54)发明名称

一种基于人工智能技术的项目文档审查系统及方法

(57)摘要

本发明公开了一种基于人工智能技术的项目文档审查系统及方法,涉及文档审查技术领域。该系统包括:配置模块、管理模块和服务模块。本发明采用人工智能算法和自然语言处理技术训练模型,能够快速识别项目文档中存在的问题和缺陷,帮助用户更好地发现和解决潜在风险,提升项目的质量和可靠性,实现对项目文档的自动化审查,能够有效提高审查效率,提升审查质量和准确性。本发明文档审查系统还具有规范化和标准化特点,可以根据用户定义的审查规则和标准进行智能审查,帮助确保文档的合规性和符合行业标准。这有助于企业建立统一的项目审查流程,提高工作标准化水平,减少错误和遗漏。




CN 116703337 B

1.一种基于人工智能技术的项目文档审查系统,其特征在于,包括:配置模块、管理模块和服务模块,其中,

所述配置模块包括数据库配置模块,用于对项目数据源进行资源配置;

项目文档配置模块,用于对项目类型及项目关联的文档类型进行配置;

文档评审规则配置模块,用于根据审查任务配置的不同,按照项目类型配置文档的审查要点及审查项,形成不同的审查规则;

智能化配置模块,用于对自动审查模块的参数进行配置;

所述管理模块包括知识库管理模块,用于对审查相关知识文档进行管理;

项目知识图谱模块,用于基于项目文档配置,通过挖掘项目之间的潜在关联信息,构建项目知识图谱,并利用构建的项目知识图谱对项目进行查询、检索;

所述服务模块包括文档信息提取模块,用于根据审查任务要求提取文档中的关键信息;

大语言模型模块,用于采用机器学习的方法根据审查任务对应的审查规则进行大语言模型训练;

自动审查模块,用于根据自动审查模块配置的参数对提取的关键信息利用大语言模型进行自动审查;

所述对项目数据源进行资源配置包括配置数据库中数据源信息或数据接口信息,并且基于已配置的数据源,配置数据采集方式以及待采集数据和目标数据库的映射关系;

所述文档评审规则配置模块,支持按照项目类型配置文档的审查要点及审查项;支持规则库的自定义配置,可满足不同类型项目资料的审查需求;提供可视化的规则编排引擎,支持以画布拖拽方式配置审查项的详细判断条件,支持且、或逻辑,支持项目属性和文档关键信息“相等”、“不相等”、“包含”、“大于”、“小于”等运算逻辑配置,并支持以上内容的非空校验规则;

所述配置模块还包括文本提示配置模块,用于配置训练模型过程中的提示内容;

所述配置模块中,配置训练模型过程中的提示内容的具体过程为:

定义初始提示,用于描述文档审查任务的目标和要求;

定义文档内容提示,用于引导大语言模型理解文档内容和评审规则;

定义评审规则提示,用于将配置的文档审查规则转化为提示,引导大语言模型执行自动审查;

所述文档信息提取模块通过训练信息提取模型进行关键信息的提取;

所述通过大规模预训练的语言模型有效地实现专业文档中关键信息的提取;

具体步骤为:

文档表示学习:使用预训练的大语言模型,对输入文档进行表示学习;

上下文理解和关联:利用大语言模型的上下文理解能力,对文档中的句子、段落或整个文档进行理解和关联;

实体识别和关键信息提取:在获得文档的语义表示和上下文理解后,利用大语言模型进行实体识别和关键信息提取;

领域适应和迁移学习:为了适应不同领域和专业文档的特点,还引入了领域适应和迁移学习技术;

所述自动审查模块,结合LORA微调方法,将生成的Prompt输入配置好的大语言模型,并调用模型进行文档内容的自动审查;

模型根据Prompt的引导,结合文档内容和评审规则,分析文档中的潜在问题、错误或违规情况,最后解析出大语言模型的输出结果,识别出文档中存在的问题、错误或违规情况;根据评审规则的配置,提取出问题的类型、位置和详细描述信息;

用户可选择单个项目,按已配置的审查要点、审查项直观查询系统对该项目文档的自动审查结果。

2.如权利要求1所述的基于人工智能技术的项目文档审查系统,其特征在于,所述文档信息提取模块按照文档格式的分类采用不同的提取方式进行关键信息的提取。

3.如权利要求1所述的基于人工智能技术的项目文档审查系统,其特征在于,还包括存储模块,用于对系统内的信息进行存储;

所述存储模块包括文档存储模块,用于根据配置类型对文档进行存储;

规则存储模块,用于对配置的规则进行存储;

模型存储模块,用于对训练的模型进行存储。

4.如权利要求1所述的基于人工智能技术的项目文档审查系统,其特征在于,还包括报告生成模块,用于根据自动审查模块自动审查得到的结果,批量导出项目的审查报告。

5.如权利要求1所述的基于人工智能技术的项目文档审查系统,其特征在于,还包括可视化模块,用于将各个模块的审查过程在客户端进行可视化展示。

6.如权利要求1-5任一项所述的基于人工智能技术的项目文档审查系统的方法,其特征在于,包括以下步骤:

对项目文档进行配置,其中包括对项目数据源进行资源配置、对项目类型及项目关联的文档类型进行配置、对审查规则进行配置以及对自动审查过程的参数进行配置;

基于项目文档配置,通过挖掘项目之间的潜在关联信息,构建项目知识图谱,并利用构建的项目知识图谱对项目进行查询、检索;

根据审查任务要求提取文档中的关键信息;

采用机器学习的方法根据审查任务对应的审查规则进行大语言模型训练;

根据自动审查模块配置的参数对提取的关键信息利用大语言模型进行自动审查。

# 一种基于人工智能技术的项目文档审查系统及方法

## 技术领域

[0001] 本发明涉及文档审查技术领域,尤其涉及一种基于人工智能技术的项目文档审查系统及方法。

## 背景技术

[0002] 目前针对项目资料进行检查工作,需要组织人员对项目规划-储备-计划等各个阶段的文档进行审阅,其中字段一致性校验、签字盖章检查、命名规范性检查等重复性高、机械系强的检查项较多,人员成本耗费较高。

[0003] 当前的文档智能审查技术在语义理解方面存在一定的限制。尽管它们能够理解和解析文本中的语法和词汇,但对于上下文和语义的深层理解仍然存在挑战。尤其是在处理复杂、具有歧义性或特定领域知识的文档时会导致一些误解和错误的判断。另外,现有文档审查通常是基于预定义的规则和标准进行审核。然而,这些标准具有很强的主观性和灵活性,因为它们可能因行业、地区或组织而异。但现有技术往往难以准确地捕捉和理解这些主观标准,从而导致不一致的审查结果。另外,现有的文档智能审查技术通常是黑盒模型,其内部工作机制和决策过程缺乏透明性和解释性。这使得用户难以理解和解释审查结果的依据,从而降低了用户的信任度和可接受度。因此,如何实现高效且准确的交互式项目文档的自动审查,成为现有技术亟待解决的问题。

## 发明内容

[0004] 针对现有技术存在的不足,本发明的目的是提供一种基于人工智能技术的项目文档审查系统及方法,通过对文档审查的规则和标准的灵活的配置实现多种文档的智能审查,并且利用机器学习方法增强对文档上下文和语义的深层理解。审查的过程用户可参与其中,将机器学习与用户配置进行融合,真正实现了高效准确的交互式项目文档审查。

[0005] 为了实现上述目的,本发明是通过如下的技术方案来实现:

[0006] 本发明第一方面提供了一种基于人工智能技术的项目文档审查系统,包括:

[0007] 配置模块、管理模块和服务模块,其中,

[0008] 所述配置模块包括数据库配置模块,用于对项目数据源进行资源配置;

[0009] 项目文档配置模块,用于对项目类型及项目关联的文档类型进行配置;

[0010] 文档评审规则配置模块,用于根据审查任务配置的不同,按照项目类型配置文档的审查要点及审查项,形成不同的审查规则;

[0011] 智能化配置模块,用于对自动审查模块的参数进行配置;

[0012] 所述管理模块包括知识库管理模块,用于对审查相关知识文档进行管理;

[0013] 项目知识图谱模块,用于基于项目文档配置,通过挖掘项目之间的潜在关联信息,构建项目知识图谱,并利用构建的项目知识图谱对项目进行查询、检索;

[0014] 所述服务模块包括文档信息提取模块,用于根据审查任务要求提取文档中的关键信息;

[0015] 大语言模型模块,用于采用机器学习的方法根据审查任务对应的审查规则进行大语言模型训练;

[0016] 自动审查模块,用于根据自动审查模块配置的参数对提取的关键信息利用大语言模型进行自动审查。

[0017] 进一步的,对项目数据源进行资源配置包括配置数据库中数据源信息或数据接口信息,并且基于已配置的数据源,配置数据采集方式以及待采集数据和目标数据库的映射关系。

[0018] 进一步的,所述配置模块还包括文本提示配置模块,用于配置训练模型过程中的提示内容。

[0019] 更进一步的,文所述配置模块中,配置训练模型过程中的提示内容的具体过程为:

[0020] 定义初始提示,用于描述文档审查任务的目标和要求;

[0021] 定义文档内容提示,用于引导大语言模型理解文档内容和评审规则;

[0022] 定义评审规则提示,用于将配置的文档审查规则转化为提示,引导大语言模型执行自动审查。

[0023] 进一步的,所述文档信息提取模块按照文档格式的分类采用不同的提取方式进行关键信息的提取。

[0024] 进一步的,所述文档信息提取模块通过训练信息提取模型进行关键信息的提取。

[0025] 进一步的,还包括存储模块,用于对系统内的信息进行存储;

