(19)国家知识产权局

(12)发明专利申请

(10)申请公布号 CN 115115227 A
(43)申请公布日 2022.09.27

(21)申请号 202210753820.2

(22)申请日 2022.06.28

(71)申请人 华南理工大学
地址 510640 广东省广州市天河区五山路381号

(72)发明人 满奕 李继庚 张欢欢 洪蒙纳

(74)专利代理机构 成都方圆聿联专利代理事务所(普通合伙) 51241
专利代理师 苟铭

(51)Int.Cl.
G06Q 10/06(2012.01)
G06F 16/35(2019.01)
G06F 16/36(2019.01)
G06Q 50/04(2012.01)

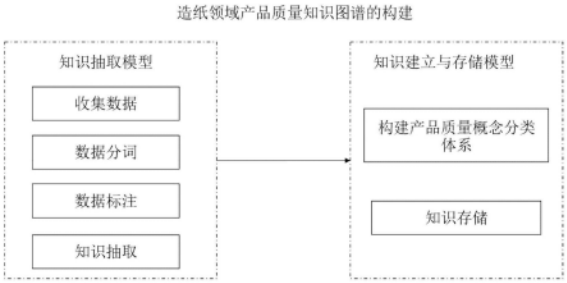
权利要求书2页 说明书4页 附图2页

(54)发明名称

一种用于造纸领域的产品质量知识图谱构建方法

(57)摘要

本发明公开了一种用于造纸领域的产品质量知识图谱构建方法,基于已有造纸领域结构化数据和互联网数据生成产品质量的相关数据,获取的数据通过收集、筛选、分析、汇总形成造纸领域产品质量的基础数据;根据采集的数据信息,进行分词处理后形成造纸领域产品质量语料库;在语料库中选取部分数据为训练集,通过人工方式进行标注作为训练数据;利用标注好的训练数据来迭代训练命名实体识别模型来实现知识的抽取;本发明通过对相关书籍、网页、论坛等信息获取,得到综合造纸领域的产品质量相关数据资料,构建基于造纸领域的产品质量知识分类体系,通过图数据库的形式进行存储;本发明提供的技术方案还可以从造纸行业泛化到其他行业。



CN 115115227 A

1.一种用于造纸领域的产品质量知识图谱构建方法,其特征在于,包括如下步骤:

步骤(1)、收集数据:

基于已有造纸行业结构化数据、互联网数据以及书籍数据生成造纸行业;获取的数据通过收集、筛选、分析、汇总后形成造纸领域产品质量的基础数据;

步骤(2)、数据分词:

根据上述步骤(1)采集的数据信息,运用分词模型进行分词处理,最后形成造纸领域产品质量语料库;

步骤(3)、数据标注:

根据上述步骤(2)中的语料库,选取部分数据以汉语每个字为识别单位进行人工标注,然后将标注好的数据作为训练集;

在人工标注过程中,标注的分类包含故障类型、故障名、故障设备名称、故障描述、故障原因以及故障解决办法;

步骤(4)、知识抽取

根据步骤(3)中的训练集建立命名实体识别模型并进行模型训练,用训练好的模型对所有文档进行知识的抽取;

步骤(5)、构建产品质量知识图谱分类体系

运用自顶而下的方式,采用人工构建的方式来构建造纸领域产品质量概念和关系的分类体系;

步骤(6)、知识存储

根据构建的造纸领域产品质量概念和关系分类体系,将抽取出来的知识分类对应存储于Neo4j图数据库中。

2.根据权利要求1所述的一种用于造纸领域的产品质量知识图谱构建方法,其特征在于,步骤(1)中所述的相关数据,包括已有相关设备和工艺的结构化数据,以及通过爬虫在相关造纸企业网站、造纸故障类网站、造纸产品质量问题相关网站采集的造纸产品质量问题相关文档信息;这些造纸产品质量相关文档信息包括造纸产品质量标准类文档信息、政策标准、专利、报告、百科。

3.根据权利要求1中所述的一种用于造纸领域的产品质量知识图谱构建方法,其特征在于,在步骤(5)包括以下子步骤:

5.1、定义造纸产品质量问题的知识分类体系,设计了6类造纸故障概念,分别是产品质量问题、产生原因、现象、解决办法、部位、检测;

