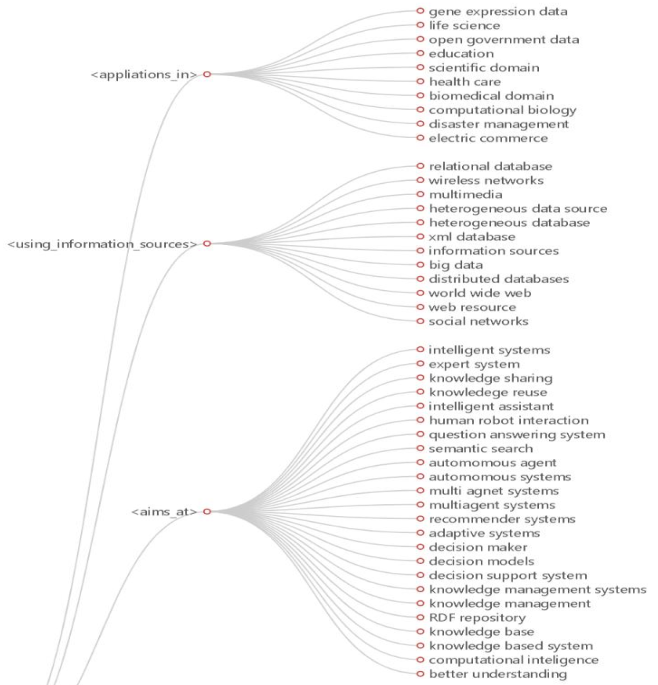
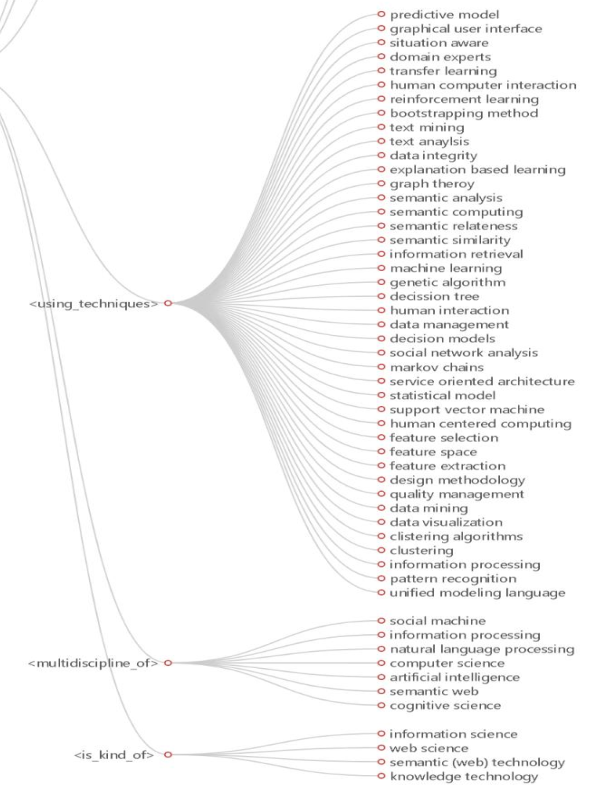
## 参考资料：

