(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 113377963 A
(43)申请公布日 2021.09.10

(21)申请号 202110719605.6

(22)申请日 2021.06.28

(71)申请人 中国科学院地质与地球物理研究所
地址 100029 北京市朝阳区北土城西路19号

(72)发明人 田飞 底青云 郑文浩 王中兴 杨永友 张文秀 裴仁忠

(74)专利代理机构 北京三友知识产权代理有限公司 11127
代理人 张德斌 姚亮

(51)Int.Cl.
G06F 16/36(2019.01)
G06F 40/253(2020.01)
G06F 40/295(2020.01)
G06Q 50/02(2012.01)

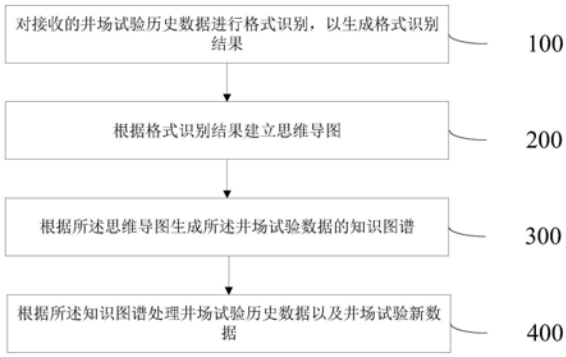
权利要求书2页 说明书13页 附图8页

(54)发明名称

一种基于知识图谱的井场试验数据处理方法及装置

(57)摘要

本发明提供了一种基于知识图谱的井场试验数据处理方法及装置，基于知识图谱的井场试验数据处理方法包括：对接收的井场试验历史数据进行格式识别，以生成格式识别结果；根据格式识别结果建立思维导图；根据所述思维导图生成所述井场试验数据的知识图谱；根据所述知识图谱处理井场试验历史数据以及井场试验新数据。本发明提供的基于知识图谱的井场试验数据处理方法及装置，可实现井场试验全流程的数据存储、管理、共享、查询、集中展示等功能，建立数据之间关系，提高数据查询效率，为技术攻关提供支撑。



CN 113377963 A

1.一种基于知识图谱的井场试验数据处理方法,其特征在于,包括:

对接收的井场试验历史数据进行格式识别,以生成格式识别结果;

根据格式识别结果建立思维导图;

根据所述思维导图生成所述井场试验数据的知识图谱;

根据所述知识图谱处理井场试验历史数据以及井场试验新数据。

2.如权利要求1所述的基于知识图谱的井场试验数据处理方法,其特征在于,所述对接收的井场试验历史数据进行格式识别,以生成格式识别结果,包括:

接收用户的井场试验数据处理请求;

从所述处理请求中提取文件名、操作对象类别以及文件格式;

扫描目标目录是否存在与所述文件格式相应的文件夹结构对象,以生成所述格式识别结果。

3.如权利要求2所述的基于知识图谱的井场试验数据处理方法,其特征在于,所述根据格式识别结果建立思维导图包括:

根据所述格式识别结果确定所述井场试验数据的关键词;

根据多个关键词以及预设的井场试验术语词典建立具有多层级关系的数据存储库;

根据所述数据存储库建立所述思维导图。

4.如权利要求3所述的基于知识图谱的井场试验数据处理方法,其特征在于,所述识别结果包括:结构化数据、半结构化数据以及非结构化数据;

所述根据所述格式识别结果确定所述井场试验数据的关键词,包括:

对所述结构化数据进行语法分析,以确定所述结构化数据的关键词;

对所述半结构化数据以及非结构化数据的标签进行标定,以确定所述半结构化数据以及非结构化数据的关键词。

5.如权利要求4所述的基于知识图谱的井场试验数据处理方法,其特征在于,所述对所述结构化数据进行语法分析,以确定所述结构化数据的关键词,包括:

根据所述井场试验术语词典对所述结构化数据进行术语抽取;

在抽取结果中,选取出现频数大于预设次数的术语;

根据所述出现频数大于预设次数的术语生成所述结构化数据的特征向量;

根据所述特征向量生成所述结构化数据的关键词。

6.如权利要求4所述的基于知识图谱的井场试验数据处理方法,其特征在于,所述对所述半结构化数据以及非结构化数据的标签进行标定,以确定所述半结构化数据以及非结构化数据的关键词,包括:

计算所述半结构化数据以及非结构化数据与所述井场试验术语词典的字面文本相似度;

根据所述字面文本相似度,从所述半结构化数据以及非结构化数据中选取一部分进行标定。

7.如权利要求1所述的基于知识图谱的井场试验数据处理方法,其特征在于,所述根据所述思维导图生成所述井场试验数据的知识图谱包括:

根据所述思维导图对所述格式识别结果进行粒度实体识别,以生成识别结果;

根据所述识别结果建立所述井场试验数据的知识层级;

根据所述识别结果提取所述井场试验数据的实体数据；

根据所述知识层级以及实体数据生成所述知识图谱。

8.一种基于知识图谱的井场试验数据处理装置，其特征在于，包括：

识别结果生成模块，用于对接收的井场试验历史数据进行格式识别，以生成格式识别结果；

思维导图建立模块，用于根据格式识别结果建立思维导图；

知识图谱生成模块，用于根据所述思维导图生成所述井场试验数据的知识图谱；

数据处理模块，用于根据所述知识图谱处理井场试验历史数据以及井场试验新数据。

9.一种电子设备，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，其特征在于，所述处理器执行所述程序时实现权利要求1至7任一项所述基于知识图谱的井场试验数据处理方法的步骤。

10.一种计算机可读存储介质，其上存储有计算机程序，其特征在于，该计算机程序被处理器执行时实现权利要求1至7任一项所述基于知识图谱的井场试验数据处理方法的步骤。

一种基于知识图谱的井场试验数据处理方法及装置

技术领域

[0001] 本发明涉及石油天然气钻探技术领域，具体涉及一种基于知识图谱的井场试验数据处理方法及装置。

背景技术

[0002] 现有技术中，井场试验流程包括试验规划、仪器装配、性能测试、下井前功能测试、实钻井试验、试验总结等，不仅流程众多，并且每个流程存在对应规范标准、研究材料、总结报告、参考文献、试验数据等结构化、非结构化和半结构化数据，可见在整个井场试验流程中，存在数据文档管理复杂，查找有价值文档困难的技术难点，从而导致个人成果记录及试验经验缺乏规范性管理等问题，直接导致知识及有价值的文档无法快速定位、共享、应用。

[0003] 近些年，信息技术的发展及应用改变了企业文档管理的方式，在提升管理效率的同时也保证了文档资料的多样性与完整性。目前涌现出了一些专业的文档及知识库管理方法(平台)，例如xyplorer、tagLyst、语雀等；同时一些思维导图的应用软件也广泛发展，例如Xmind、MindMaster等。参见表1，当前流行的文档、知识管理方法偏向大众化应用，各有优缺点、缺少针对性定制，难以对知识图谱与文件管理进行有效融合，无法解决石油行业井场试验中多种仪器、大量数据、文档及对应知识的管理。

[0004] 表1

[0005]

软件名称	软件用途	软件优点	软件局限性
xyplorer	文件管理器	可以从大部分程度上替代原生文件管理器，有浏览器快捷键和鼠标快捷。	多用户协同性差，数据可视化及分析功能弱
tagLyst	文件资料管理器	给一个文件打上多个标签，来实现对文件的多维管理，并通过标签和关键字对归档的文件进行高效检索。	文档之间关联性差，数据可视化及分析功能弱
语雀	文档与知识管理工具	提供项目文档、学习笔记管理功能，碎片化、结构化知识梳理。	云存储、空间受限、安全性低，数据可视化及分析功能弱
Xmind	思维导图绘制软件	可以绘制思维导图，还能绘制鱼骨图、二维图、树形图、逻辑图、组织结构图	文档管理功能弱
MindMaster	思维导图软件	进行项目管理、知识管理、会议管理、读书笔记内容梳理	文档管理功能弱

发明内容

[0006] 针对现有技术中的问题，本发明提供的基于知识图谱的井场试验数据处理方法及装置，可实现井场试验全流程的数据存储、管理、共享、查询、集中展示等功能，建立数据之间关系，提高数据查询效率，为技术攻关提供支撑。