5.2、根据上述步骤5.1中6种已定义的造纸过程故障概念,将泛化的概念共现关系按照语义类型进一步细分为7大类的概念关系以及14小类的概念关系。

4.根据权利要求3中所述的一种用于造纸领域的产品质量知识图谱构建方法,其特征在于,7大类的概念关系包括:故障诊断、故障表现、作用部位、质量检测、故障部位、检测结果、诊断依据;

14小类的概念关系包括:“产品质量问题”与“现象”、“现象”与“产生原因”、“故障决策”与“现象”之间的关系类型定义为“故障表现”;“产品质量问题”与“发生部位”、“部位”与“现象”之间的关系类型定义为“故障部位”;“产品质量问题”与“检测”、“检测”与“现象”之间的关系类型定义为“质量检测”;“产品质量问题”与“产生原因”、“产生原因”与“解决办法”、

“解决办法”与“产品质量问题”之间的关系类型定义为“故障诊断”；“检测”与“部位”、“部位”与“解决办法”之间的关系类型定义为“作用部位”；“检测”与“现象”之间的关系类型定义为“检测结果”；“现象”与“产品质量问题”之间的关系类型定义为“诊断依据”。

一种用于造纸领域的产品质量知识图谱构建方法

技术领域

[0001] 本发明涉及知识图谱构建技术领域，尤其涉及一种用于造纸领域的产品质量知识图谱构建方法。

背景技术

[0002] 知识图谱是以概念、实体为中心，表达概念与概念之间，实体与实体之间的关系。知识图谱能够表达关系复杂的知识，从知识图谱中可以清楚的看到影响因素的关联关系、变量链接的传播路径、数据的层次性。同时通过知识图谱状态的变化能够发现故障时延。知识图谱的这些特性正好能够解决复杂工业过程(如造纸)关系难以表达的难题。

[0003] 现有技术的缺陷和不足：目前，知识图谱在造纸行业的应用还是一片空白。造纸生产过程规模庞大、结构复杂、生产单元之间的耦合性极强，使得造纸生产过程故障诊断的难度越来越大。由于造纸工业过程无法建立精确的数学模型，所以利用数学模型诊断的方法不适用；造纸生产过程中各个设备变量存在错综复杂的关联关系，基于数据的方法由于不能很好的表达这些关联关系导致诊断能力不足。而造纸产品质量问题以及造纸生产过程的故障诊断都需要这种能够表达复杂关联关系的知识库来帮助相关人员进行故障诊断。鉴于此，需要一种或多种方法针对造纸等复杂工业过程进行知识抽取与知识图谱的建立。

发明内容

[0004] 本发明要解决的问题是造纸领域知识图谱体系空白问题，为解决上述技术问题，本发明提供了一种用于造纸领域的产品质量知识图谱构建方法。

[0005] 本发明的目的通过以下的技术方案来实现：

[0006] 一种用于造纸领域的知识抽取及知识图谱构建方法，包括如下步骤：

[0007] 步骤(1)、收集数据：

[0008] 基于已有造纸行业结构化数据、互联网数据以及书籍数据生成造纸行业相关数据，这些数据包括已有相关设备和工艺的结构化数据，以及通过爬虫在相关造纸企业网站、造纸故障类网站、造纸产品质量问题相关网站采集的造纸产品质量问题相关文档信息；这些造纸产品质量相关文档信息包括造纸产品质量标准类文档信息、政策标准、专利、报告、百科；获取的数据通过收集、筛选、分析、汇总后形成造纸领域产品质量的基础数据；