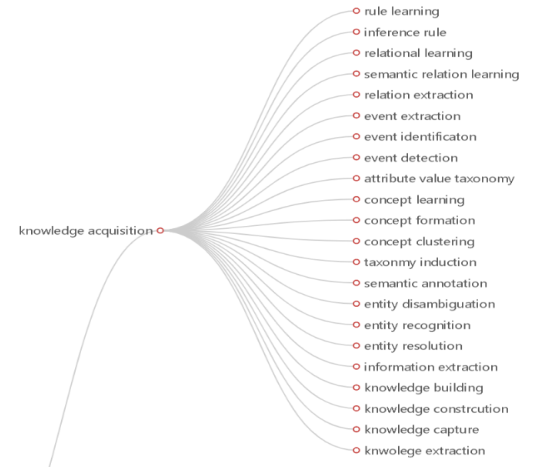
清华AMiner-2019年第二期《人工智能之知识图谱》

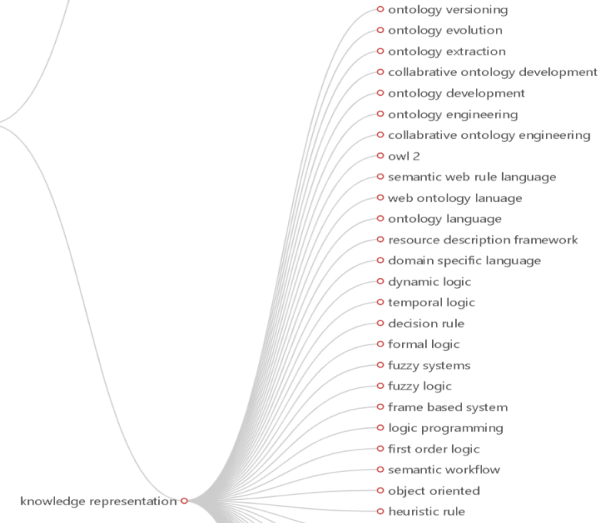
## 知识工程的相关概念：

注：图中带“<>”的节点表示关系，没有标”<>” 的标明的节点关系是上下位关系。



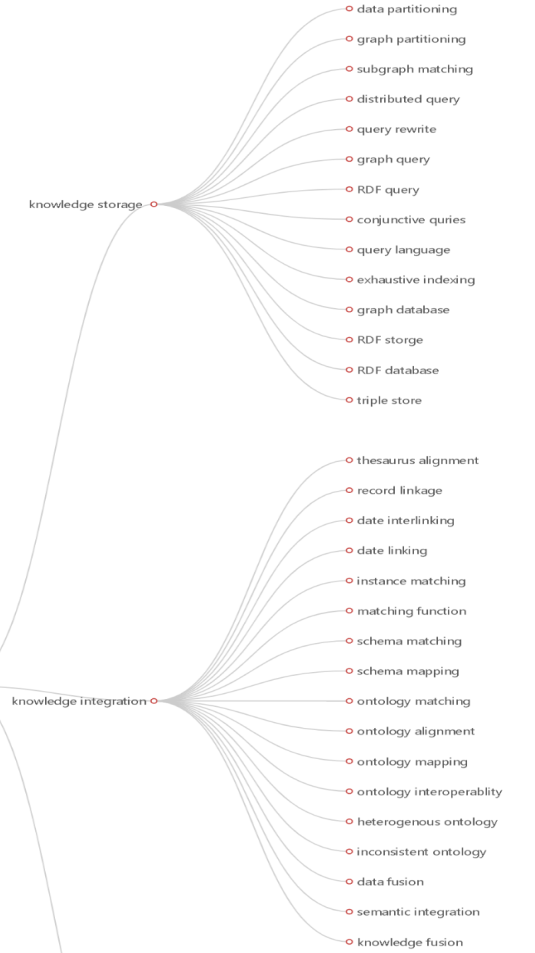


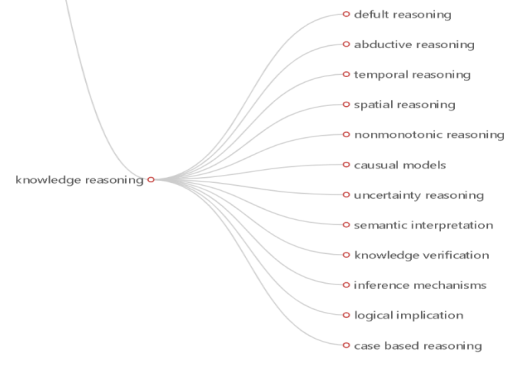












|  |  |
| --- | --- |
| **<using\_information\_source>** | **< using\_information\_source >** |
| relational database | 关系数据库 |
| wireless networks | 无线网络 |
| multimedia | 多媒体 |
| hererogerneous data source | hererogerneous数据源 |
| heterogerneous database | heterogerneous数据库 |
| xml database | xml数据库 |
| information sources | 信息来源 |
| big data | 大数据 |
| distributed databases | 分布式数据库 |
| world wide web | 万维网 |
| web resource | 网络资源 |
| social networks | 社交网络 |
|  |  |
| **<aims\_at>** | **< aims\_at >** |
| intelligent systems | 智能系统 |
| expert system | 专家系统 |
| knowledge sharing | 知识共享 |
| knowledge reuse | 知识重用 |
| intelligent assistant | 智能助理 |
| human robot interaction | 人类机器人交互 |
| question answering system | 问答系统 |
| semantic search | 语义搜索 |
| automomous agent | 自动代理 |
| automomous systems | 自动系统 |
| multi agent systems | 多代理系统 |
| recommender systems | 推荐系统 |
| adaptive systems | 自适应系统 |
| decision maker | 决策者 |
| decision models | 决策模型 |
| decision support system | 决策支持系统 |
| knowledge management systems | 知识管理系统 |
| knowledge management | 知识管理 |
| RDF repository | RDF存储库 |
| knowledge base | 知识库 |
| knowledge based system | 基于知识的系统 |
| computational intelligence | 计算智能 |
| better understanding | 更好的理解 |
|  |  |
| **<using\_techniques>** | **< using\_techniques >** |
| predictive model | 预测模型 |
| graphical user interface | 图形用户界面 |
| situation aware | 情况了解 |
| domain experts | 领域专家 |
| transfer learning | 转移学习 |
| human computer interactionre | 人机interactionre |
| inforcement learning | 结构加固学习 |
| bootstrapping method | 引导方法 |
| text mining | 文本挖掘 |
| text anaylsis | 文本分析 |
| data integrity | 数据完整性 |
| exp lanation based learni nggraph theroy | 基于经验的学习图理论 |
| se mantic analysis | se占卜的分析 |
| semantic computing | 语义计算 |
| semantic relate ness | 语义相关的湖水 |
| semantic similarity | 语义相似度 |
| information retrieval | 信息检索 |
| machine learning | 机器学习 |
| genetic algorithm | 遗传算法 |
| dedission tree | dedission树 |
| human interaction | 人机交互 |
| data management | 数据管理 |
| decision models | 决策模型 |
| social network analysis | 社会网络分析 |