[0026] 所述存储模块包括文档存储模块,用于根据配置类型对文档进行存储;

[0027] 规则存储模块,用于对配置的规则进行存储;

[0028] 模型存储模块,用于对训练的模型进行存储。

[0029] 进一步的,还包括报告生成模块,用于根据自动审查模块自动审查得到的结果,批量导出项目的审查报告。

[0030] 进一步的,还包括可视化模块,用于将各个模块的审查过程在客户端进行可视化展示。

[0031] 本发明第二方面提供了一种基于人工智能技术的项目文档审查方法,包括以下步骤:

[0032] 对项目文档进行配置,其中包括对项目数据源进行资源配置、对项目类型及项目关联的文档类型进行配置、对审查规则进行配置以及对自动审查过程的参数进行配置;

[0033] 基于项目文档配置,通过挖掘项目之间的潜在关联信息,构建项目知识图谱,并利用构建的项目知识图谱对项目进行查询、检索;

[0034] 根据审查任务要求提取文档中的关键信息;

[0035] 采用机器学习的方法根据审查任务对应的审查规则进行大语言模型训练;

[0036] 根据自动审查模块配置的参数对提取的关键信息利用大语言模型进行自动审查。

[0037] 以上一个或多个技术方案存在以下有益效果:

[0038] 本发明公开了一种基于人工智能技术的项目文档审查系统及方法,能够有效提高审查效率,提升审查质量和准确性。本发明采用人工智能算法和自然语言处理技术训练模型,能够快速识别项目文档中存在的问题和缺陷,如内容不合规、数据不一致、信息遗漏等,这帮助用户更好地发现和解决潜在风险,提升项目的质量和可靠性,实现对项目文档的自

动化审查。

[0039] 本发明文档审查系统具有规范化和标准化特点,可以根据用户定义的审查规则和标准进行智能审查,帮助确保文档的合规性和符合行业标准。这有助于企业建立统一的项目审查流程,提高工作标准化水平,减少错误和遗漏。

[0040] 本发明附加方面的优点将在下面的描述中部分给出,部分将从下面的描述中变得明显,或通过本发明的实践了解到。

## 附图说明

[0041] 构成本发明的一部分的说明书附图用来提供对本发明的进一步理解,本发明的示意性实施例及其说明用于解释本发明,并不构成对本发明的不当限定。

[0042] 图1为本发明实施例一中基于人工智能技术的项目文档审查系统框架图;

[0043] 图2为本发明实施例二中基于人工智能技术的项目文档审查方法流程图。

## 具体实施方式

[0044] 应该指出,以下详细说明都是示例性的,旨在对本发明提供进一步的说明。除非另有指明,本文使用的所有技术和科学术语具有与本发明所属技术领域的普通技术人员通常理解的相同含义。

[0045] 应当说明的是,本发明实施例中,涉及到项目文档等相关的数据,当本发明以上实施例运用到具体产品或技术中时,需要获得用户许可或者同意,且相关数据的收集、使用和处理需要遵守相关国家和地区的相关法律法规和标准。

[0046] 需要注意的是,这里所使用的术语仅是为了描述具体实施方式,而非意图限制根据本发明的示例性实施方式。如在这里所使用的,除非上下文另外明确指出,否则单数形式也意图包括复数形式,此外,还应当理解的是,当在本说明书中使用术语“包含”和/或“包括”时,其指明存在特征、步骤、操作、器件、组件和/或它们的组合;

[0047] 术语解释:

[0048] 1.LLM:大语言模型(Large Language Model,LLM),也称大型语言模型,是一种人工智能模型,旨在理解和生成人类语言。

[0049] 2.OCR:OCR(optical character recognition)文字识别是指电子设备(例如扫描仪或数码相机)检查纸上打印的字符,然后用字符识别方法将形状翻译成计算机文字的过程。

[0050] 3.NLP:NLP(Natural Language Processing)为自然语言处理,是一门研究计算机与人类自然语言之间交互的学科领域。它涉及使用计算机来理解、处理、生成和操纵人类语言的方法和技术。

[0051] 4.Prompt:在大语言模型技术中,"Prompt"(提示)是指输入给模型的文本片段或问题,用于引导模型生成相应的输出。Prompt可以是一个完整的句子、一个问题、一个描述性的开头,甚至只是一些关键词。

[0052] 实施例一:

[0053] 本发明实施例一提供了一种基于人工智能技术的项目文档审查系统,如图1所示,包括配置模块、管理模块、服务模块、存储模块、报告生成模块和可视化模块。

[0054] 其中,配置模块包括数据库配置模块、项目文档配置模块、文档评审规则配置模块、智能化配置模块和文本提示配置模块。

[0055] 数据库配置模块用于对项目数据源进行资源配置;具体的,对项目数据源进行资源配置包括配置数据库中数据源信息或数据接口信息,实现项目信息采集的数据源配置。