[0009] 步骤(2)、数据分词：

[0010] 根据上述步骤(1)采集的数据信息，运用分词模型进行分词处理，最后形成造纸领域产品质量语料库；

[0011] 步骤(3)、数据标注：

[0012] 根据上述步骤(2)中的语料库，选取部分数据以汉语每个字为识别单位进行人工标注，然后将标注好的数据作为训练集；

[0013] 在人工标注过程中，标注的分类包含故障类型、故障名、故障设备名称、故障描述(现象)、故障原因以及故障解决办法。

[0014] 步骤(4)、知识抽取

[0015] 根据步骤(3)中的训练集建立命名实体识别模型并进行模型训练，用训练好的模型对所有文档进行知识的抽取。

[0016] 步骤(5)、构建产品质量知识图谱分类体系

[0017] 运用自顶而下的方式，采用人工构建的方式来构建基于造纸领域产品质量知识图谱的概念和关系分类体系。

[0018] 构建基于造纸领域产品质量知识图谱的概念和关系分类体系，包括：

[0019] 5.1、定义造纸产品质量问题的知识分类体系，设计了6类故障概念，分别是产品质量问题、产生原因、现象、解决办法、部位、检测；

[0020] 5.2、根据上述步骤5.1中6种已定义的造纸产品质量的故障概念，将泛化的概念共现关系按照语义类型进一步细分为7大类以及14小类的概念关系；

[0021] 7大类概念关系包括：故障诊断、故障表现、作用部位、质量检测、故障部位、检测结果、诊断依据；14小类的概念关系包括："产品质量问题"与"现象"、"现象"与"产生原因"、"故障决策"与"现象"之间的关系类型定义为"故障表现"；"产品质量问题"与"发生部位"、"部位"与"现象"之间的关系类型定义为"故障部位"；"产品质量问题"与"检测"、"检测"与"现象"之间的关系类型定义为"质量检测"；"产品质量问题"与"产生原因"、"产生原因"与"解决办法"、"解决办法"与"产品质量问题"之间的关系类型定义为"故障诊断"；"检测"与"部位"、"部位"与"解决办法"之间的关系类型定义为"作用部位"；"检测"与"现象"之间的关系类型定义为"检测结果"；"现象"与"产品质量问题"之间的关系类型定义为"诊断依据"。

[0022] 步骤(6)、知识存储

[0023] 根据构建的造纸领域产品质量概念和关系分类体系，将抽取出来的知识分类对应存储于Neo4j图数据库中。

[0024] 其中，知识图谱旨在描述真实世界中存在的各种实体或概念及其关系，节点表示实体或概念，边则由属性或关系构成。实体指的是具有可区别性且独立存在的某种事物。概念是具有同种特性的实体构成的集合，如产品质量、设备、工艺等。概念关系是用来描述两个概念之间的语义关系，是结构化知识的重要组成元素。

[0025] 与现有技术相比，本发明的技术方案具有如下有益效果：

[0026] 采用本发明的技术方案，通过对相关网站、书籍等的信息获取，得到造纸领域产品质量的基础数据，构建基于造纸领域产品质量知识图谱的概念和关系分类体系，将抽取出来的知识分类对应存储于Neo4j图数据库中，形成基于造纸领域产品质量知识图谱；本发明提供的技术方案还可以从造纸行业泛化到及其他复杂过程工业。

附图说明

[0027] 图1是本发明的整体工作示意图；

[0028] 图2是本发明产品质量故障知识图谱概念和分类体系图；

[0029] 图3是本发明的具体技术方案流程示意图；

具体实施方式

[0030] 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，所描述实施例仅仅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。

[0031] 如图1、图3所示，本发明提供了一种用于造纸领域的产品质量知识图谱构建方法的具体实例，包括如下步骤：

[0032] 步骤(1)、收集数据：

[0033] 基于已有造纸行业结构化数据、互联网数据以及书籍数据生成造纸行业相关数据，这些数据包括已有相关设备和工艺的结构化数据，以及通过爬虫在相关造纸企业网站、造纸故障类网站、造纸产品质量问题相关网站采集的造纸产品质量问题相关文档信息；这些造纸产品质量相关文档信息包括造纸产品质量标准类文档信息、政策标准、专利、报告、百科；获取的数据通过收集、筛选、分析、汇总后形成造纸领域产品质量的基础数据；

[0034] 步骤(2)、数据分词：

[0035] 根据上述步骤(1)采集的数据信息，运用分词模型进行分词处理，最后形成造纸领域产品质量语料库；

[0036] 以Thulac分词模型为例进行说明，Thulac分词模型的具体步骤如下：首先导入Thulac工具包；之后再读取收集到的造纸领域产品质量基础数据；然后调用Thulac工具包中的Thulac模块对收集的造纸领域产品质量的基础数据进行分词处理；接着将分词处理后的数据进行保存；最后对保存的数据进行去除标点符号和非法字符操作并以文本的形式重新保存；

[0037] 步骤(3)、数据标注：

[0038] 根据上述步骤(2)中的语料库，选取部分数据以汉语每个字为识别单位进行人工标注，然后将标注好的数据作为训练集；

[0039] 在人工标注过程中，标注的分类包含故障类型、故障名、故障设备名称、故障描述(现象)、故障原因以及故障解决办法。

[0040] 对语料库中出现的故障类型、故障名、故障设备名称、故障描述(现象)、故障原因以及故障解决办法等相关词分别标注为故障类型实体、故障名实体、故障设备名称实体、故障描述(现象)实体、故障原因实体、故障解决办法实体。以BMES四位序列标注法进行人工标注为例进行说明：B表示词首，M表示词中，E标注词尾，S表示单个词。

