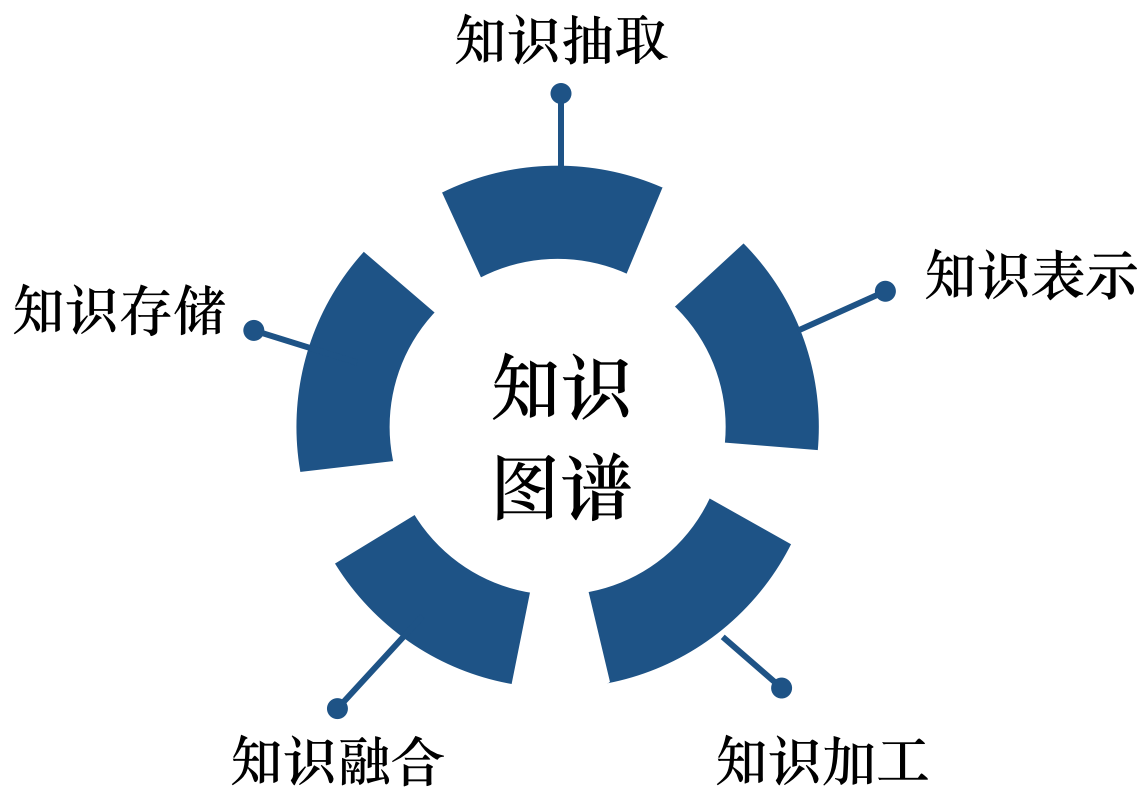
#### 主题知识构建

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。

主要建设目标是实现从基础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。



图错误! 文档中没有指定样式的文字。‑3知识图谱的构建阶段

实现知识图谱实时构建，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。实现自动化验证新知识和旧知识之间的不一致，以及对冗余知识，二义性表述知识的归一处理。

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. 国内外情况综述

知识管理领域已经蓬勃发展了近三十年，这一领域涉及的范围宽广,不仅包括计算机科学、工程学、运筹学、管理科学、图书情报学等领域,还延伸到医学、化学、航天等专业性很强的领域,甚至在制造业、电气电子、环境、机械和石油工业等部门发生了边缘应用。知识图谱就是知识库，对知识的展示直观、自然、直接和高效，对于大数据智能具有重要意义，将对自然语言处理、信息检索和人工智能等领域产生深远影响。所以,基于知识图谱进行知识管理研究有重要的学术意义和现实意义。

自2012年，google正式推出知识图谱搜索，为了让用户能够更快更简单的发现新的信息和知识，Google搜索将发布“知识图谱”（Knowledge Graph）——可以将搜索结果进行知识系统化，任何一个关键词都能获得完整的知识体系。站在用户的角度，谷歌的创新的确提供了更加快捷的搜索体验——只需轻轻一敲，信息尽在眼前。

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，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。

1. 知识计算：知识计算包括关联分析、知识聚类、知识搜索、知识抽取、知识加工、知识融合、实体对齐等技术。
2. 关联分析：关联分析又称关联挖掘，就是在交易数据、关系数据或其他信息载体中，查找存在于项目集合或对象集合之间的频繁模式、关联、相关性或因果结构。关联分析是一种简单、实用的分析技术，就是发现存在于大量数据集中的关联性或相关性，从而描述了一个事物中某些属性同时出现的规律和模式。
3. 知识聚类：所谓知识聚类是指根据数据的内在性质将数据分成一些聚合类，每一聚合类中的元素尽可能具有相同的特性，不同聚合类之间的特性差别尽可能大。聚类分析的目的是分析数据是否属于各个独立的分组，使一组中的成员彼此相似，而与其他组中的成员不同。它对一个数据对象的集合进行分析，但与分类分析不同的是，所划分的类是未知的，因此，聚类分析也称为无指导或无监督的(Unsupervised)学习。聚类分析的一般方法是将数据对象分组为多个类或簇(Cluster)，在同一簇中的对象之间具有较高的相似度，而不同簇中的对象差异较大。由于聚类分析的上述特征，在许多应用中，对数据集进行了聚类分析后，可将一个簇中的各数据对象作为一个整体对待。
4. 知识搜索：知识搜索的本质是通过数学来拜托当今搜索中使用的猜测和近似，并为词语的含义以及它们如何关联到我们在搜索引擎输入框中所找的东西引进一种清晰的理解方式。
5. 知识抽取：知识图谱的构建是后续应用的基础，而且构建的前提是需要把数据从不同的数据源中抽取出来。对于垂直领域的知识图谱来说，它们的数据源主要来自两种渠道：一种是业务本身的数据，这部分数据通常包含在公司内的数据库表并以结构化的方式存储；另一种是网络上公开、抓取的数据，这些数据通常是以网页的形式存在所以是非结构化的数据。前者一般只需要简单预处理即可以作为后续AI系统的输入，但后者一般需要借助于自然语言处理等技术来提取出结构化信息。
6. 知识加工：对于经过融合的新知识，需要经过质量评估之后（部分需要人工参与甄别），才能将合格的部分加入到知识库中，以确保知识库的质量。
7. 知识融合：通过知识提取，实现了从非结构化和半结构化数据中获取实体、关系以及实体属性信息的目标。但是由于知识来源广泛，存在知识质量良莠不齐、来自不同数据源的知识重复、层次结构缺失等问题，所以必须要进行知识的融合。知识融合是高层次的知识组织，使来自不同知识源的知识在同一框架规范下进行异构数据整合、消歧、加工、推理验证、更新等步骤，达到数据、信息、方法、经验以及人的思想的融合，形成高质量的知识库。
8. 知识对齐：实体对齐 (entity alignment) 也称为实体匹配 (entity matching)或实体解析(entity resolution)或者实体链接（entity linking），主要是用于消除异构数据中实体冲突、指向不明等不一致性问题，可以从顶层创建一个大规模的统一知识库，从而帮助机器理解多源异质的数据，形成高质量的知识。
9. 知识应用：知识图谱的应用场景很多，除了问答、搜索和个性化推荐外，在不同行业不同领域也有广泛应用。