[0007] 为解决上述技术问题，本发明提供以下技术方案：

[0008] 第一方面，本发明提供一种基于知识图谱的井场试验数据处理方法，包括：

[0009] 对接收的井场试验历史数据、文本文件、音频文件、图片和视频文件等进行格式识别，以生成格式识别结果；

[0010] 根据格式识别结果建立文件的思维导图；

[0011] 根据所述思维导图生成所述井场试验数据的知识图谱；

[0012] 根据所述知识图谱处理井场试验历史数据、文本文件、音频文件、图片和视频文件等以及井场试验新数据、文本文件、音频文件、图片和视频文件等。

[0013] 一实施例中，所述对接收的井场试验历史数据进行格式识别，以生成格式识别结果，包括：

[0014] 接收用户的井场试验数据处理请求；

[0015] 从所述处理请求中提取文件名、操作对象类别以及文件格式；

[0016] 扫描目标目录是否存在与所述文件格式相应的文件夹结构对象，以生成所述格式识别结果。

[0017] 一实施例中，所述根据格式识别结果建立思维导图包括：

[0018] 根据所述格式识别结果确定所述井场试验数据的关键词；

[0019] 根据多个关键词以及预设的井场试验术语词典建立具有多层级关系的数据存储库；

[0020] 根据所述数据存储库建立所述思维导图。

[0021] 一实施例中，所述识别结果包括：结构化数据、半结构化数据以及非结构化数据；

[0022] 所述根据所述格式识别结果确定所述井场试验数据的关键词，包括：

[0023] 对所述结构化数据进行语法分析，以确定所述结构化数据的关键词；

[0024] 对所述半结构化数据以及非结构化数据的标签进行标定，以确定所述半结构化数据以及非结构化数据的关键词。

[0025] 一实施例中，所述对所述结构化数据进行语法分析，以确定所述结构化数据的关键词，包括：

[0026] 根据所述井场试验术语词典对所述结构化数据进行术语抽取；

[0027] 在抽取结果中，选取出现频数大于预设次数的术语；

[0028] 根据所述出现频数大于预设次数的术语生成所述结构化数据的特征向量；

[0029] 根据所述特征向量生成所述结构化数据的关键词。

[0030] 一实施例中，所述对所述半结构化数据以及非结构化数据的标签进行标定，以确定所述半结构化数据以及非结构化数据的关键词，包括：

[0031] 计算所述半结构化数据以及非结构化数据与所述井场试验术语词典的字面文本相似度；

[0032] 根据所述字面文本相似度，从所述半结构化数据以及非结构化数据中选取一部分进行标定。

[0033] 一实施例中，所述根据所述思维导图生成所述井场试验数据的知识图谱包括：

[0034] 根据所述思维导图对所述格式识别结果进行粒度实体识别，以生成识别结果；

[0035] 根据所述识别结果建立所述井场试验数据的知识层级；

[0036] 根据所述识别结果提取所述井场试验数据的实体数据；

[0037] 根据所述知识层级以及实体数据生成所述知识图谱。

[0038] 第二方面，本发明提供一种基于知识图谱的井场试验数据处理装置，包括：

[0039] 识别结果生成模块，用于对接收的井场试验历史数据进行格式识别，以生成格式

识别结果；

[0040] 思维导图建立模块，用于根据格式识别结果建立思维导图；

[0041] 知识图谱生成模块，用于根据所述思维导图生成所述井场试验数据的知识图谱；

[0042] 数据处理模块，用于根据所述知识图谱处理井场试验历史数据以及井场试验新数据。

[0043] 一实施例中，所述识别结果生成模块包括：

[0044] 处理请求接收单元，用于接收用户的井场试验数据处理请求；

[0045] 请求提取单元，用于从所述处理请求中提取文件名、操作对象类别以及文件格式；

[0046] 识别结果生成单元，用于扫描目标目录是否存在与所述文件格式相应的文件夹结构对象，以生成所述格式识别结果。

[0047] 一实施例中，所述思维导图建立模块包括：

[0048] 关键词确定单元，用于根据所述格式识别结果确定所述井场试验数据的关键词；

[0049] 数据存储库建立单元，用于根据多个关键词以及预设的井场试验术语词典建立具有多层级关系的数据存储库；

[0050] 思维导图建立单元，用于根据所述数据存储库建立所述思维导图。

[0051] 一实施例中，所述识别结果包括：结构化数据、半结构化数据以及非结构化数据；

[0052] 所述关键词确定单元包括：

[0053] 数据语法分析单元，用于对所述结构化数据进行语法分析，以确定所述结构化数据的关键词；

[0054] 标签标定单元，用于对所述半结构化数据以及非结构化数据的标签进行标定，以确定所述半结构化数据以及非结构化数据的关键词。

[0055] 一实施例中，所述数据语法分析单元包括：

[0056] 术语抽取单元，用于根据所述井场试验术语词典对所述结构化数据进行术语抽取；

[0057] 术语选取单元，用于在抽取结果中，选取出现频数大于预设次数的术语；

[0058] 特征向量生成单元，用于根据所述出现频数大于预设次数的术语生成所述结构化数据的特征向量；

[0059] 关键词生成单元，用于根据所述特征向量生成所述结构化数据的关键词。

[0060] 一实施例中，所述标签标定单元包括：

[0061] 相似度计算单元，用于计算所述半结构化数据以及非结构化数据与所述井场试验术语词典的字面文本相似度；

[0062] 部分标定单元，用于根据所述字面文本相似度，从所述半结构化数据以及非结构化数据中选取一部分进行标定。

[0063] 一实施例中，所述知识图谱生成模块包括：

[0064] 粒度识别单元，用于根据所述思维导图对所述格式识别结果进行粒度实体识别，以生成识别结果；

[0065] 知识层级建立单元，用于根据所述识别结果建立所述井场试验数据的知识层级；

[0066] 实体数据提取单元，用于根据所述识别结果提取所述井场试验数据的实体数据；

[0067] 知识图谱生成单元，用于根据所述知识层级以及实体数据生成所述知识图谱。

[0068] 第三方面，本发明提供一种电子设备，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，处理器执行程序时实现基于知识图谱的井场试验数据处理方法的步骤。

[0069] 第四方面，本发明提供一种计算机可读存储介质，其上存储有计算机程序，该计算机程序被处理器执行时实现基于知识图谱的井场试验数据处理方法的步骤。

[0070] 从上述描述可知，本发明实施例提供的基于知识图谱的井场试验数据处理方法及装置，首先对接收的井场试验历史数据进行格式识别，以生成格式识别结果；根据格式识别结果建立思维导图；接着，根据思维导图生成井场试验数据的知识图谱；最后根据知识图谱处理井场试验历史数据以及井场试验新数据。本发明提供的基于知识图谱的井场试验数据处理方法及装置，可实现井场试验全流程的数据存储、管理、共享、查询、集中展示等功能，建立数据之间关系，提高数据查询效率，为技术攻关提供支撑。

附图说明

[0071] 为了更清楚地说明本发明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，下面描述中的附图是本发明的一些实施例，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。

[0072] 图1为本发明的实施例中基于知识图谱的井场试验数据处理方法的流程示意图；

[0073] 图2为本发明的实施例中步骤100的流程示意图；

[0074] 图3为本发明的实施例中步骤100的思维导图；

[0075] 图4为本发明的实施例中步骤200的流程示意图；

[0076] 图5为本发明的实施例中步骤201的流程示意图；

[0077] 图6为本发明的实施例中步骤2011的流程示意图；

[0078] 图7为本发明的实施例中步骤2012的流程示意图；

[0079] 图8为本发明的实施例中步骤300的流程示意图；

[0080] 图9为本发明的具体应用实例中基于知识图谱的井场试验数据处理方法的流程示意图；

[0081] 图10为本发明的实施例中基于知识图谱的井场试验数据处理装置的结构框图；

[0082] 图11为本发明的实施例中识别结果生成模块10的结构示意图；

[0083] 图12为本发明的实施例中思维导图建立模块20的结构示意图；

[0084] 图13为本发明的实施例中关键词确定单元201的结构示意图；

[0085] 图14为本发明的实施例中数据语法分析单元2011的结构示意图；

[0086] 图15为本发明的实施例中标签标定单元2012的结构示意图；

[0087] 图16为本发明的实施例中知识图谱生成模块30的结构示意图；

[0088] 图17为本发明的实施例中的电子设备的结构示意图。

具体实施方式

[0089] 为使本发明实施例的目的、技术方案和优点更加清楚，下面将结合本发明实施例中的附图，对本发明实施例中的技术方案进行清楚、完整的描述，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作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。

[0090] 需要说明的是，在不冲突的情况下，本申请中的实施例及实施例中的特征可以相互组合。下面将参考附图并结合实施例来详细说明本申请。

[0091] 本发明的实施例提供一种基于知识图谱的井场试验数据处理方法的具体实施方式，参见图1，该方法具体包括如下内容：

[0092] 步骤100：对接收的井场试验历史数据进行格式识别，以生成格式识别结果。

[0093] 具体地，发送/接收并解析用户通过网络请求发送的文件二进制流及附加的操作命令；从命令参数中提取文件名、操作对象类别以及文件格式的信息；扫描目标目录是否存在与提取的文件格式相应的文件夹结构对象，如果确定不存在与提取的文件目录结构相同的文件夹结构，则创建，并且在创建的文件夹下写入相应非结构文件对象(包括：技术文档、图片/音视频、仪器设备台账、实钻数据等，另外可以理解的是，其数据类型多、文件关系复杂是井场试验文档管理的难点，也是目前现有软件无法解决的。)。

[0094] 步骤200：根据格式识别结果建立思维导图。

[0095] 可以理解的是，思维导图是使用一个中央关键词或想法引起形象化的构造和分类的想法；它用一个中央关键词或想法以辐射线形连接所有的代表字词、想法、任务或其它关联项目的图解方式。

[0096] 具体地，根据数据存储库的文件夹的层级关系生成思维导图，思维导图的节点保留对应文件夹中的所对应的文件；数据存储库中文件夹的标签可表示文件夹的属性，通过统计不同属性文件夹内文件的数目可实现图表等综合展示；可对文件预览并记录笔记，生产报告；导出的文件具有数据存储库中文件夹的层级结构。

[0097] 步骤300：根据所述思维导图生成所述井场试验数据的知识图谱。

[0098] 知识图谱(Knowledge Graph)，是一种知识域可视化或知识领域映射地图，是显示知识发展进程与结构关系的一系列各种不同的图形，用可视化技术描述知识资源及其载体，挖掘、分析、构建、绘制和显示知识及它们之间的相互联系。进一步地，是通过将应用数学、图形学、信息可视化技术、信息科学等学科的理论与方法与计量学引文分析、共现分析等方法结合，并利用可视化的图谱形象地展示学科的核心结构、发展历史、前沿领域以及整体知识架构达到多学科融合目的的现代理论。

[0099] 用户可自定义一套项目库结构，可构建目录结构、文档类别、主题大纲；数据存储库构建完成后可保存为知识图谱模板，可多次引用，尤其在流程化的项目开展中更加方便快捷。例如，在新项目中，可将已有的思维导图形成模板，建立新的思维导图时，可直接应用该模板作为思维导图，或者在该模板的基础上进行调整，进一步地：知识图谱是基于文件层级结构得到的思维导图建立的，同时还可以对导入的每一个文件建立标签，根据标签建立新的思维导图和知识图谱，从而可以根据时间进度、事件进程等建立思维导图，并形成相应的知识图谱。