[0041] 步骤(4)、知识抽取

[0042] 根据步骤(3)中的训练集建立命名实体识别模型并进行模型训练，用训练好的模型对所有文档进行知识的抽取。

[0043] 以采用基于双向长短时记忆网络和条件随机场模型(简称Bi-LSTM+CRF模型)进行命名实体识别模型的训练，实现知识的抽取为例进行说明：首先将标注好的汉字映射为词向量作为模型的输入；之后将词向量输入到BiLSTM层，输出每个单词对应于每个标签的得分概率；然后在CRF层通过学习标签之间的顺序依赖信息得到最终的预测结果，并输出每个单词的预测的序列标注；最后，将文档数据输入到训练好的Bi-LSTM+CRF模型中进行知识的抽取。

[0044] 步骤(5)、构建产品质量知识图谱分类体系

[0045] 运用自顶而下的方式，采用人工构建的方式来构建基于造纸领域产品质量知识图谱的概念和关系分类体系。

[0046] 构建基于造纸领域产品质量知识图谱的概念和关系分类体系，包括：

[0047] 5.1、定义造纸产品质量问题的知识分类体系，设计了6类故障概念，分别是产品质量问题、产生原因、现象、解决办法、部位、检测；

[0048] 5.2、根据上述步骤5.1中6种已定义的造纸产品质量的故障概念，将泛化的概念共现关系按照语义类型进一步细分为7大类以及14小类的概念关系；

[0049] 5.1、定义造纸产品质量问题的知识分类体系，设计了6类故障概念，分别是产品质量问题、产生原因、现象、解决办法、部位、检测；

[0050] 5.2、根据上述步骤5.1中6种已定义的造纸产品质量的故障概念，将泛化的概念共现关系按照语义类型进一步细分为7大类以及14小类的概念关系；如图2所示，7大类概念关系包括：故障诊断、故障表现、作用部位、质量检测、故障部位、检测结果、诊断依据；如图2所示，14小类的概念关系包括：“产品质量问题”与“现象”、“现象”与“产生原因”、“故障决策”与“现象”之间的关系类型定义为“故障表现”；“产品质量问题”与“发生部位”、“部位”与“现象”之间的关系类型定义为“故障部位”；“产品质量问题”与“检测”、“检测”与“现象”之间的关系类型定义为“质量检测”；“产品质量问题”与“产生原因”、“产生原因”与“解决办法”、“解决办法”与“产品质量问题”之间的关系类型定义为“故障诊断”；“检测”与“部位”、“部位”与“解决办法”之间的关系类型定义为“作用部位”；“检测”与“现象”之间的关系类型定义为“检测结果”；“现象”与“产品质量问题”之间的关系类型定义为“诊断依据”。

[0051] 步骤(6)、知识存储

[0052] 根据构建的造纸领域产品质量概念和关系分类体系，将抽取出来的知识分类对应存储于Neo4j图数据库中。

[0053] 通过采用本发明公开的上述技术方案，得到了如下有益的效果：采用本发明的技术方案，通过对相关网站、书籍等的信息获取，得到造纸领域产品质量的基础数据，构建基于造纸领域产品质量知识图谱的概念和关系分类体系，将抽取出来的知识分类对应存储于Neo4j图数据库中，形成基于造纸领域产品质量知识图谱；本发明提供的技术方案还可以从造纸行业泛化到及其他复杂过程工业。

[0054] 本文中应用了具体个例对本发明的原理及实施方式进行了阐述，以上实例的说明只是用于帮助理解本发明的方法及其核心思想，以上所述仅是本发明的优选实施方式，应当指出，由于文字表达的有限性，而客观上存在无限的具体结构，对于本技术领域的普通技术人员来说，在不脱离本发明原理的前提下，还可以做出若干改进、润饰或变化，也可以将上述技术特征以适当的方式进行组合；这些改进润饰、变化或组合，或未经改进将发明的构思和技术方案直接应用于其他场合的，均应视为本发明的保护范围。