| markov chains | 马尔可夫链 |
| service oriented architecturestatistical model | 面向服务的体系结构统计模型 |
| support vector machine | 支持向量机 |
| human centered computingfeature selection | 以人为中心的计算特征选择 |
| feature space | 特征空间 |
| feature extraction | 特征提取 |
| design methodology | 设计方法 |
| quality management | 质量管理 |
| data mining | 数据挖掘 |
| data visualization | 数据可视化 |
| clistering algorithms | clistering算法 |
| clustering | 聚类 |
| information processing | 信息处理 |
| pattern recognition | 模式识别 |
| unified modeling language | 统一建模语言 |
|  |  |
| **<multidiscipline\_of>** | **< multidiscipline\_of >** |
| social machine | 社会机 |
| information processing | 信息处理 |
| natural language processing | 自然语言处理 |
| computer science | 计算机科学 |
| artificial intelligence | 人工智能 |
| semantic web | 语义网 |
| cognitive science | 认知科学 |
|  |  |
| **<is\_kind\_of>** | **< is\_kind\_of >** |
| information science | 信息科学 |
| web science | 网络科学 |
| semantic (web) technologyknowledge technology | 语义(web)技术知识技术 |
|  |  |
| **Knowledge Acquisition** | **知识获取** |
| rule learning | 规则学习 |
| inference rule | 推理规则 |
| relational learning | 关系学习 |
| semantic relation learning | 语义关系学习 |
| relation extraction | 关系抽取 |
| event extraction | 事件提取 |
| event identificaton | 事件identificaton |
| event detection | 事件检测 |
| attribute value taxonomy | 属性值分类法 |
| concept learning | 概念学习 |
| concept formation | 概念形成 |
| concept clustering | 概念聚类 |
| taxonmy induction | taxonmy感应 |
| semantic annotation | 语义注释 |
| entity disambiguation | 实体消歧 |
| entity recogrition | 实体发展 |
| entity resolution | 实体解析 |
| information extraction | 信息提取 |
| knowledge building | 知识构建 |
| knowledge constrcution | 知识constrcution |
| knowledge capture | 知识获取 |
| knwolege extraction | knwolege提取 |
|  |  |
| **Knowledge Representation** | **知识表示** |
| ontology versioning | 本体版本管理 |
| ontology evolution | 本体进化 |
| ontology extraction | 本体抽取 |
| collabrative ontology development | collabrative本体开发 |
| ontology deve lopment | 本体开发lopment |
| ontology engineering | 本体工程 |
| collabrative ontology engineering | collabrative本体工程 |
| owl 2 | owl 2 |
| semantic web rule language | 语义web规则语言 |
| web ontology lanuage | 网络本体语言 |
| ontology language | 本体语言 |
| resource descri ption framework | 资源描述框架 |
| domain specific language | 领域特定语言 |
| dynamic logic | 动态逻辑 |
| temporal logic | 时序逻辑 |
| dedision rule | dedision规则 |