[0100] 步骤400：根据所述知识图谱处理井场试验历史数据以及井场试验新数据。

[0101] 针对目前井场试验存在数据文档管理复杂，查找有价值文档困难，个人成果记录及试验经验缺乏规范性管理等问题，直接导致知识及有价值的文档无法迅速定位、共享的问题。本发明利用步骤400所生成的知识图谱可实现井场试验全流程的数据存储、管理、共

享、查询、集中展示等功能，建立数据之间关系，提高数据查询效率，为技术攻关提供支撑。

[0102] 从上述描述可知，本发明实施例提供的基于知识图谱的井场试验数据处理方法，首先对接收的井场试验历史数据进行格式识别，以生成格式识别结果；根据格式识别结果建立思维导图；接着，根据思维导图生成井场试验数据的知识图谱；最后根据知识图谱处理井场试验历史数据以及井场试验新数据。本发明提供的基于知识图谱的井场试验数据处理方法及装置，可实现井场试验全流程的数据存储、管理、共享、查询、集中展示等功能，建立数据之间关系，提高数据查询效率，为技术攻关提供支撑。

[0103] 一实施例中，参见图2，步骤100进一步包括：

[0104] 步骤101：接收用户的井场试验数据处理请求；

[0105] 步骤102：从所述处理请求中提取文件名、操作对象类别以及文件格式；

[0106] 步骤103：扫描目标目录是否存在与所述文件格式相应的文件夹结构对象，以生成所述格式识别结果。

[0107] 参见图3，步骤101至步骤103中，如果确定存在与提取的文件目录结构相同的文件夹名称，并且在相应的文件夹下不存在与提取的文件名相同的文件，则在所述相应的文件夹下写入提取的文件对象；如果确定存在与提取的文件名相同的文件，并且所述与提取的文件名相同的文件与提取的操作对象内容相同，则通知用户系统已存在该操作对象；如果确定存在与提取的文件名相同的文件，并且所述与提取的文件名相同的文件与提取的操作对象的内容不同，则按照预定的版本升级规则将与提取的文件名相同的文件重新命名，并且在所述相应的文件夹下写入提取的操作对象。

[0108] 相对地，如果确定存在与提取的文件格式相应的文件夹，并且在所述相应的文件夹下不存在与提取的文件名相同或者内容相同的文件，则在所述相应的文件夹下写入提取的文件操作对象；如果确定存在与提取的文件格式相应的文件夹，并且在所述相应的文件夹下存在与提取的文件名不同但与提取的操作对象内容相同的文件，则可以提示用户选择使用已有文件的文件名还是提取的文件名作为所述操作对象的文件名称；如果确定存在与提取的文件格式相应的文件夹，并且在所述相应的文件夹下存在与提取的文件名相同但与提取的操作对象内容不同的文件，则可以按照预定的版本升级规则将与提取的文件名相同的文件重新命名，并且在所述相应的文件夹下写入提取的操作对象；

[0109] 一实施例中，参见图4，步骤200进一步包括：

[0110] 步骤201：根据所述格式识别结果确定所述井场试验数据的关键词；

[0111] 步骤202：根据多个关键词以及预设的井场试验术语词典建立具有多层级关系的数据存储库；

[0112] 步骤203：根据所述数据存储库建立所述思维导图。

[0113] 在步骤201至步骤203中，根据抽取的数据将文件归类，根据井场文件包含的主要关键词，共设计14张数据库表84个字段，建立了数据存储结构表。根据关键词分类和存储结构的不同，所对应的数据存储部中文件夹类型不同。一个文件所抽取的数据，将其与数据库表终部分表及字段对应，从而存储该文件所对应的标签、类别、基本信息等，并按照关键词匹配到数据存储库对应的文件夹中。这些预设的文件夹存在层级关系，是根据项目的事件维度的逻辑关系设定的文件夹，已赋予了实际含义，即文件夹的名称，多个文件夹以及文件夹名称即形成了数据存储库，最后根据数据存储库中多个节点的层级关系以及节点所代表

的内容生成思维导图。

[0114] 一实施例中，所述识别结果包括：结构化数据、半结构化数据以及非结构化数据；

[0115] 一实施例中，参见图5，步骤201进一步包括：

[0116] 步骤2011：对所述结构化数据进行语法分析，以确定所述结构化数据的关键词；

[0117] 步骤2012：对所述半结构化数据以及非结构化数据的标签进行标定，以确定所述半结构化数据以及非结构化数据的关键词。

[0118] 具体地，步骤2011以及步骤2012在实施时，主要包括以下几个部分，结构化文件数据抽取：语法分析处理部件，接着，进行结构化文件的术语抽取。最后半/非结构化文件数据抽取：由于半\非结构化文件无法通过语法分析来抽取数据，采用添加标签的方式，作为半/非结构化文件的数据。

[0119] 井场试验中大量出现的是半/非结构化文件，优选地，采用2种方式进行数据抽取。(1)人工定义抽取方法，按照井场试验的术语词典2，通过人工的方式添加半/非结构化文件的标签，实现数据的抽提；(2)计算字面文本相似度，自动匹配抽取。提取半/非结构化文件的文件名、建立时间、创建人、创建位置等数据，计算每个数据同井场试验的术语词典2中术语的字面文本相似度，

[0120] 一实施例中，参见图6，步骤2011进一步包括：

[0121] 步骤20111：根据所述井场试验术语词典对所述结构化数据进行术语抽取；

[0122] 步骤20112：在抽取结果中，选取出现频数大于预设次数的术语；

[0123] 在步骤20111以及步骤20122中，根据用户机提供的结构化文件储存请求，对存储对象的结构化文件进行语法分析，实现数据的抽取。首先是建立一个井场试验的术语辞典。基于专业术语库、国家标准、行业规范等建立井场试验的术语词典1，然后对10-20次的井场试验进行统计和归纳，根据高频词汇出现的频数排序，截取频数范围超过预设阈值的的词汇(优选地，该预设阈值为2-4次)，形成井场试验的术语词典2。

[0124] 步骤20123：根据所述出现频数大于预设次数的术语生成所述结构化数据的特征向量；

[0125] 以结构化文件为统计样本，统计出井场试验的术语词典2中术语出现的频次，并进行列表，得到该结构化文件的术语抽取特征向量。从而实现了结构化文件的数据抽取工作。

[0126] 步骤20124：根据所述特征向量生成所述结构化数据的关键词。

[0127] 一实施例中，参见图7，步骤2012进一步包括：

[0128] 步骤20121：计算所述半结构化数据以及非结构化数据与所述井场试验术语词典的字面文本相似度；

[0129] 步骤20122：根据所述字面文本相似度，从所述半结构化数据以及非结构化数据中选取一部分进行标定。

[0130] 参见公式1，选定字面文本相似度最高的作为本数据所对应的标签，从而实现半/非结构化文件数据抽取。

[0131] $$sim=60\times\left(\frac{xsword}{ctrlword}+\frac{xsword}{keyword}\right)\Big/2+40\times dp\times\left(\sum\frac{c\_xsword\ (i)}{\sum ctrlword\ (i)}+\sum\frac{k\_xsword\ (i)}{\sum keyword\ (i)}\right)\Big/2 \quad (1)$$

[0132] 其中xsword代表两个词汇拥有相同的字的个数；ctrlword代表被匹配词A所含有的字的个数；keyword代表待匹配词B所含有的字的总个数；dp代表位置系数，表示被匹配词

A与待匹配词B的总字数的比值；$\sum \frac{c_xsword\ (i)}{\sum ctrlword\ (i)}$代表两个词A与B拥有的相同的字在A中所处位置的权重之和；$\sum \frac{k_xsword\ (i)}{\sum keyword\ (i)}$代表两个词A与B拥有的相同的字在B中所处位置的权重之和。

[0133] 一实施例中，参见图8，步骤300进一步包括：

[0134] 步骤301：根据所述思维导图对所述格式识别结果进行粒度实体识别，以生成识别结果；

[0135] 优选地，粒度实体识别主要分为三个阶段，首先开展井场试验数据解析，分解文本档案、音视频档案、档案元数据、XML数据等资源中各类异构数据格式。在数据解析的基础上，进行井场试验数据知识层面的实体描述，通过井场试验数据元数据架构设计以及井场试验数据知识层级构建来共同揭示井场试验数据资源。最后参考科技类相关字典(也可以用术语词典2)，结合井场试验数据实体的词性特征等要素，建立基于井场试验数据的实体抽取规则，通过深度学习模型等完成井场试验数据实体抽取。通过对井场试验数据数据解析、深层次揭示以及实体抽取，完成井场试验数据细粒度的实体识别，为井场试验数据语义关联研究提供数据支持。

[0136] 步骤302：根据所述识别结果建立所述井场试验数据的知识层级；

[0137] 井场试验数据资源除井场试验数据资源元数据外，还包括井场试验数据资源自身的知识，故井场试验数据资源粒度加工应该能够识别到内容层级，即井场试验数据资源知识层面的研究任务、研究思路以及施工方案、考核指标等实体，在井场试验数据知识层面的数据揭示过程中，需要对井场试验数据语料定义句的语法-语义进行剖析，借助语义技术进行实体识别，提取井场试验数据知识层面的高频词汇以及关键词。