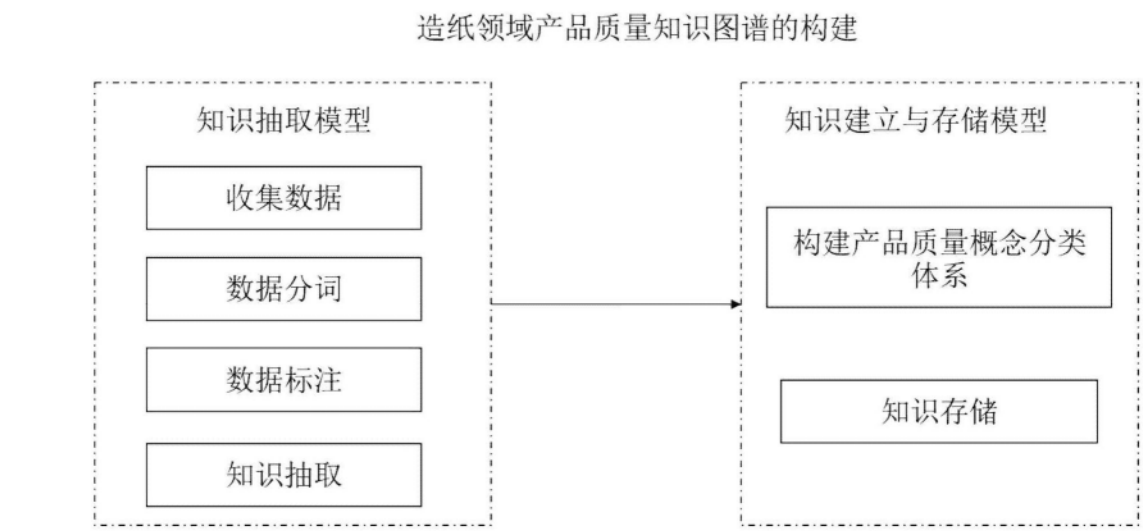

造纸领域产品质量知识图谱的构建
知识抽取模型
收集数据
数据分词
数据标注
知识抽取
知识建立与存储模型
构建产品质量概念分类体系
知识存储

图1

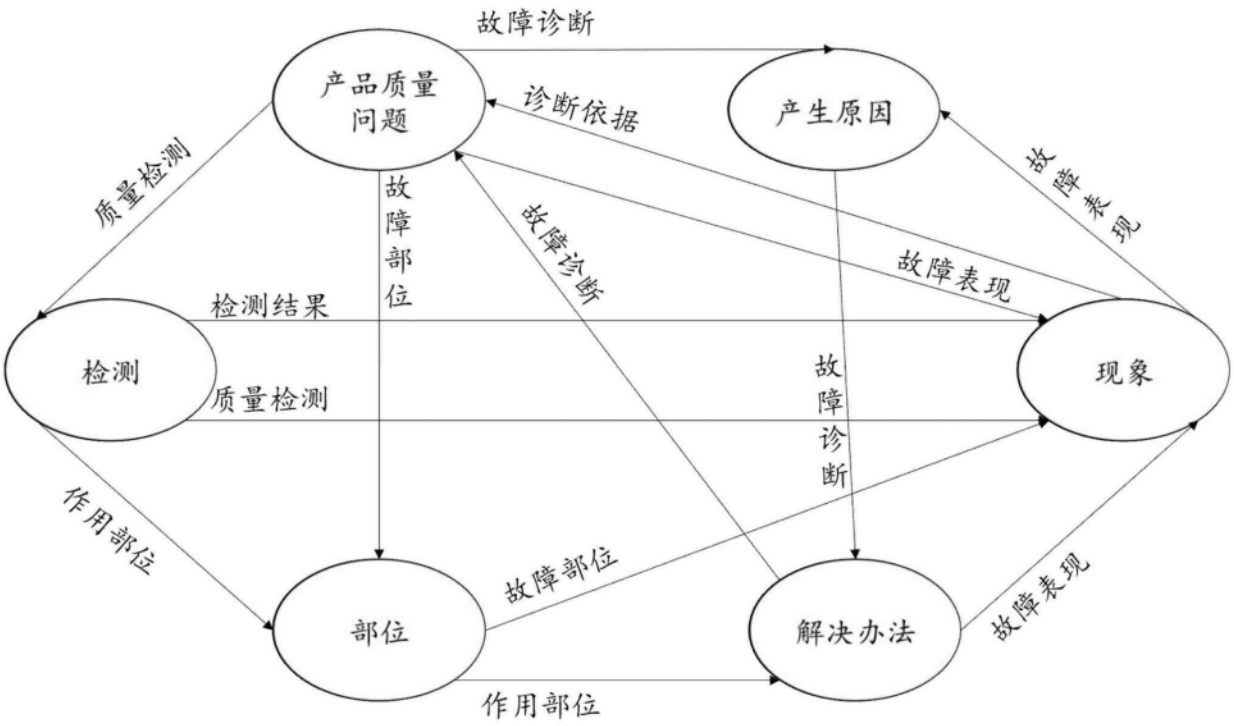

故障诊断
产品质量问题
诊断依据
产生原因
质量检测
故障部位
故障诊断
故障表现
故障表现
检测结果
检测
现象
质量检测
故障诊断
作用部位
故障部位
部位
解决办法
故障表现
作用部位

图2