| formal logic | 形式逻辑 |
| fuzzy systems | 模糊系统 |
| fuzzy logic | 模糊逻辑 |
| frame based system | 基于框架的系统 |
| logic programming | 逻辑编程 |
| first order logic | 一阶逻辑 |
| se mantic workflow | se占卜的工作流 |
| object oriented | 面向对象的 |
| heuristic rule | 启发式规则 |
| rule based system | 基于规则的系统 |
| rule based | 基于规则的 |
| fuzzy sets | 模糊集 |
| concept map | 概念图 |
| rough set theroy | 粗糙集理论 |
| rough set | 粗糙集 |
| ontology | 本体 |
| se mantic network | se占卜的网络 |
| domain knowledge | 领域知识 |
| background knowledge | 背景知识 |
| web of data | 数据网络 |
| world knowledge | 世界知识 |
| commonsense knowledge | 常识性的知识 |
| formal semantics | 正式的语义 |
| formal meaning prepresentation | 正式意义prepresentation |
| formal specification | 正式规范 |
| structured data | 结构化数据 |
| knowledge model | 知识模型 |
| semantic model | 语义模型 |
| conceptual model | 概念模型 |
| concept model | 概念模型 |
| concept modelling | 概念模型 |
| data model | 数据模型 |
|  |  |
| **Knowledge Storage** | **知识存储** |
| data partitioning | 数据分区 |
| graph partitioning | 图分区 |
| subgraph matching | 子图匹配 |
| distributed query | 分布式查询 |
| query rewrite | 查询重写 |
| graph query | 图查询 |
| RDF query | RDF查询 |
| conjunctive quries | 连接quries |
| query language | 查询语言 |
| exhaustive indexing | 详尽的索引 |
| graph database | 图形数据库 |
| RDF storge | RDF仓储费 |
| RDF database | RDF数据库 |
| triple store | 三重存储 |
|  |  |
| **Knowledge Integration** | **知识集成/融合** |
| thesaurus alignment | 同义词典对齐 |
| record linkage | 记录链接 |
| date interlinking | 日期互连 |
| date linking | 日期链接 |
| instance matching | 实例匹配 |
| matching function | 匹配的函数 |
| schema matching | 模式匹配 |
| schema mapping | 模式映射 |
| ontology matching | 本体匹配 |
| ontology alignment | 本体对齐 |
| ontology mapping | 本体映射 |
| ontology interoperablity | 本体interoperablity |
| hete rogenous ontology | hete rogenous本体 |
| inconsistent ontology | 不一致的本体 |
| data fusion | 数据融合 |
| se mantic integration | se占卜的集成 |
| knowledge fusion | 知识的融合 |
|  |  |
| **Knowledge Reasoning** | **知识推理** |
| defult reasoning | defult推理 |
| abductive reasoning | 诱导的推理 |
| temporal reasoning | 时态推理 |
| spatial reasoning | 空间推理能力 |
| nonmonotonic reasoning | 非单调推理 |
| causual modeIs | causual modeIs |
| uncertainty reasoning | 不确定性推理 |
| semantic interpretation | 语义解释 |
| knowledge verification | 知识验证 |
| inference mechanisms | 推理机制 |
| logical implication | 逻辑含义 |
| case based reasoning | 基于案例推理 |