[0138] 步骤303：根据所述识别结果提取所述井场试验数据的实体数据；

[0139] 为开展语义关联的井场试验数据管理研究，在井场试验数据数据解析与深层次揭示的基础上，须抽取井场试验数据实体，以作为最小的井场试验数据知识单元与其他知识单元建立关联关系。因此，在井场试验数据细粒度实体识别的最后阶段，可利用命名实体识别、自然语言处理等技术完成井场试验数据实体的抽取。在井场试验数据资源分类以及属性定义的基础上，通过数据解析从井场试验数据资源中辨别和析出深层次揭示的实体的实例数据。根据井场试验数据资源的数据结构和特点，为了提高井场试验数据资源实体抽取的性能，可引入科技类相关字典，结合词性特征等要素，建立基于井场试验数据的实体抽取规则。目前常用的知识抽取模型有CRF模型、BiLSTM模型等(修晓蕾，2019)。结合井场试验数据语义词典，基于命名实体识别等技术可获得井场试验数据资源语义层面的关键词或高频词，如研究任务C Research Mission、实施方案(Ttnplementation Plan、考核指标(Target)、经费预算(Budget)等实体。由于数据资源的不同，在实体抽取时获得的实体也会有变化，这4个实体仅为井场试验数据资源实体抽取中的通用实体，在具体到某个井场试验数据的抽取时，需要根据数据特点进一步细化实体。在实体抽取后，还可利用语义理解、机构知识库等知识库中的名词解释等进行智能校对井场试验数据实体，再由人工审核入知识库，将错误率降到最低。

[0140] 步骤304:根据所述知识层级以及实体数据生成所述知识图谱。

[0141] 首先需要选取合适的构建知识图谱的语言,在本发明中采用OWL Lite语言构建知识图谱,接着,利用OWL Lite语言以及步骤304中的实体数据,在知识层级的整体框架下,构建知识图谱,另一方面,构建好的知识图谱支持文件、项目、项目库的分享。导出的文件和项目会以数据存储库中设定的文件夹的层级结构导出。在浏览文档知识库时,可在文档(WORD、EXCEL、PPT等)中选择抓取关键信息,形成笔记记录,并通过整理将其生成结果性文件如实验报告、数据报表等。

[0142] 为进一步地说明本方案,本发明还提供一种基于知识图谱的井场试验数据处理方法的具体应用实例,具体包括如下内容,参见图9。

[0143] S1:建立异构融合数据库。

[0144] 具体地,对文件的格式识别,将文件划分为结构化文件和半/非结构化文件,接着,进行数据抽取,根据用户机提供的结构化文件储存请求,对存储对象的结构化文件进行语法分析,实现数据的抽取。

[0145] S2:基于异构融合数据库建立井场试验数据的知识图谱。

[0146] 可以理解的是,在步骤S2中,需要实现全流程的结构化、非结构化、半结构化试验资料数据的上传,通过项目库形式进行统一管理,可单次/批量井场试验数据入库模式,在上传文档前或之后创建项目库。具体地:第一步构建项目库基本信息,包括项目名称、周期、简介等;第二步设置项目默认导图维度分类及名称;第三步为上传的单个文件或批量文件添加标签,以当前结构生成思维导图或构建新的思维导图;文档阅读标记管理:实现文档知识点标记、批注等保存及入库;文档数据权限设置:仅自己可见、部分用户可见、共享所有用户等;数据标签管理及输出:针对已上传试验数据,实现数据管理,包括数据标签增加、编辑、删除,数据关系编辑,数据材料编辑及删除,数据输出等功能。

[0147] S3:根据知识图谱对井场试验数据进行可视化展示。

[0148] 根据预设数据存储库的文件夹的层级关系,可自动生成相同框架的思维导图,以思维导图形式对项目可视化。思维导图的每个节点处保留对应文件夹中的所对应的文件,节点名称就是以文件夹的名称命名。文件夹的不同的标签所反映的是文件夹内文件的不同的属性。对不同属性的文件数量进行统计。采用类Office UI样式,图表等组合形式进行井场试验规划、进度、关键指标、有形化成果等数据库资料的可视化。采用类似RDF的文档资源模型,可快速根据用户输入检索到精确结果;按照修改日期、用户级别标记、类型等检索可见知识库;支持知识库标题、关键字、创建时间、知识所有人、标签等不同类别的查询检索。

[0149] 从上述描述可知,本发明实施例提供的基于知识图谱的井场试验数据处理方法,对文件的格式识别,将文件划分为结构化文件和半/非结构化文件;通过对结构化文件进行语法分析和对半/非结构化文件的标签标定,得到文件的关键词;根据井场文件包含的主要关键词,建立具有层级关系的多个文件夹作为数据存储库,并将对应的关键词作为文件夹的名称;将文件的关键词与文件夹的名称匹配,将文件放入其中;可对数据存储库中的文件夹及文件进行标签标定、增加、删除等管理功能;

[0150] 基于同一发明构思,本申请实施例还提供了基于知识图谱的井场试验数据处理装置,可以用于实现上述实施例所描述的方法,如下面的实施例。由于基于知识图谱的井场试验数据处理装置解决问题的原理与基于知识图谱的井场试验数据处理方法相似,因此基于

知识图谱的井场试验数据处理装置的实施可以参见基于知识图谱的井场试验数据处理方法实施，重复之处不再赘述。以下所使用的，术语“单元”或者“模块”可以实现预定功能的软件和/或硬件的组合。尽管以下实施例所描述的系统较佳地以软件来实现，但是硬件，或者软件和硬件的组合的实现也是可能并被构想的。

[0151] 本发明的实施例提供一种能够实现基于知识图谱的井场试验数据处理方法的基于知识图谱的井场试验数据处理装置的具体实施方式，参见图10，基于知识图谱的井场试验数据处理装置具体包括如下内容：

[0152] 识别结果生成模块10，用于对接收的井场试验历史数据进行格式识别，以生成格式识别结果；

[0153] 思维导图建立模块20，用于根据格式识别结果建立思维导图；

[0154] 知识图谱生成模块30，用于根据所述思维导图生成所述井场试验数据的知识图谱；

[0155] 数据处理模块40，用于根据所述知识图谱处理井场试验历史数据以及井场试验新数据。

[0156] 一实施例中，参见图11，所述识别结果生成模块10包括：

[0157] 处理请求接收单元101，用于接收用户的井场试验数据处理请求；

[0158] 请求提取单元102，用于从所述处理请求中提取文件名、操作对象类别以及文件格式；

[0159] 识别结果生成单元103，用于扫描目标目录是否存在与所述文件格式相应的文件夹结构对象，以生成所述格式识别结果。

[0160] 一实施例中，参见图12，所述思维导图建立模块20包括：

[0161] 关键词确定单元201，用于根据所述格式识别结果确定所述井场试验数据的关键词；

[0162] 数据存储库建立单元202，用于根据多个关键词以及预设的井场试验术语词典建立具有多层级关系的数据存储库；

[0163] 思维导图建立单元203，用于根据所述数据存储库建立所述思维导图。

[0164] 一实施例中，所述识别结果包括：结构化数据、半结构化数据以及非结构化数据；

[0165] 一实施例中，参见图13，所述关键词确定单元201包括：

[0166] 数据语法分析单元2011，用于对所述结构化数据进行语法分析，以确定所述结构化数据的关键词；

[0167] 标签标定单元2012，用于对所述半结构化数据以及非结构化数据的标签进行标定，以确定所述半结构化数据以及非结构化数据的关键词。

[0168] 一实施例中，参见图14，所述数据语法分析单元2011包括：

[0169] 术语抽取单元20111，用于根据所述井场试验术语词典对所述结构化数据进行术语抽取；

[0170] 术语选取单元20112，用于在抽取结果中，选取出现频数大于预设次数的术语；

[0171] 特征向量生成单元20113，用于根据所述出现频数大于预设次数的术语生成所述结构化数据的特征向量；

[0172] 关键词生成单元20114，用于根据所述特征向量生成所述结构化数据的关键词。

[0173] 一实施例中，参见图15，所述标签标定单元2012包括：

[0174] 相似度计算单元20121，用于计算所述半结构化数据以及非结构化数据与所述井场试验术语词典的字面文本相似度；

[0175] 部分标定单元20122，用于根据所述字面文本相似度，从所述半结构化数据以及非结构化数据中选取一部分进行标定。

[0176] 一实施例中，参见图16，所述知识图谱生成模块30包括：

[0177] 粒度识别单元301，用于根据所述思维导图对所述格式识别结果进行粒度实体识别，以生成识别结果；

[0178] 知识层级建立单元302，用于根据所述识别结果建立所述井场试验数据的知识层级；

[0179] 实体数据提取单元303，用于根据所述识别结果提取所述井场试验数据的实体数据；

[0180] 知识图谱生成单元304，用于根据所述知识层级以及实体数据生成所述知识图谱。

[0181] 从上述描述可知，本发明实施例提供的基于知识图谱的井场试验数据处理装置，首先对接收的井场试验历史数据进行格式识别，以生成格式识别结果；根据格式识别结果建立思维导图；接着，根据思维导图生成井场试验数据的知识图谱；最后根据知识图谱处理井场试验历史数据以及井场试验新数据。本发明提供的基于知识图谱的井场试验数据处理方法及装置，可实现井场试验全流程的数据存储、管理、共享、查询、集中展示等功能，建立数据之间关系，提高数据查询效率，为技术攻关提供支撑。

[0182] 上述实施例阐明的装置、模块或单元，具体可以由计算机芯片或实体实现，或者由具有某种功能的产品来实现。一种典型的实现设备为电子设备，具体的，电子设备例如可以为个人计算机、膝上型计算机、蜂窝电话、相机电话、智能电话、个人数字助理、媒体播放器、导航设备、电子邮件设备、游戏控制台、平板计算机、可穿戴设备或者这些设备中的任何设备的组合。

[0183] 在一个典型的实例中电子设备具体包括存储器、处理器以及存储在存储器上并可在处理器上运行的计算机程序，该处理器执行该程序时实现上述基于知识图谱的井场试验数据处理方法的步骤，该步骤包括：

[0184] 步骤100：对接收的井场试验历史数据进行格式识别，以生成格式识别结果；

[0185] 步骤200：根据格式识别结果建立思维导图；

[0186] 步骤300：根据所述思维导图生成所述井场试验数据的知识图谱；

[0187] 步骤400：根据所述知识图谱处理井场试验历史数据以及井场试验新数据。

[0188] 下面参考图17，其示出了适于用来实现本申请实施例的电子设备600的结构示意图。

[0189] 如图17所示，电子设备600包括中央处理单元(CPU)601，其可以根据存储在只读存储器(ROM)602中的程序或者从存储部分608加载到随机访问存储器(RAM))603中的程序而执行各种适当的工作和处理。在RAM603中，还存储有系统600操作所需的各种程序和数据。CPU601、ROM602、以及RAM603通过总线604彼此相连。输入/输出(I/O)接口605也连接至总线604。