[0056] 在一种具体的实施方式中,可支持数据源的新增、编辑和删除,支持配置mysql、oracle、pgsql等数据库的连接地址、端口、用户名和密码信息,以及get/post接口的请求参数和响应格式等信息。

[0057] 基于已配置的数据源,配置数据采集方式以及待采集数据和目标数据库的映射关系。

[0058] 在一种具体的实施方式中,支持以在线编写sql的方式从源数据库检索需要同步的项目信息。支持数据采集任务的创建、编辑和删除,支持sql代码的在线编写、预览和运行,可导出查询的结果集。

[0059] 项目文档配置模块,用于对项目类型及项目关联的文档类型进行配置。

[0060] 在一种具体的实施方式中,支持配置项目类型及项目关联的文档类型,文档格式支持doc、docx、wps文档,excel表格及PDF扫描件。根据已配置的项目文档类型,选择项目并上传相应的项目资料。支持项目文档的上传、预览、下载和删除。

[0061] 文档评审规则配置模块,用于根据审查任务配置的不同,按照项目类型配置文档的审查要点及审查项,形成不同的审查规则。

[0062] 在一种具体的实施方式中,支持按照项目类型配置文档的审查要点及审查项。支持规则库的自定义配置,可满足不同类型项目资料的审查需求。

[0063] 提供可视化的规则编排引擎,支持以画布拖拽方式配置审查项的详细判断条件,支持且、或逻辑,支持项目属性和文档关键信息“相等”、“不相等”、“包含”、“大于”、“小于”等运算逻辑配置,并支持以上内容的非空校验规则。

[0064] 除应用规则引擎进行上述文档评审规则配置模块外,用户可以根据自身的文档评审规则,在文本提示配置模块配置大语言模型的prompt,实现自动审查功能。

[0065] 智能化配置模块,用于对自动审查模块的参数进行配置。

[0066] 文本提示配置模块,用于配置训练模型过程中的提示(prompt)内容,定义文档评审规则,包括格式要求、合规性要求、内容要求等。对不同类型的文档,可以配置相应的审查规则集合,以适应不同的审查需求。

[0067] 在一种具体的实施方式中,配置训练模型过程中的提示内容的具体过程为:

[0068] 定义初始提示(Initial Prompt),用于描述文档审查任务的目标和要求。

[0069] 定义文档内容提示(Document Content Prompt),用于引导大语言模型理解文档内容和评审规则。

[0070] 定义评审规则提示(Review Rule Prompt),用于将配置的文档审查规则转化为提示,引导大语言模型执行自动审查。

[0071] 管理模块包括知识库管理模块和项目知识图谱模块。

[0072] 知识库管理模块用于对审查相关知识文档进行管理。

[0073] 在一种具体的实施方式中,支持项目管理条例、指导意见、审查要点等审查相关知识文档的管理,支持文档的上传、审核、下载、删除,以及全文检索、在线预览功能。

[0074] 项目知识图谱模块,用于通过挖掘项目之间的潜在关联信息,构建项目知识图谱,并利用构建的项目知识图谱对项目进行查询、检索。

[0075] 在一种具体的实施方式中,挖掘项目之间的潜在关联信息如专业、规模成效、立项依据、关联设备等。支持对项目内容进行相似度比对。辅助发现重复立项。

[0076] 服务模块包括文档信息提取模块、大语言模型模块和自动审查模块。文档信息提取模块提取文档中的关键信息。

[0077] 文档信息提取模块可通过两种方法根据审查任务要求提取文档中的关键信息:

[0078] 方法一,文档信息提取模块按照文档格式的分类采用不同的提取方式进行关键信息的提取。

[0079] 在一种具体的实施方式中,对word(doc、docx、wps)格式的项目文档,支持通过在线配置提取章节、文本、数值和日期等信息。通过章节名称的相似度及短文本匹配设置,可以提取文档中的章节信息;通过正则表达式、表格元素配置、算法模型调用等方式,实现文本、数值、日期等关键信息的提取和存储。

[0080] 在一种具体的实施方式中,以提取“项目必要性”章节为例,需要预先维护一批与目标章节名称类似的短文本,比如“必要性”、“必要性分析”、“项目必需性”等。在执行章节提取任务时,首先用python-docx等工具将word文档转为纯文本,去除特殊格式、标点符号和标记。再基于配置中维护的这些短文本,利用文本匹配方法或Word2Vec、FastText模型来计算文本相似度,提取相似度较高的词或短句,再通过制定规则约束(比如筛选掉不包含大写或小写数字的词或短句),提取出所需的章节名称。

[0081] 对excel(xls、xlsx)格式的项目文档,支持通过在线配置提取sheet页、文本、数值和日期等信息。通过sheet页名称匹配或序列匹配,可以定位到目标sheet页;通过行定位器、列定位器及偏移量配置,提取目标单元格的信息。