## 知识表示与建模

### 知识表示模型

知识表示技术的变化，大致可以分为三个阶段：

1）**基于符号逻辑进行知识表示和推理**，主要包括**逻辑表示法**（如一阶逻辑、描述逻辑）、**产生式表示法**和**框架表示**等。逻辑表示与人类的自然语言比较接近，是最早使用的一种知识表示方法；

2）万维网内容的知识表示技术，包括基于标签的半结构置标语言XML、基于万维网资源语义元数据描述框架RDF和基于描述逻辑的本体描述语言OWL等。当前在工业界大规模应用的多为基于**RDF三元组**的表示方法；

3）随着自然语言处理领域词向量等嵌入（Embedding）技术手段的出现，采用**连续向量方式来表示知识**的研究（TransE翻译模型、SME、SLM、NTN、MLP，以及NAM神经网络模型等）正在与上述以符号逻辑为基础知识表示方法相融合，成为现阶段知识表示的研究热点。更为重要的是，**知识图谱嵌入也通常作为一种类型的先验知识辅助输入到很多深度神经网络模型中，用来约束和监督神经网络的训练过程**。

### 知识表示学习

知识表示学习将实体和关系表示为**稠密的低维向量**实现了对实体和关系的**分布式表示**，已经成为**知识图谱语义链接预测和知识补全**的重要方法。

知识表示学习能够显著**提升计算效率，有效缓解数据稀疏，实现异质信息融合并有助于实现知识融合**，因此对知识库的构建、推理和应用具有重要意义。

**2种建模方法：**

* 复杂关系建模：TransE、TransH、TransR、TransD、TranSparse模型
* 关系路径建模：Path-based TransE（PTransE）模型

## 知识获取

### 实体识别与链接

实体识别与链接是海量文本分析、知识图谱的核心技术。

实体识别是文本理解意义的基础，也就是识别文本中指定类别实体的过程，可以检测文本中的新实体，并将其加入到现有知识库中。

实体链接是识别出文本中提及实体的词或者短语并与知识库中对应实体进行链接的过程，通过发现现有实体在文本中的不同出现，可以针对性的发现关于特定实体的新知识。

三种统计模型方法中的实体识别与链接：

**（1）传统统计模型方法**

**实体识别：**

最大熵分类模型、SVM 模型、隐马尔可夫模型、条件随机场模型等统计方法都曾被用来抽取文本中的实体识别，其中的**条件随机场模型**作为实体识别的代表性统计模型能够**将实体识别问题转化为序列标注问题**。

**实体链接：**

实体链接在传统模型中的核心在于**挖掘可用于识别提及目标实体相互关联的证据信息**，目前主要使用的证据信息包括**实体统计信息、名字统计信息、上下文词语分布、实体关联度、文章主题等信息**。同时，考虑到一段文本中实体之间的相互关联，相关的全局推理算法也被提出用来寻找全局最优决策。

**（2）深度学习方法**

**实体识别：**

目前存在两类用于命名实体识别NER的典型深度学习架构。

一种是 **NN-CRF 架构**，在该架构中，CNN/LSTM 被用来学习每一个词位置处的向量表示，基于该向量表示 NNCRF 解码该位置处的最佳标签。

第二种是采用**滑动窗口分类的思想**，使用神经网络学习句子中的每一个 N-Gram 的表示，然后预测该 N-Gram 是否是一个目标实体。

**实体链接：**

实体链接在深度学习的核心是构建多类型多模态上下文及知识的统一表示，并建模不同信息、不同证据之间的相互交互，通过将不同类型的信息映射到相同的特征空间，并提供高效的端到端训练算法。

**（3）文本挖掘方法**

文本挖掘方法应用于半结构 Web 数据源上的语义知识获取，工作核心是从特定结构（如列表、 Infobox）构建实体挖掘的特定规则，代表性文本挖掘抽取系统包括 DBPedia、 YAGO、BabelNet、 NELL 和 Kylin 等。由于规则本身可能带有不确定性和歧义性，同时目标结构可能会有一定的噪音，文本挖掘方法往往基于特定算法来对语义知识进行评分和过滤。

此外，人们发现结构化数据源只包含有限类别的实体，对长尾类别覆盖不足。

另一方面，实体获取技术往往采用 Bootstrapping 策略，充分利用大数据的冗余性，开放式的从 Web 中获取指定类型的实体。该部分的代表性的工作包括 TextRunner 系统和 Snowball 系统。

开放式实体集合扩展的主要问题是语义漂移问题，近年来的主要工作集中在解决该问题。

具体技术包括互斥 Bootstrapping 技术、 Co-Training 技术和 Co-Bootstrapping 技术。

### 实体关系学习

实体关系定义为两个或多个实体间的某种联系，用于描述客观存在的事物之间的关联关系。

实体关系学习就是自动从文本中检测和识别出实体之间具有的某种语义关系，也称为**关系抽取/关系识别**。

实体关系抽取分为预定义关系抽取和开放关系抽取。

* **预定义关系抽取**是指系统所抽取的关系是预先定义好的，如上下位关系、国家—首都关系等。
* **开放式关系抽取**不预先定义抽取的关系类别，由系统自动从文本中发现并抽取关系。