[0190] 以下部件连接至I/O接口605：包括键盘、鼠标等的输入部分606；包括诸如阴极射

线管(CRT)、液晶显示器(LCD)等以及扬声器等的输出部分607;包括硬盘等的存储部分608;以及包括诸如LAN卡,调制解调器等的网络接口卡的通信部分609。通信部分609经由诸如因特网的网络执行通信处理。驱动器610也根据需要连接至I/O接口605。可拆卸介质611,诸如磁盘、光盘、磁光盘、半导体存储器等等,根据需要安装在驱动器610上,以便于从其上读出的计算机程序根据需要被安装如存储部分608。

[0191] 特别地,根据本发明的实施例,上文参考流程图描述的过程可以被实现为计算机软件程序。例如,本发明的实施例包括一种计算机可读存储介质,其上存储有计算机程序,该计算机程序被处理器执行时实现上述基于知识图谱的井场试验数据处理方法的步骤,该步骤包括:

[0192] 步骤100:对接收的井场试验历史数据进行格式识别,以生成格式识别结果;

[0193] 步骤200:根据格式识别结果建立思维导图;

[0194] 步骤300:根据所述思维导图生成所述井场试验数据的知识图谱;

[0195] 步骤400:根据所述知识图谱处理井场试验历史数据以及井场试验新数据。

[0196] 在这样的实施例中,该计算机程序可以通过通信部分609从网络上被下载和安装,和/或从可拆卸介质611被安装。

[0197] 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。按照本文中的界定,计算机可读介质不包括暂存电脑可读媒体(transitory media),如调制的数据信号和载波。

[0198] 为了描述的方便,描述以上装置时以功能分为各种单元分别描述。当然,在实施本申请时可以把各单元的功能在同一个或多个软件和/或硬件中实现。

[0199] 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。

[0200] 这些计算机程序指令也可存储在能引导计算机或其他可编程数据处理设备以特定方式工作的计算机可读存储器中,使得存储在该计算机可读存储器中的指令产生包括指令装置的制造品,该指令装置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。

[0201] 这些计算机程序指令也可装载到计算机或其他可编程数据处理设备上,使得在计算机或其他可编程设备上执行一系列操作步骤以产生计算机实现的处理,从而在计算机或其他可编程设备上执行的指令提供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步骤。

[0202] 还需要说明的是，术语“包括”、“包含”或者其任何其他变体意在涵盖非排他性的包含，从而使得包括一系列要素的过程、方法、商品或者设备不仅包括那些要素，而且还包括没有明确列出的其他要素，或者是还包括为这种过程、方法、商品或者设备所固有的要素。在没有更多限制的情况下，由语句“包括一个……”限定的要素，并不排除在包括该要素的过程、方法、商品或者设备中还存在另外的相同要素。

[0203] 本领域技术人员应明白，本申请的实施例可提供为方法、系统或计算机程序产品。因此，本申请可采用完全硬件实施例、完全软件实施例或结合软件和硬件方面的实施例的形式。而且，本申请可采用在一个或多个其中包含有计算机可用程序代码的计算机可用存储介质(包括但不限于磁盘存储器、CD-ROM、光学存储器等)上实施的计算机程序产品的形式。

[0204] 本申请可以在由计算机执行的计算机可执行指令的一般上下文中描述，例如程序模块。一般地，程序模块包括执行特定任务或实现特定抽象数据类型的例程、程序、对象、组件、数据结构等等。也可以在分布式计算环境中实践本申请，在这些分布式计算环境中，由通过通信网络而被连接的远程处理设备来执行任务。在分布式计算环境中，程序模块可以位于包括存储设备在内的本地和远程计算机存储介质中。

[0205] 本说明书中的各个实施例均采用递进的方式描述，各个实施例之间相同相似的部分互相参见即可，每个实施例重点说明的都是与其他实施例的不同之处。尤其，对于系统实施例而言，由于其基本相似于方法实施例，所以描述的比较简单，相关之处参见方法实施例的部分说明即可。

[0206] 以上该仅为本申请的实施例而已，并不用于限制本申请。对于本领域技术人员来说，本申请可以有各种更改和变化。凡在本申请的精神和原理之内所作的任何修改、等同替换、改进等，均应包含在本申请的权利要求范围之内。

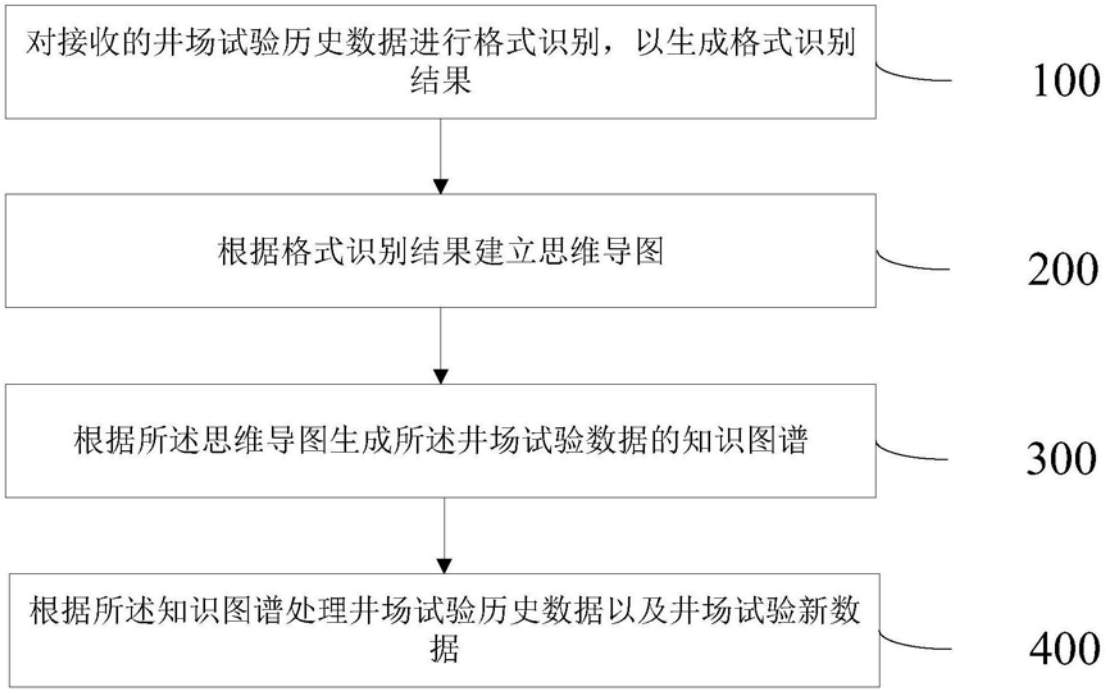

对接收的井场试验历史数据进行格式识别，以生成格式识别结果
100
根据格式识别结果建立思维导图
200
根据所述思维导图生成所述井场试验数据的知识图谱
300
根据所述知识图谱处理井场试验历史数据以及井场试验新数据
400

图1

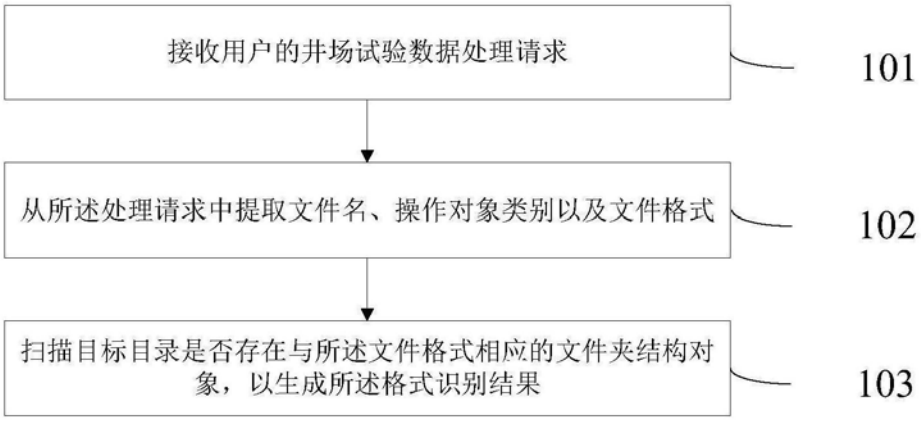

接收用户的井场试验数据处理请求
101
从所述处理请求中提取文件名、操作对象类别以及文件格式
102
扫描目标目录是否存在与所述文件格式相应的文件夹结构对象，以生成所述格式识别结果
103