[0082] 对pdf扫描件类型的文档,通过配置后台OCR服务及返回键值,可从文档中提取签字、盖章、文本、数值、日期和表格等信息。

[0083] 方法二,文档信息提取模块通过训练信息提取模型进行关键信息的提取:

[0084] 除了方法一中传统文档信息提取方法,本实施例还包括配置大语言模型的prompt,定义输出参数格式,实现文档信息提取的功能。具体功能包括:

[0085] 1、输入处理:接收用户提供的文档内容或文档路径,并将其格式化为模型可接受的输入格式。

[0086] 2、Prompt生成:根据用户需求和任务进行配置,具体为根据Prompt配置规则生成完整的Prompt,包括初始提示、输入文档提示、提取目标提示和上下文提示。

[0087] 3、大语言模型调用:将生成的Prompt作为输入传递给配置好的大语言模型,并调用模型进行文档信息提取。可以利用生成模型的特性,如条件生成或文本分类,生成结构化的输出结果。

[0088] 4、输出解析:根据预先定义的输出参数格式,解析大语言模型的输出结果,并提取出提取的实体、关键信息、关联信息和附加信息。

[0089] 5、结果返回:将解析后的结果以适当的格式返回给用户,如JSON格式、表格格式或自定义的数据结构。

[0090] 大语言模型模块可以支持用户与审查结果进行交互式查询和反馈。本发明应用的

大语言模型为清华开源的ChatGLM-6B模型,ChatGLM-6B 是一个支持中英双语的对话语言模型,基于 GLM (General Language Model),具有 62 亿个参数。该模型采用的是编码器-解码器架构,具备以下特点:

[0091] (1)充分的中英双语预训练:ChatGLM-6B 在 1:1 比例的中英语料上训练了 1T 的 token 量,兼具双语能力。

[0092] (2)优化的模型架构和大小:吸取 GLM-130B 训练经验,修正了二维 RoPE 位置编码实现,使用传统FFN结构。6B(62亿)的参数大小,也使得研究者和个人开发者自己微调和部署 ChatGLM-6B 成为可能。

[0093] (3)较低的部署门槛:FP16 半精度下,ChatGLM-6B 需要至少 13GB 的显存进行推理,结合模型量化技术,这一需求可以进一步降低到 10GB(INT8) 和 6GB(INT4), 使得 ChatGLM-6B 可以部署在消费级显卡上。

[0094] (4)更长的序列长度:相比 GLM-10B(序列长度1024),ChatGLM-6B 序列长度达 2048,支持更长对话和应用。

[0095] (5)人类意图对齐训练:使用了监督微调(Supervised Fine-Tuning)、反馈自助(Feedback Bootstrap)、人类反馈强化学习(Reinforcement Learning from Human Feedback) 等方式,使模型初具理解人类指令意图的能力。输出格式为 markdown,方便展示。

[0096] 因此,ChatGLM-6B 具备了一定条件下较好的对话与问答能力。用户可以向模型提出特定问题,要求进一步解释模型的决策依据。模型可以根据用户的查询生成相应的解释性文本或展示相关的证据。通过这种方式,用户可以主动参与解释过程,获得更准确和有针对性的解释。相比传统的人工审查方法,这种方法可以大幅提高审查速度和效率,节省大量人力资源。

[0097] 自动审查模块,用于根据自动审查模块配置的参数对提取的关键信息利用大语言模型进行自动审查。本实施例通过训练机器学习模型,并结合大语言模型的propmt及微调,形成专业化的自动审查算法。本发明采用LORA(Low-Rank Adaptation)微调方法,实现原理在于,冻结预训练模型权重,并将可训练的秩分解矩阵注入到Transformer层的每个权重中,大大减少了下游任务的可训练参数数量。

[0098] 在一种具体的实施方式中,以项目建议书中“设备现状”内容的评估为例。首先,从大量历史文档中获取关于“设备现状”的描述,通过人工标注或规则约束,将文本标注为“高质量”或“低质量”,构建训练数据集和验证集;使用文本分类的机器学习算法,如朴素贝叶斯分类器,使用训练集对模型进行训练,并在验证集上进行模型性能评估和调优;使用训练好的模型对新的“设备现状”的描述进行自动评估,输入文本数据到模型中,模型会输出评估结果,如“高质量”或“低质量”。

[0099] 在具体的微调实施方式中,以设备现状内容的评估为例,需要收集业务相关的文本数据,如项目建议书、可研报告等,并对原始文本数据进行清洗和标注,整理成带有相关prompt的对话或问题数据集,示例如下:

[0100] (1)文本:“1号主变高压侧1101断路器型号为LW250,生产厂家西安高压开关厂,投运于2000年。”

[0101] 问:“以上内容是否按规定描述了设备的生产厂家?”