**（1）限定域关系抽取和开放域关系抽取**

**限定域关系抽取**是指系统所抽取的关系是预先定义好的，预定义关系个数有限。这类抽取可以抽取语义化的实体关系三元组，方便用于辅助其它任务。

**开放域关系抽取**是指不预先定义关系，由系统自动从文本中发现、抽取关系。由于开放域关系抽取**难以抽取语义化三元组**，近年来，越来越多的研究者关注限定域关系抽取。

**（2）基于规则的关系抽取和基于机器学习的关系抽取**

**基于规则的关系抽取**方法是指首先由通晓语言学知识的专家根据抽取任务的要求设计出一些包含词汇、句法和语义特征的**手工规则**（或称为模式），然后在文本分析的过程中寻找与这些模式相匹配的实例，从而推导出实体之间的语义关系。

基于机器学习的关系抽取可分成三大类：无监督关系抽取，有监督关系抽取、弱监督关系抽取。

* **无监督关系抽取**希望把表示相同关系的模版聚合起来，不需要人工标注的数据。
* 有监督关系抽取使用人工标注的训练语料进行训练。有监督关系抽取目前可以取得最好的抽取效果，但是由于其需要费时费力的人工标注，难以应用到大规模场景。
* 因此有学者提出了**利用知识库回标文本来自动获得大量的弱监督数据**，目前弱监督关系抽取是关系抽取领域的一大热点。

### 事件知识学习

事件是促使事物状态和关系改变的条件，是动态的、结构化的知识。目前已存在的知识资源（如谷歌知识图谱）所描述多是实体以及实体之间的关系，缺乏对事件知识的描述。

事件知识学习，就是**将非结构化文本中自然语言所表达的事件以结构化的形式呈现**，对于知识表示、理解、计算和应用意义重大。

知识图谱中的事件知识隐含互联网资源中，包括已有的结构化的语义知识、数据库的结构化信息资源、半结构化信息资源以及非结构化资源，不同性质的资源有不同的知识获取方法。

**（1）事件识别和抽取**

根据抽取方法，事件抽取可以分为基于模式匹配的事件抽取和基于机器学习的事件抽取。

**基于模式匹配的事件抽取方法：**

是指对某种类型事件的识别和抽取是在一些模式的指导下进行的， 模式匹配的过程就是事件识别和抽取的过程。采用模式匹配的方法进行事件抽取的过程一般可以分为两个步骤：模式获取和模式匹配。模式准确性是影响整个方法性能的重要因素，按照模式构建过程中所需训练数据的来源可细分为基于人工标注语料的方法和弱监督的方法。

**基于机器学习的事件抽取方法：**

建立在统计模型基础上，一般将事件抽取建模成多分类问题，因此研究的重点在于**特征和分类器的选择**。根据利用信息的不同可以分为基于特征、基于结构和基于神经网络三类主要方法。

* **基于特征的方法**：研究重点在于**如何提取和集成具有区分性的特征**，从而产生描述事件实例的各种局部和全局特征，作为特征向量输入分类器。该类方法多用于阶段性的管道抽取，即顺序执行事件触发词识别和元素抽取。
* **基于结构的方法**：**将事件结构看作依存树**，抽取任务则相应地转化为依存树结构预测问题，触发词识别和元素抽取可以同时完成。
* **神经网络的方法**：利用**卷积神经网络模型**抽取特征来完成两阶段的识别任务以便更好地考虑事件内部结构和各个元素间的关系。**将联合抽取模型与 RNN 相结合，利用带记忆的双向 RNN 抽取句子中的特征，并联合预测事件触发词和事件元素**，进一步提升了抽取效果。

**（2）事件检测和追踪**

**基于相似度的方法：**

首先需要**定义相似度度量**，而后基于此进行**聚类或者分类**。 Yang 等提出在 TDT 中用向量空间模型（Vector Space Model， VSM）对文档进行表示，并提出了组平均聚类（Group Average Clustering， GAC）和单一通过法（Single Pass Algorithm， SPA）两种聚类算法。 GAC 只适用于历史事件发现，它利用分治策略进行聚类。 SPA 可以顺序处理文档并增量式产生聚类结果，能同时应用于历史事件发现和在线事件发现。

**概率统计方法：**

通常使用生成模型，由于需要大量数据的支持，所以这种方法更加适用于历史事件检测。对比基于相似度聚类的模型，这类模型虽然复杂，但当数据量充足时，通常可以取得更高的准确率。基于概率的方法是目前 TDT 中的研究热点，主要分成两个方向，一是针对新闻等比较正式的规范文档，另一个则用于不规则或没有规律的非规范文档。