图2

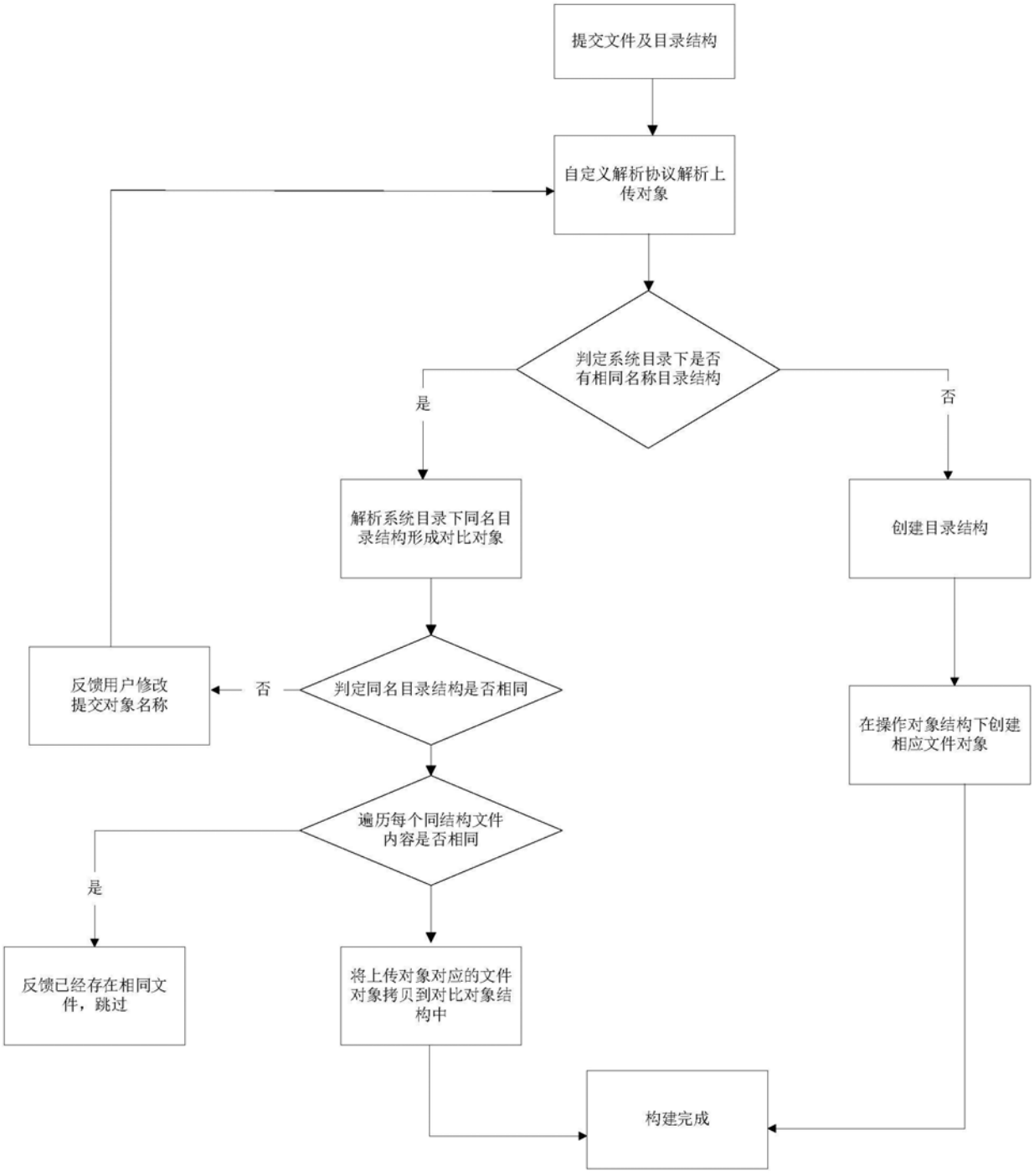

提交文件及目录结构
自定义解析协议解析上传对象
判定系统目录下是否有相同名称目录结构
是
否
解析系统目录下同名目录结构形成对比对象
创建目录结构
判定同名目录结构是否相同
否
反馈用户修改提交对象名称
在操作对象结构下创建相应文件对象
遍历每个同结构文件内容是否相同
是
反馈已经存在相同文件，跳过
将上传对象对应的文件对象拷贝到对比对象结构中
构建完成

图3

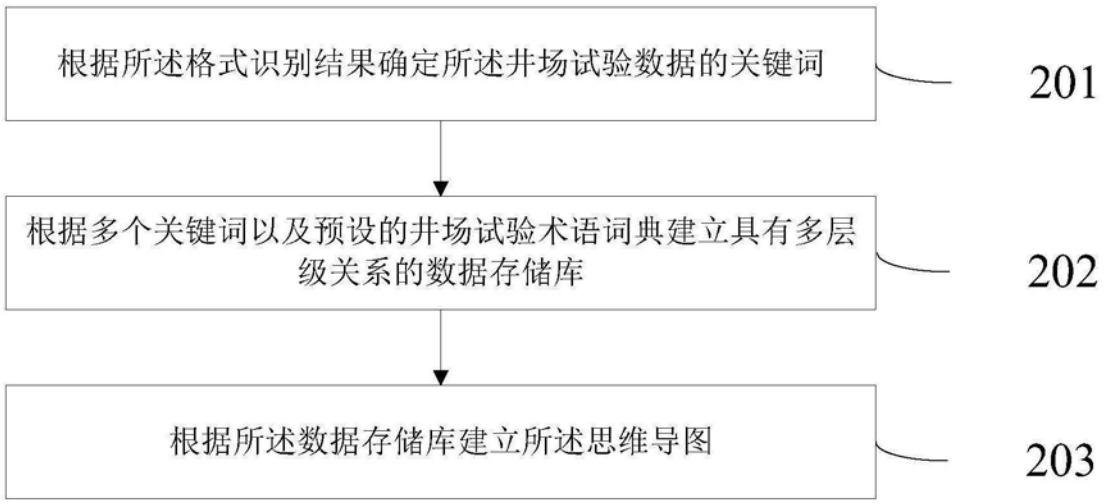

根据所述格式识别结果确定所述井场试验数据的关键词
201
根据多个关键词以及预设的井场试验术语词典建立具有多层级关系的数据存储库
202
根据所述数据存储库建立所述思维导图
203

图4

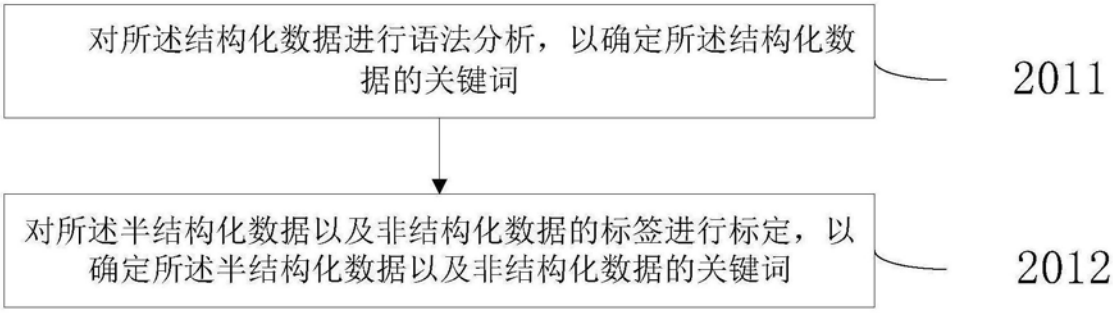

对所述结构化数据进行语法分析，以确定所述结构化数据的关键词
2011
对所述半结构化数据以及非结构化数据的标签进行标定，以确定所述半结构化数据以及非结构化数据的关键词
2012

图5

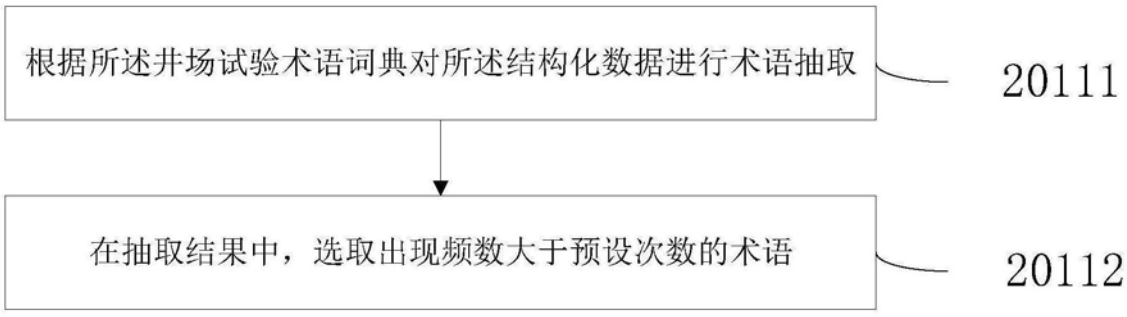

根据所述井场试验术语词典对所述结构化数据进行术语抽取
20111
在抽取结果中，选取出现频数大于预设次数的术语
20112

图6

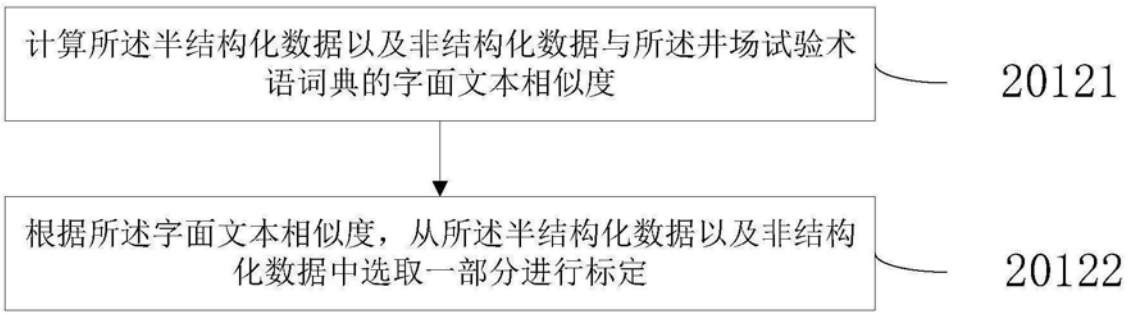

计算所述半结构化数据以及非结构化数据与所述井场试验术语词典的字面文本相似度
20121
根据所述字面文本相似度，从所述半结构化数据以及非结构化数据中选取一部分进行标定
20122