[0102] 答:“是的,1号主变高压侧1101断路器的生产厂家是西安高压开关厂。”

[0103] (2)文本:“1号主变高压侧1101断路器型号为LW250,生产厂家西安高压开关厂。”

[0104] 问:“以上内容是否按规定描述了设备的投运时间?”

[0105] 答:“上述内容未描述1号主变高压侧1101断路器的投运时间,因此不符合规定。”

[0106] 通过以上步骤,结合将LORA算法应用于大模型微调,可以提高模型在内容审查方面的性能和效果。

[0107] 在一种具体的实施方式中,将生成的Prompt输入配置好的大语言模型,并调用模型进行文档内容的自动审查。模型根据Prompt的引导,结合文档内容和评审规则,分析文档中的潜在问题、错误或违规情况。最后解析出大语言模型的输出结果,识别出文档中存在的问题、错误或违规情况。根据评审规则的配置,提取出问题的类型、位置和详细描述等信息。

[0108] 更为具体的,用户可选择单个项目,按已配置的审查要点、审查项直观查询系统对该项目文档的自动审查结果。可通过审查项类型、审查结果对列表数据进行筛选,列表展示当前选中的审查要点下的全部审查项信息,包括审查是否通过及具体的原因。

[0109] 用户可通过文档预览视图,对项目关联的文档进行在线预览,可在左侧列表切换文档目录,右侧展示系统从该文档中自动提取的关键信息和相关审查结果,展示审查通过或不通过的原因,可通过点击关键字(超链接)自动定位的文档的相应段落。

[0110] 存储模块用于对系统内的信息进行存储;存储模块包括文档存储模、规则存储模块和模型存储模块。

[0111] 文档存储模块,用于根据配置类型对文档进行存储;

[0112] 规则存储模块,用于对配置的规则进行存储;

[0113] 模型存储模块,用于对训练的模型进行存储。

[0114] 报告生成模块,用于根据自动审查模块自动审查得到的结果,批量导出项目的审查报告。

[0115] 在一种具体的实施方式中,审查报告格式为excel或pdf。内容包括审查要点、审查项、审查项类型(必改、疑似、提示)、审查内容、审查结果(通过、不通过)及整改意见。

[0116] 可视化模块,用于将各个模块的审查过程在客户端进行可视化展示。

[0117] 本实施例借助OCR、NLP、LLM技术,能够快速而准确地从文档中提取关键信息,并智能化地进行分析和判断。这减少了人为主观因素的干扰,提高了审查结果的一致性、可靠性和准确性。大语言模型可以生成解释性文本来解释审查结果。当一个文档的内容存在合规性问题时,模型可以生成一段解释性的文本,解释为什么该文档被判定为有问题,并提供相关的法规、政策或规范作为支持。这样的解释可以帮助用户理解模型的决策过程,增加可信度和可接受性。

[0118] 实施例二:

[0119] 本发明实施例二提供了一种基于人工智能技术的项目文档审查方法,如图2所示,包括对项目文档进行配置,构建项目知识图谱,根据审查任务要求提取文档中的关键信息,进行大语言模型训练,并利用训练好的大语言模型进行自动审查,最后根据审查结果生成评估报告。具体包括以下步骤:

[0120] 步骤1,对项目文档进行配置,其中包括对项目数据源进行资源配置、对项目类型及项目关联的文档类型进行配置、对审查规则进行配置以及对自动审查过程的参数进行配

置。

[0121] 通过动态配置数据同步服务接口,从运行中的项目管理系统中自动同步项目基本属性和文件信息。通过配置定时任务,可实现项目相关信息的定期增量更新,保证自动审查结果的时效性。在同步项目信息的同时,基于项目本体库自动构建项目知识图谱,并生成项目之间的关联信息。

[0122] 步骤2,基于项目文档配置,通过挖掘项目之间的潜在关联信息,构建项目知识图谱,并利用构建的项目知识图谱对项目进行查询、检索。

[0123] 步骤3,根据审查任务要求提取文档中的关键信息。

[0124] 依据后台配置逻辑,从文档中获取待审查的关键信息,包括文本(数值、日期)、标题、段落、图片等。传统的文件关键信息提取方法包括:相似度匹配、实体识别、正则匹配、OCR目标检测、上下文定位法等等。然而,这些方法存在局限性,如对语义理解的不足、准确性受限和处理速度较慢等。因此,需要一种新的方法来优化和改良文档信息提取过程。本实施例在以上提取方法的基础上进行了改良,通过大规模预训练的语言模型有效地实现专业文档中关键信息的提取。

[0125] 具体步骤为:

[0126] 1、文档表示学习:首先,使用预训练的大语言模型,对输入文档进行表示学习。通过将文档转化为词嵌入或句子嵌入表示,大语言模型可以捕捉到文本中的丰富语义信息。