**（3）事件知识库构建**

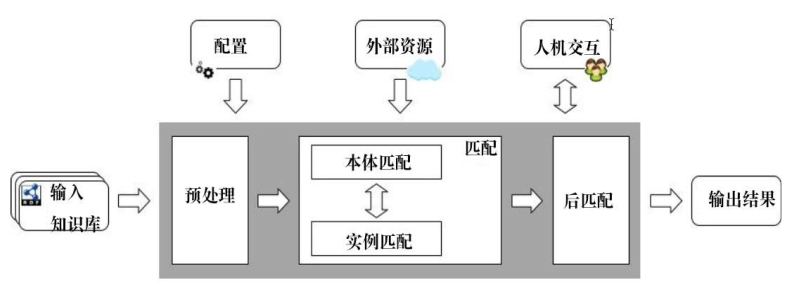
已有知识图谱，如 DBpedia， YAGO 和 Wikidata 等均**侧重于实体的客观属性及实体间的静态关联，缺乏结构化的事件数据**。事件知识学习的最终**目的就是从非结构化的文本数据中抽取结构化的事件表示，构建事件知识库弥补现有知识图谱的动态事件信息缺失问题**。

目前事件知识库构建的研究处于起步阶段，基础就是上述两方面研究，**基于句子级的事件抽取和文档级的事件发现**。

## 知识融合

语义集成的提出是为了能够将不同的知识图谱融合为一个统一、一致、简洁的形式，为使用不同知识图谱的应用程序间的交互建立操作性。

常用的技术包括**本体匹配（也称为本体映射）、实力匹配（也称为实体对齐、对象公指消解）以及知识融合**等。



一个语义集成的常见流程，主要包括：输入、预处理、匹配、知识融合和输出 5 个环节。语义集成的输入包括待集成的若干个知识库以及配置、外部资源等。

**待集成的知识库**格式一般为 RDF/OWL 数据文件或 SPARQL 端点（endpoint）。外部资源是语义集成过程中使用到的背景知识，例如字/辞典背景知识（例如 WordNet）、常识背景知识（例如 Cyc）、实时背景知识（例如搜索引擎）等。

**预处理**主要包括对输入知识库进行清洗和后续步骤的准备。

**清洗**主要是为了解决输入质量问题，与自有文本不同，知识库通常基于 RDF/OWL 语言构建，质量较好。

**后续步骤的准备**分为配置和数据两方面。根据匹配对象的不同，匹配一般分为本体匹配和实例匹配两方面。

**文本相似性度是发现匹配的最基础方法**，大致可分为四种类型：**基于字符的**（例如 Leven-shtein 编辑距离）、**基于单词的**（例如 Jaccard 系数）、**混合型**（例如 soft TF-IDF）和**基于语义的**（例如 WordNet）。

在匹配的基础上，知识融合一般通过**冲突检测、真值发现**等技术消解知识集成过程中的冲突，再对知识进行关联与合并，最终形成一个一致的结果。语义集的输出是一个统一的、一致的、简洁的知识库。

### 本体匹配

现有大多数本体匹配方法处理的是成对的本体，但是成对匹配方法在同时匹配多个本体时会产生的最主要问题是它们得到的结果从全局看可能存在冲突。 **LPHIM** 是一种多文本全体匹配方法，能够在匹配多个本体的同时保证结果是全局最优解。

随着多语言知识库的发展，跨语言本体匹配方法的重要性已经凸显。较有代表性的工作包括： EAFG 和双语主题模型。

### 实例匹配

众包和主动学习等人机协作方法是目前实例匹配的研究热点。这些方法雇佣普通用户，通过付出较小的人工代价来获得丰富的先验数据，从而提高匹配模型的性能。

随着表示学习技术在诸如图像、视频、语言、自然语言处理等领域的成功，一些研究人员开始着手研究面向知识图谱的表示学习技术，**将实体、关系等转换成一个低维空间中的实质向量（即分布式语义表示）**，并在知识图谱补全、知识库问答等应用中取得了不错的效果。

近年来，如何在语义集成中运用强化学习逐渐成为新的动向。

## 知识图谱查询和推理

### 知识推理

知识推理从给定的知识图谱推导出新的实体跟实体之间的关系。知识图谱推理可以分为**基于符号的推理**和**基于统计的推理**。

* **基于符号的推理：**一般是基于经典逻辑（一阶谓词逻辑或者命题逻辑）或者经典逻辑的变异（比如说缺省逻辑）。
* **基于符号的推理：**可以从**一个已有的知识图谱推理出新的实体间关系，可用于建立新知识或者对知识图谱进行逻辑的冲突检测**。基于统计的方法一般指关系机器学习方法，即通过统计规律从知识图谱中学习到新的实体间关系。