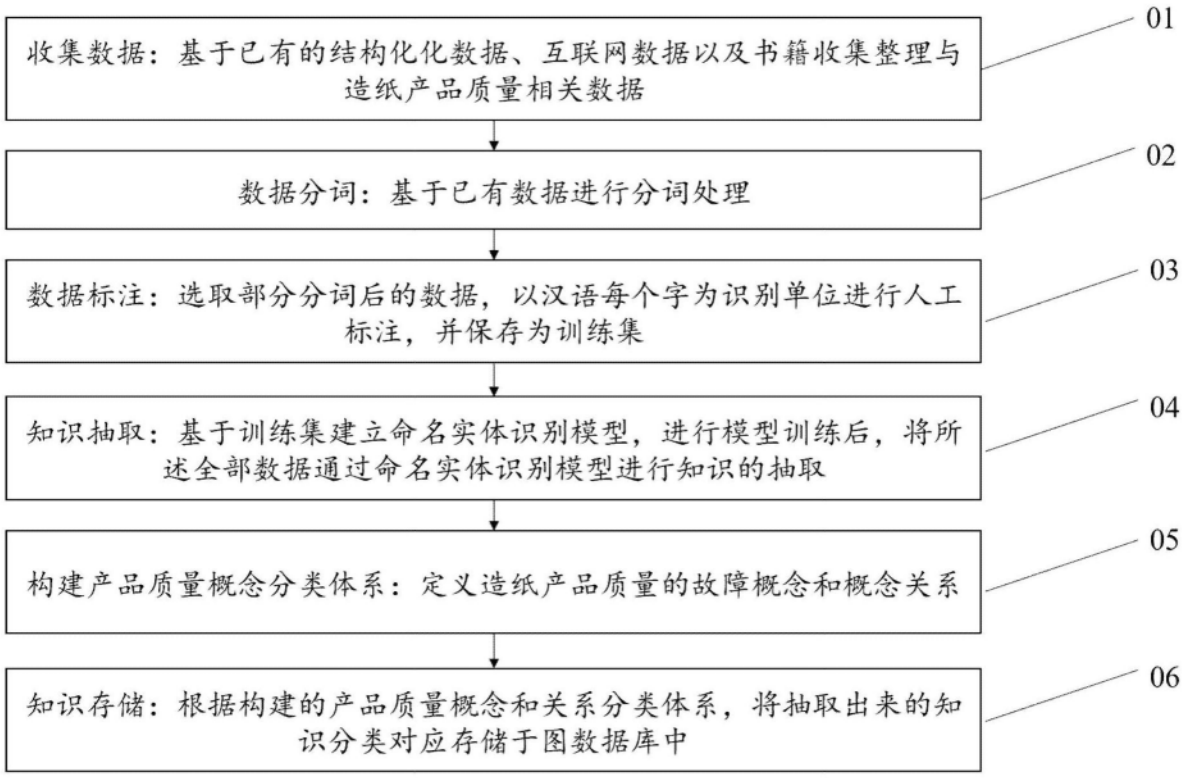

收集数据：基于已有的结构化化数据、互联网数据以及书籍收集整理与造纸产品质量相关数据
01
数据分词：基于已有数据进行分词处理
02
数据标注：选取部分分词后的数据，以汉语每个字为识别单位进行人工标注，并保存为训练集
03
知识抽取：基于训练集建立命名实体识别模型，进行模型训练后，将所述全部数据通过命名实体识别模型进行知识的抽取
04
构建产品质量概念分类体系：定义造纸产品质量的故障概念和概念关系
05
知识存储：根据构建的产品质量概念和关系分类体系，将抽取出来的知识分类对应存储于图数据库中
06

图3