**知识管理平台－知识图谱构建及应用建设**

1. 建设背景

知识图谱 (Knowledge Graph) 是一组实体以及这些实体间的关系，用于描述真实世界存在的概念以及概念之间的关系，从而令计算机能够通过模拟人类的知识体系，实现智能化知识存储、索引、分析以及处理能力。目前，最终用户的有害内容识别以及情报搜索等相关业务，基于关键字搜索技术，存在概念混淆、推送效能较低、缺乏情报推理与预测能力等问题。因此，本项目拟基于知识图谱构建技术研究，构建知识管理平台，从而提供更加智能化的服务。

知识图谱的技术构成主要在构建阶段，构建分为自顶向下和自底向上两种方式。自顶向下方式是指利用例如wikidata等经过详细整理的结构化数据构建知识图谱，自底向上的方式则是指从采集的原始数据中提取实体、关系、属性，从而构建知识图谱的过程。知识管理平台的主要作用是，通过整合和实现数据采集、自然语言处理、实体抽取、关系提取等相关技术，分别从自顶向下、自底向上两方面实现从原始情报数据到知识图谱的转化，为后续基于知识图谱的网信业务的智能化研究及应用提供基础支撑。

1. 建设目标

知识图谱构建技术研究需要先构建领域知识图谱。因此，本平台的主要建设目标是实现从原始情报数据到知识图谱的转化及可视化，从而为基于知识图谱的业务应用相关研究与应用提供基础支撑。知识图谱构建阶段主要包括知识表示、知识抽取、加工、融合以及知识存储。 知识表示是通过符号或向量表示方式，将原始数据转换成可计算的数据形态；知识抽取用于将原始数据中的实体、关系以及属性进行提取；知识存储是指通过三元组方式将提取出的实体、关系和属性进行存储，形成利于查询计算的图结构。实现半自动化的知识图谱构建过程。

知识图谱应用场景包括不限于有害识别、情报搜索、事件推理、突发事件识别、社交关系分析，以下服务内容需至少针对三种以上场景进行设计与实现。

1. 服务内容

能够实现从原始数据到知识图谱的转化，满足知识图谱存储、可视化及应用需求。知识图谱构建过程，主要包括知识图谱存储设计、知识图谱加工分析组件，知识图谱检索组件及智能支撑组件四部分建设内容。

* 1. 知识图谱存储设计

知识图谱可分为通用知识图谱与领域知识图谱。通用知识图谱存储了全部数据，存在搜索效率与准确率相对较低的问题。而领域知识图谱主要存储领域相关知识，在限定领域内有更优良的表现，但难以实现跨领域知识处理。由于网信业务横跨多个领域，如何界定领域范围和存储数据种类十分重要。网信领域情报图数据库设计与实现，需要针对网信实际业务需求，总结最小领域范围以及知识存储规则，在尽可能覆盖所有相关领域的前提下实现最优化的知识索引与处理。

需要满足三种以上的业务场景的设计与实现。

* 1. 知识图谱加工分析组件

本组件结合知识图谱、大数据分析、人工智能等技术，实现知识图谱实时构建和离线知识计算与分析。主要从本体构建、实体提取、关系抽取、知识清理、知识融合及知识推理几方面进行建设。

* + 1. 本体构建的设计与实现

构建本体的目的是识别、描述和表示相关领域的知识，提供对该领域知识的共同理解，确定领域内共同认可的对象模型，并从不同层次的形式化模式上给出这些对象和对象间相互关系的明确定义。

* + 1. 实体提取模块设计与实现

在原始数据到知识图谱的转化过程中，实体提取和关系抽取是最重要的工作内容。本平台通过整合自然语言处理及其他相关技术，对原始文本数据进行实体提取与预处理工作。

* + 1. 关系抽取模块设计与实现

实体与实体之间存在关系。通过实体和关系，计算机能够对不同概念进行区分、对不同事物进行识别。本平台设计和实现关系抽取模块，对实体间的关系进行自动化构建。

* + 1. 知识清理模块设计与实现

由于原始数据的复杂性与不确定性，平台需要对获取到的原始数据进行清理过滤，包括对同一概念的不同表达方式进行归一化处理、对于实体进行概念抽象、根据置信度对知识进行选择等，从而确保知识图谱的数据可靠性。

对新加入知识图谱的知识进行验证，确保新知识与现有知识图谱的一致性。

* + 1. 知识融合的设计与实现

知识融合指的是将多个数据源抽取的知识进行融合后集成到知识图谱中。在进行知识融合时，需要解决多种类型的数据冲突问题，包括一个短语对应多个实体、实体属性名不一致、实体属性缺失、实体属性值不一致、实体属性值一对多映射等情况。知识融合阶段主要对数据进行本体对齐和实体匹配。

* + 1. 知识推理

知识推理通过算法及其他方式整合，从现有知识中发现、获取新的有效知识或者结论，实现知识的自动补充。

* 1. 知识图谱检索组件

本组件结合知识图谱本身及应用场景，合理利用图数据库及搜索领域技术，实现对知识图谱的检索需求封装为高性能、可扩展的检索组件。主要从图查询接口进行建设。

* + 1. 图查询接口设计与实现

知识图谱采用图结构进行存储，其数据查询方式与传统数据库有很大区别。需要设计与实现统一且高效的图数据查询接口，用于本平台及应用场景下进行知识图谱的查询。实现亿级以上知识图谱的秒级查询，并支持向下兼容。

* 1. 智能支撑组件

本组件结合知识图谱和人工智能领域技术，及知识图谱本身及应用场景，实现人工智能相关技术支撑组件。主要从自然语言处理、图片处理几方面进行建设。

* + 1. 自然语言处理服务

提供文本处理服务设计与建设，主要结合人工智能技术、语言学、计算机科学，提供自然语言处理服务，为知识图谱加工分析组件提供存算法服务支撑。知识图谱令计算机能够通过模拟人类的知识体系，自然语言处理作为底层算法支撑，对文本信息进行识别处理。包括不限于语种识别、句子边界识别、分词、分句、关键词提取、命名实体识别、文本分类等功能。

本服务支持语种包括维、汉、哈、土、乌等六个语种。

* + 1. 图片处理服务

提供图片处理服务设计与实现，主要结合图像处理技术、人像处理技术等领域相关技术，为知识图谱加工分析组件提供图像处理服务支撑。包括不限于图像特征提取，图像相似度对比等功能，为知识清理、知识消歧等功能提供技术支撑。

1. 预期成果

建设面向网信领域的知识管理平台知识图谱构建与应用，完成知识图谱相关的设计与实现。