图7

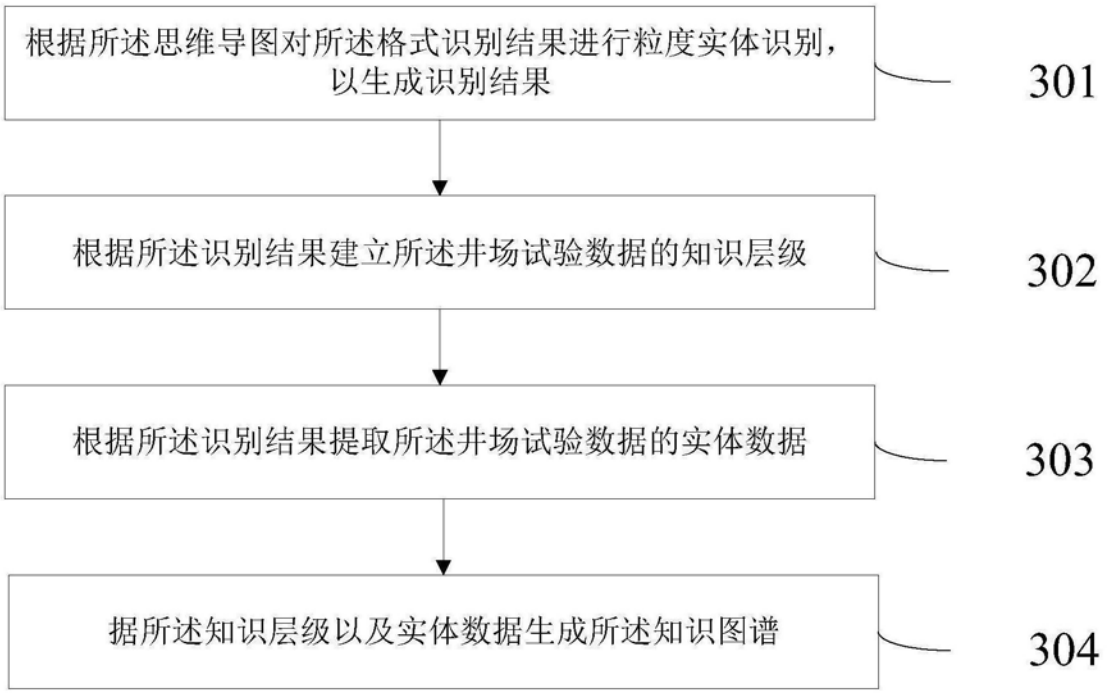

根据所述思维导图对所述格式识别结果进行粒度实体识别，以生成识别结果
301
根据所述识别结果建立所述井场试验数据的知识层级
302
根据所述识别结果提取所述井场试验数据的实体数据
303
据所述知识层级以及实体数据生成所述知识图谱
304