[0127] 2、上下文理解和关联:利用大语言模型的上下文理解能力,对文档中的句子、段落或整个文档进行理解和关联。通过考虑上下文信息,大语言模型可以更好地把握实体之间的关系、事件的发展以及特定领域的背景知识。

[0128] 3、实体识别和关键信息提取:在获得文档的语义表示和上下文理解后,利用大语言模型进行实体识别和关键信息提取。通过对文档进行标注或生成模型的训练,大语言模型可以准确地识别出文档中的关键实体,如人物、地点、事件等,并提取出与该实体相关联的重要信息。例如,针对“某变电站内装设220kV主变压器2台,容量均为180MVA,均为2018年投运”这段文字,需要提取变电站、设备、容量、投运时间等关键信息,大语言模型可以一次性提取这些关键信息,不仅限于实体识别,还能提取出相关属性和上下文信息。相比之下,传统方法需要分别进行实体识别和属性提取,可能需要额外的规则和步骤,需要手动标注大量的训练数据,且无法轻松适应新的数据集或文本类型。另外,通过不断调整和改进prompt的配置,可以提高大语言模型的准确性和适应性。而传统方法则需要手动调整和优化规则或模型,工作量较大且较为繁琐。

[0129] 4、领域适应和迁移学习:为了适应不同领域和专业文档的特点,本方法还引入了领域适应和迁移学习技术。通过在特定领域的数据集上微调大语言模型,使其能够更好地理解和提取该领域的专业术语、短语和上下文关系,从而提高文档信息提取的准确性和适应性。

[0130] 在一种具体的实施方式中,1号主变高压侧1101断路器型号为LW250,生产厂家西安高压开关厂,投运于2000年,如果想从这段文本中提取设备型号、投运年限、生产厂家等关键信息,可以通过构建合适的prompt来实现,构建一个包含关键词的问题模板,以引导模型正确回答问题,例如:“请问该设备的型号、投运年限和生产厂家是什么?”或者让模型按照json、xml等结构化格式返回结果,方便对数据进行解析。

[0131] 为了提升大模型对这些关键信息提取的完整率和准确率,本发明采用基于LORA(Low-Rank Adaptation)的大模型微调方法,收集包含设备信息的相关文本数据,如设备说明书、技术规范等,并对原始文本数据进行清洗和标注,将设备型号、投运年限和生产厂家等关键信息用特定的标记进行标注,形成类似于下面的格式:

[0132] 输入文本:1号主变高压侧1101断路器型号为LW250,生产厂家西安高压开关厂,投运于2000年。

[0133] 标注文本:设备型号为<LW250>,生产厂家为<西安高压开关厂>,投运年限为<2000年>。

[0134] 再通过Q&A数据集构建工具,将已标注的文本组织成带有相关prompt的对话或问题数据集,对话或问题数据集形式例如:

[0135] (1)文本“1号主变高压侧1101断路器型号为LW250,生产厂家西安高压开关厂,投运于2000年。”

[0136] 问:“设备型号是什么?”

[0137] 答:“1号主变高压侧1101断路器,型号为LW250。”

[0138] (2)文本“1号主变高压侧1101断路器型号为LW250,生产厂家西安高压开关厂,投运于2000年”

[0139] 问:“设备什么时候投运的?”

[0140] 答:“该设备于2000年投运。”

[0141] 通过以上步骤,结合将LORA算法应用于大模型微调,可以提高模型在设备信息提取任务上的性能和效果。

[0142] 5、结果生成和评估:根据实际应用需求,利用大语言模型生成提取结果。可以使用生成模型,如条件语言模型或生成对抗网络(GAN),结合大语言模型的生成能力,生成具有一定结构的提取结果,使其更符合实际应用场景。同时,还可以采用人工标注或领域专家评估等方法,对提取结果进行评估和优化。

[0143] 步骤4,采用机器学习的方法根据审查任务对应的审查规则进行大语言模型训练,根据自动审查模块配置的参数对提取的关键信息利用大语言模型进行自动审查。

[0144] 针对从文档中提取出的结构化信息,传统方法是基于规则引擎,应用相应的规则表达式或逻辑,验证字段是否满足要求;或应用机器学习技术,对待审查的文档进行分类、标记或打分,依据模型给出的评估分数和特定的后处理规则(人工干预),给出文档的自动审查结果。本实施例在以上传统方法的基础上,提供了应用大语言模型进行内容审查的思路:

[0145] 首先定义评估对象和Prompt,确定需要评估的文档内容或特定领域的语义要求,定义评估对象的相关特征、属性或要求,并设计Prompt,即向语言模型提供的指令或问题,以引导其对文档内容进行分析和评估;将文档内容作为输入数据,调用API或相应的库函数,传递给大语言模型进行语义理解和分析;获取语言模型生成的输出结果,对输出结果进行解析和处理,提取有关评估对象相关的信息和语义理解结果。

