### 基于Stanza的病毒知识图谱构建方法研究

#### 摘要

本文提出了一种基于结构化与非结构化病毒信息的知识图谱构建方法。该方法通过对病毒信息的分层处理，逐步提取和关联病毒特征、属性、关系信息，并将其构建成知识图谱。我们采用正则表达式、自然语言处理工具（Stanza）以及基于 Neo4j 的图数据库构建病毒知识图谱，以便更全面地表达和展示病毒信息。本文详细描述了病毒特征信息提取、实体关系解析与知识图谱构建过程。

### 1. 引言

在信息化时代，病毒信息的复杂性和多样性使得对其进行全面管理和展示具有重要的应用价值。知识图谱作为一种能够系统化表示信息的技术手段，为病毒特征信息的关联与呈现提供了新的方法。本文提出一种基于病毒特征信息的知识图谱构建方法，通过解析病毒信息中的文本特征，构建病毒、特征和关系的图谱结构。具体而言，我们以结构化的 JSON 数据为输入源，逐步提取病毒实体、特征和属性，并最终生成用于展示的知识图谱数据。

### 2. 数据加载与预处理

本文的数据来源为一组病毒信息 JSON 文件，每条记录包含病毒名称、发现日期、感染症状等字段。首先，读取 JSON 文件，逐条加载病毒信息，形成包含多条病毒数据的列表。在预处理中，我们对每条记录进行结构化提取，确保后续的特征解析和关系提取步骤可以顺利进行。

### 3. 特征信息提取方法

病毒特征信息是知识图谱构建的基础。我们设计了的具体提取方法如下：

**感染症状提取**：通过正则表达式匹配，从文本中提取症状信息，以字符串的形式保存。

**文件长度变化信息提取**：使用正则表达式匹配文本中的文件类型和长度增加量。提取的结果形成字典结构，以文件类型为键，文件长度增加量为值，以便后续使用。

该步骤确保了在病毒特征文本中的关键属性信息可以有效解析。

### 4. 基于自然语言处理的实体与关系提取

本文使用 Stanza 自然语言处理工具从病毒特征文本中自动提取实体、关系和属性。为了实现这一目标，我们设计了 TextProcessor 类来处理每条病毒特征文本，具体过程包括：

**实体提取**：基于 Stanza 工具，通过命名实体识别（NER）模块，提取文本中的实体。提取的实体类型包括 ORG（组织）、PRODUCT（产品）、QUANTITY（数量）、CARDINAL（基数）和 DISEASE（疾病）等，并以二元组的形式 (实体文本, 实体类型) 存储。

**关系与属性提取**：通过分析句法依存关系，识别出主-谓-宾结构中的核心关系。例如，谓语动词“感染”可以通过依存关系结构识别出来，形成三元组 (subject, relation, object)。属性信息则通过特征词的匹配提取，例如“感染状态”、“大小”等特征。

**三元组生成**：将提取的实体和属性信息格式化为 (实体, "is\_a", 实体类型) 以及 (属性名, "has", 属性值) 的结构，最终整合为知识图谱中的三元组。

### 5. 知识图谱构建

在知识图谱构建过程中，本文设计了 build\_knowledge\_graph 函数来整合各个病毒数据记录的特征信息，具体步骤如下：

**创建病毒节点**：首先将病毒名定义为主节点，每条记录初始化为 (virus\_name, "is\_a", "Virus") 形式的三元组。

**属性添加**：从 JSON 数据中提取各个属性（如 aliases、discovery\_date、length 等），将这些属性关联到病毒节点，并以 (病毒名, "has\_属性", 属性值) 的形式添加到知识图谱三元组中。

**病毒特征信息提取**：调用 extract\_virus\_characteristics 函数，获取症状信息和文件长度增加信息，进一步丰富病毒的特征内容，并形成三元组存储。

**实体、关系与属性的提取**：对于每条病毒信息，调用 TextProcessor 类的 process 方法提取出实体和关系三元组，并加入知识图谱三元组列表中。

**实时写入 JSON 文件**：为便于后续查询，将每个三元组作为 JSON 对象写入文件，每行一条记录。

**写入 Neo4j 数据库**：调用 create\_knowledge\_graph\_in\_neo4j 函数，将三元组写入 Neo4j 数据库，实现知识图谱的可视化。

### 6. 知识图谱可视化

为了实现知识图谱的可视化展示，我们利用 Neo4j 图数据库存储和呈现病毒知识图谱。具体实现包括以下步骤：

**建立数据库连接**：使用 GraphDatabase.driver 方法连接 Neo4j，设置 URI、用户名和密码。

**创建图谱节点与关系**：对每个三元组 (subject, relation, object)，通过 Cypher 语言创建或匹配节点和关系。病毒节点设置为 Virus 类型，对象节点设为 Entity 类型，关系类型根据 relation 字段动态指定。

**展示与分析**：Neo4j 提供丰富的图谱可视化工具，可以清晰展示病毒与其相关特征的关系图谱。借助该工具，用户能够直观地查看病毒特征、传播途径和可能的风险。

### 8. 结论

本文设计并实现了一种病毒信息知识图谱构建方法。该方法通过特征提取和自然语言处理技术，提取病毒信息中的重要特征、属性和关系，并基于 Neo4j 实现可视化的知识图谱构建。实验结果表明，本方法能够有效地从结构化和非结构化数据中提取病毒特征关系，并清晰地展示病毒的多维度信息。未来研究中，将进一步优化图谱展示效果，增加信息查询和分析功能。