图8

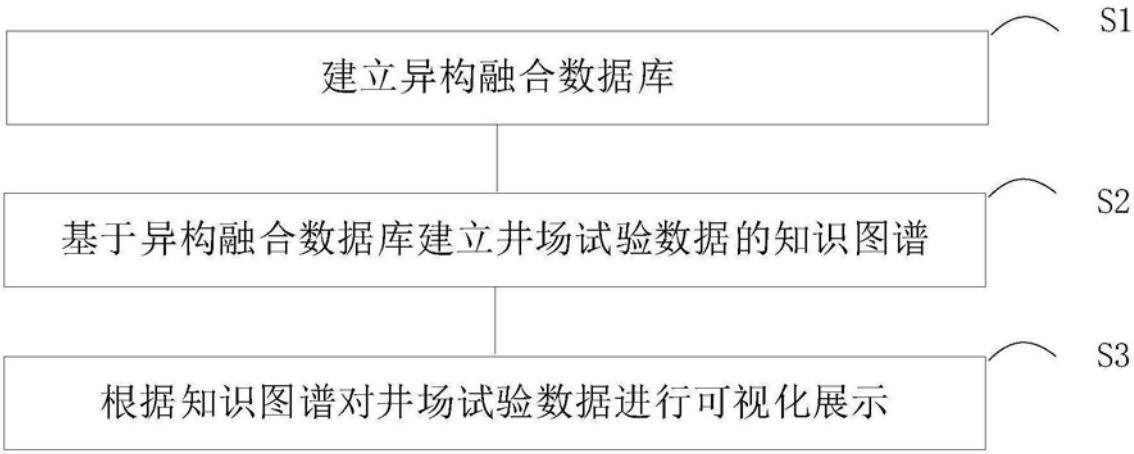

S1
建立异构融合数据库
S2
基于异构融合数据库建立井场试验数据的知识图谱
S3
根据知识图谱对井场试验数据进行可视化展示

图9

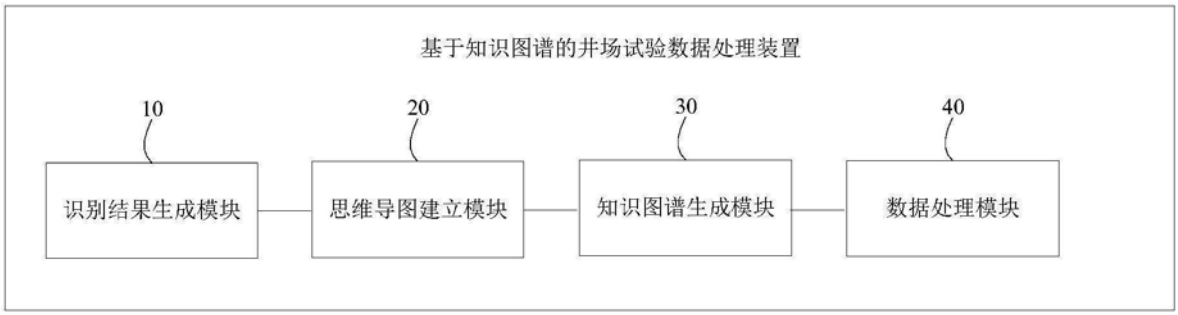

基于知识图谱的井场试验数据处理装置
10
识别结果生成模块
20
思维导图建立模块
30
知识图谱生成模块
40
数据处理模块

图10

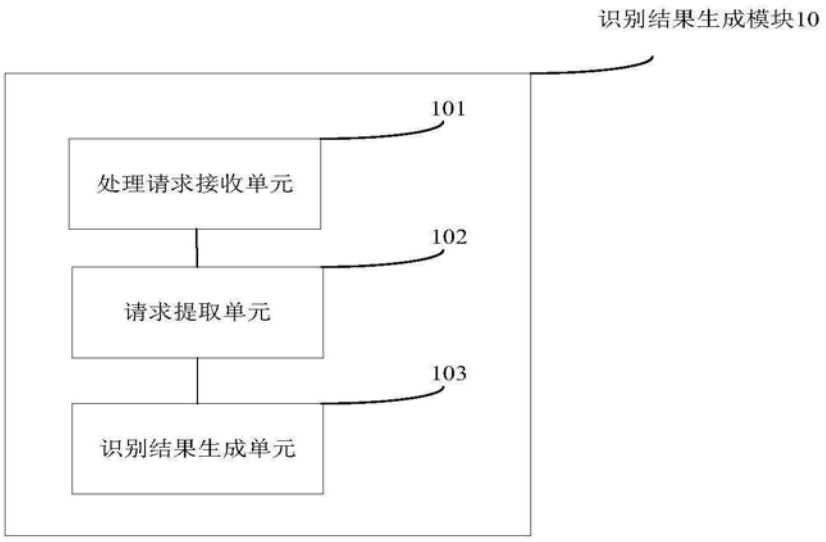

识别结果生成模块10
101
处理请求接收单元
102
请求提取单元
103
识别结果生成单元

图11

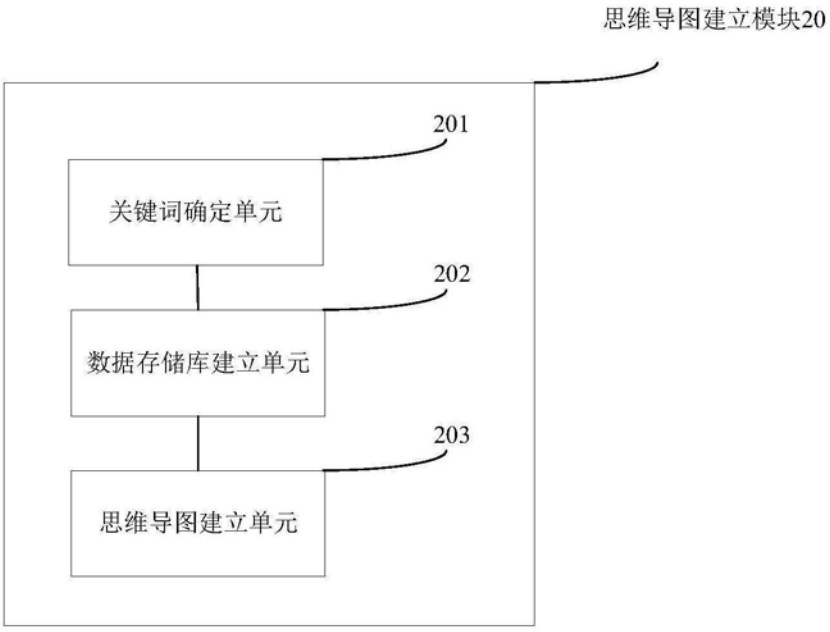

思维导图建立模块20
201
关键词确定单元
202
数据存储库建立单元
203
思维导图建立单元

图12

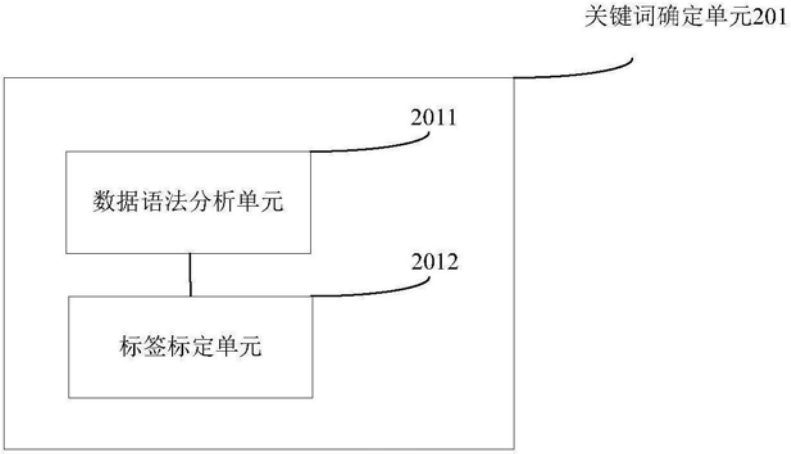

关键词确定单元201
2011
数据语法分析单元
2012
标签标定单元

图13

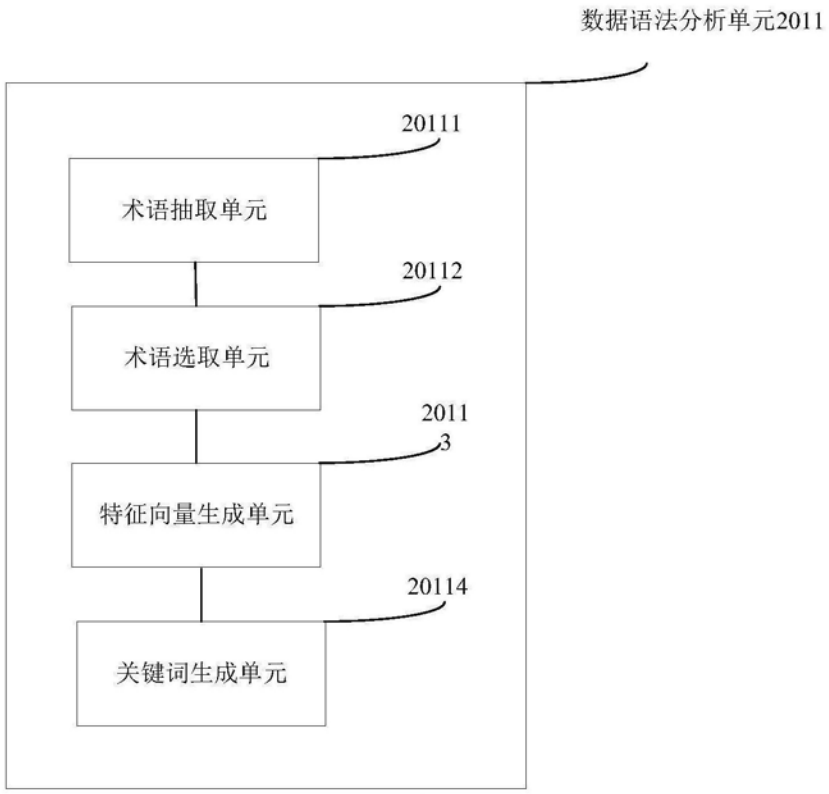

数据语法分析单元2011
20111
术语抽取单元
20112
术语选取单元
2011
3
特征向量生成单元
20114
关键词生成单元

图14

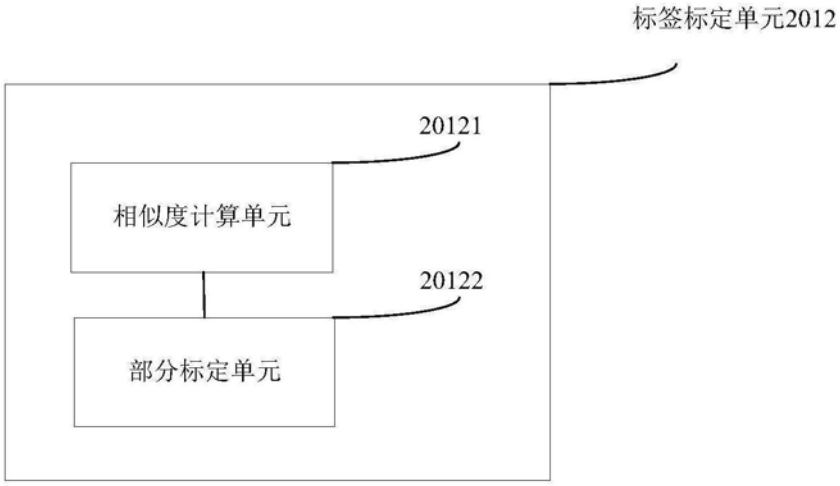

标签标定单元2012
20121
相似度计算单元
20122
部分标定单元

图15

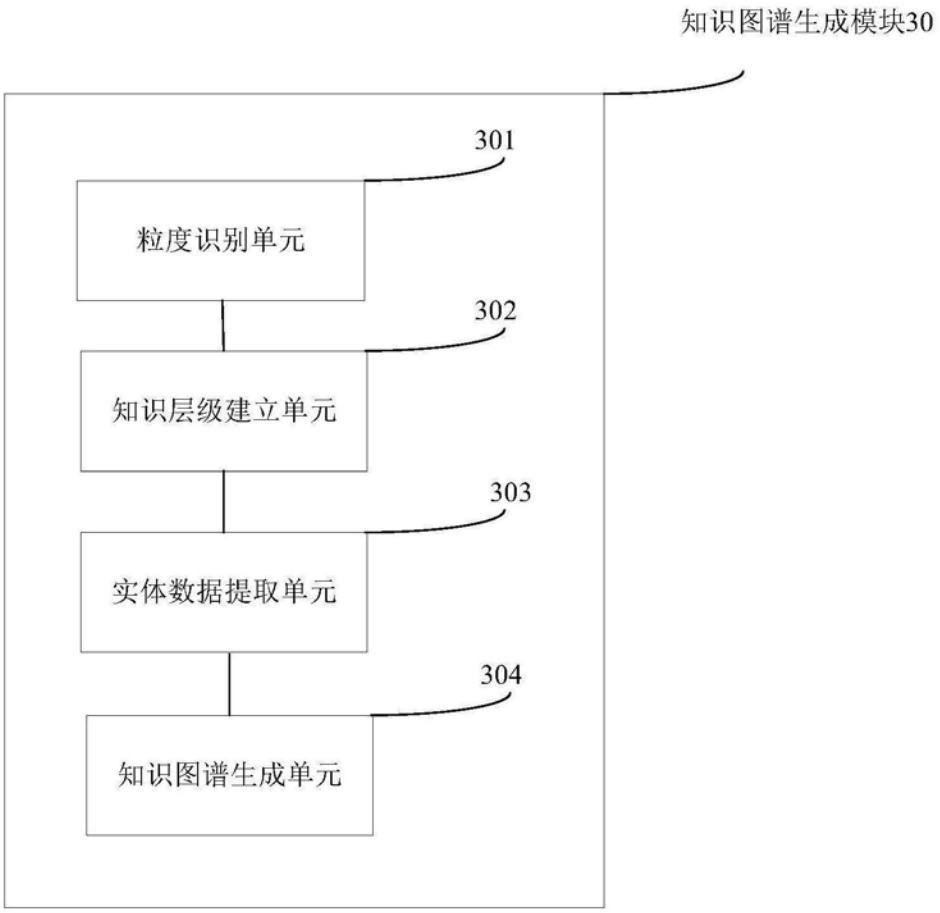

知识图谱生成模块30
301
粒度识别单元
302
知识层级建立单元
303
实体数据提取单元
304
知识图谱生成单元

图16

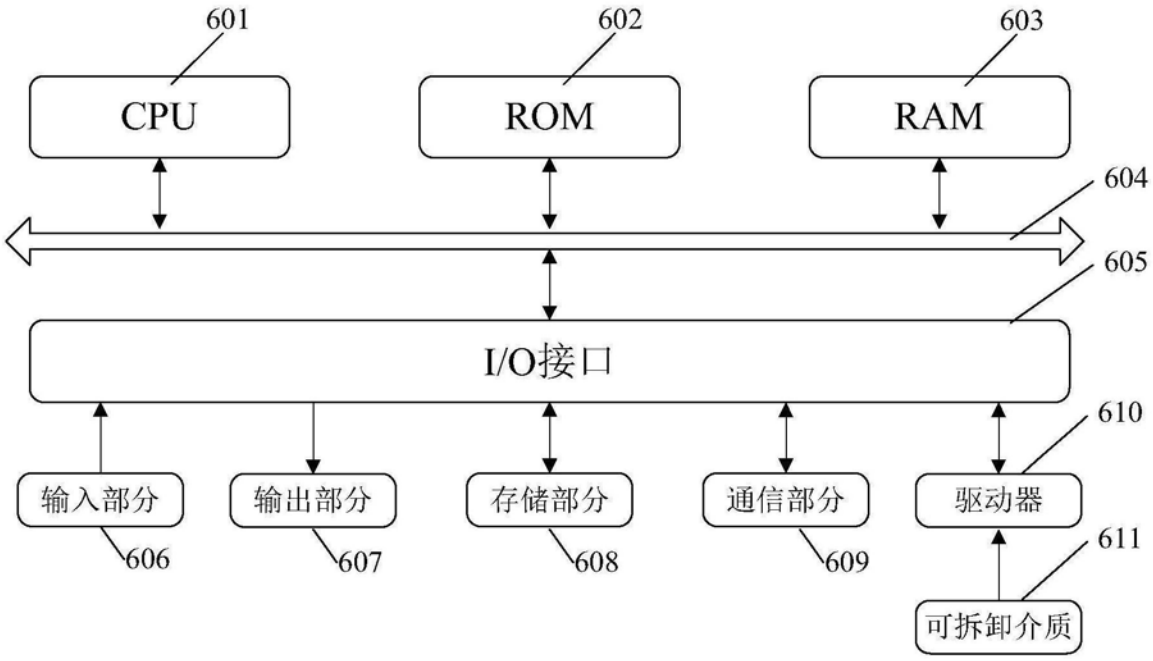

601
602
603
CPU
ROM
RAM
604
605
I/O接口
610
输入部分
输出部分
存储部分
通信部分
驱动器
611
606
607
608
609
可拆卸介质

图17