[0146] 以下是一种具体的实现方式:在程序中配置文档审查规则时,选择调用大语言模型的审查方式,并配置输入的指令参数和出参协议,例如“请帮我按照规定的json格式输出项目内容中描述的设备现状、存在问题和方案规模”,将预处理后的文档分段后,将需要审

查的“项目内容”部分输入到模型中,调用模型的生成功能,以下是大模型输出的示例:

[0147] “设备现状”:”1号主变高压侧1101断路器型号为LW250,生产厂家西安高压开关厂,投运于2000年”,

[0148] “存在问题”,”设备投运年限长,部件老化,存在安全隐患”,

[0149] “方案规模”,”无”

[0150] 基于大语言模型输出的结果,可通过规则引擎,判断关键信息项是否为空、是否包含关键字;也可以通过大语言模型的语义理解能力,再次调用模型,输入对关键内容的审查要求,例如“请分析以上内容中提取到的‘存在问题’,是否包含具体投运时间等量化描述,而非‘投运时间长’等模糊性描述”。大模型会根据输入内容响应结果,响应内容可作为智能审查规则引擎的补充提示,和规则审查内容一并在前端展示给用户。

[0151] 步骤5,根据审查结果生成评估报告。

[0152] 对于系统中存储的审查内容、审查项和审查结果等结构化数据,进行数据包装并,采用模板生成的方式或者动态组织语言生成审查报告。

[0153] 以上实施例二中涉及的各步骤与方法实施例一相对应,具体实施方式可参见实施例一的相关说明部分。

[0154] 上述虽然结合附图对本发明的具体实施方式进行了描述,但并非对本发明保护范围的限制,所属领域技术人员应该明白,在本发明的技术方案的基础上,本领域技术人员不需要付出创造性劳动即可做出的各种修改或变形仍在本发明的保护范围以内。

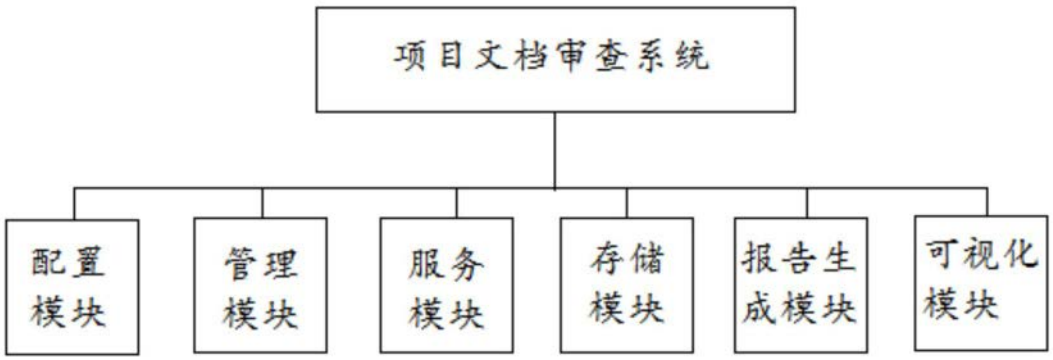

项目文档审查系统
配置模块
管理模块
服务模块
存储模块
报告生成模块
可视化模块

图1

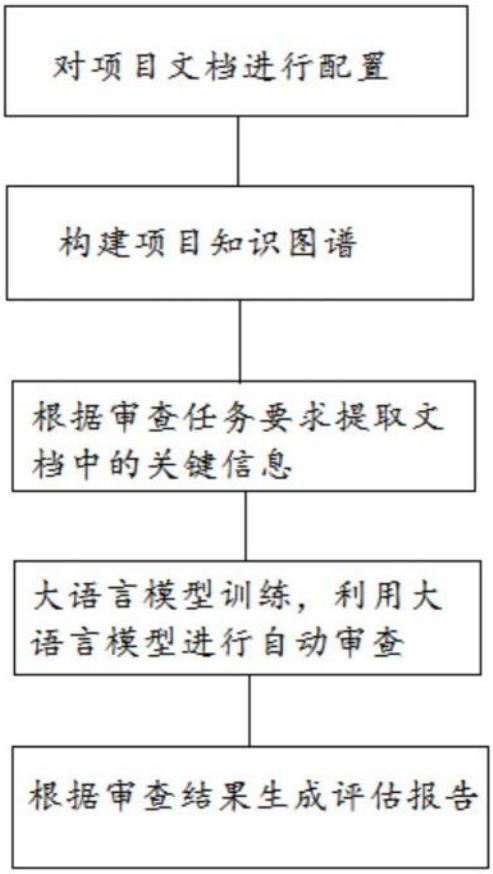

对项目文档进行配置
构建项目知识图谱
根据审查任务要求提取文档中的关键信息
大语言模型训练，利用大语言模型进行自动审查
根据审查结果生成评估报告

图2