知识推理在知识计算中具有重要作用，如知识分类、知识校验、知识链接预测与知识补全等。

**（1）基于符号的并行知识推理**

**基于多核、多处理器技术的大规模推理**：

单机环境下的并行技术以共享内存模型为特点，侧重于提升本体推理的时间效率，适用于对于实时性要求较高的应用场景，这种方法成为首选。对于表达能力较低的语言，比如 RDFS、 OWL EL，单机环境下的并行技术显著地提升了本体推理效率。

**基于分布式技术的大规模推理**：

基于分布式技术可以突破大规模数据的处理界限，这种方法利用多机搭建集群来实现本体推理，很多工作基于 MapReduce 的开源实现设计提出了大规模本体的推理方法。

**（2） 链接预测**

**基于表示学习的方法：**

知识图谱表示学习旨在于将知识图谱中的实体与关系统一映射至低维连续向量空间，以刻画它们的潜在语义特征。通过比较实体与关系在该向量空间中的分布式表示，可以推断出实体和实体之间潜在的关系。

**基于图特征的方法：**

基于图特征的方法借助从知识图谱中抽取出的图特征来预测两个实体间可能存在的不同类型的边（关系）。例如，根据两个实体“姚明”和“叶莉”在知识图谱中的联通路径可以预测出他们之间大概率具备“配偶”关系。

**（3）模式归纳方法**

**基于 ILP 的模式归纳方法：**

基于 ILP 的方法进行本体学习的早期工作给出了很好的综述。 Jens Lehmann 等提出用向下精化算子学习 ALC 的概念定义公理的方法，并在后续工作中将原有方法扩展到处理大规模知识库上。相关的算法都在本体学习工具 DL-Learner 中得到实现，并且在工作中得到进一步扩展，涉及到框架的设计和可扩展性的提升等方面。

**基于关联规则挖掘的模式归纳方法：**

利用谓词偏好因子度量方法以及谓词语义相似度学习相反和对称公理；利用模式层信息给规则的挖掘提供更多的语义；对传统关联规则挖掘技术进行了改进，事务表中用 0 到 1 之间的一个实数代替原来的 0 或者 1，使得提出的方法更符合语义数据开放的特点。

**基于机器学习的模式归纳方法：**

利用聚类的算法学习关系的定义域和值域；应用统计的方法过滤属性的使用，并找出准确、健壮的模式，用于学习属性的数量约束公理。

### 知识存储和查询

知识图谱以图（Graph）的方式来展现实体、事件及其之间的关系。

知识图谱存储和查询研究**如何设计有效的存储模式支持对大规模图数据的有效管理**，实现对知识图谱中知识高效查询。

**（1）基于关系数据模型的 RDF 数据存储和查询**

**简单三列表：**

系统通过维护一张巨大的**三元组表来管理 RDF 数据**。这张三元组表包含三列，对应存储主体、谓词和客体（或者主体、属性和属性值）。

当系统接收到用户输入的SPARQL 查询时，这些系统将 SPARQL 查询转化为 SQL 查询。然后根据所得 SQL 查询，这些系统通过对三元组表执行多次自连接操作以得到最终解。

**水平存储：**

将知识图谱中的每一个 RDF 主体（subject）表示为数据库表中的一行。表中的列包括该 RDF 数据集合中所有的属性。

这种的策略的好处在于设计简单，同时很容易回答面向某单个主体的属性值的查询，即星状查询。

存储方法的缺点也是很明显的：其一，表中存在大量的列；其二，表的稀疏性问题；其三，水平存储存在多值性的问题；其四，数据的变化可能带来很大的更新成本。

**属性表：**

为降低自连接操作次数， Jena 和 Oracle 在单张大三元组表之外还支持利用属性表进行 RDF 数据管理。

具体而言， Jena 通过聚类的方式将一些类似的三元组聚类到一起，然后将每一个聚类的三元组统一到一张属性表中进行管理，这种方式下的属性表也被称之为聚类属性表；而 Oracle 利用 RDF 资源的类型信息将三元组进行分类，相同类的三元组放到同一张表中，这种方式下的属性表也被称之为分类属性表。

**垂直划分策略：**

SW-Store 提出了对 RDF 数据按照谓词（或属性）分割成若干表的方法。

具体而言， SW-Store 将 RDF 三元组按照谓词（或属性）的不同分成不同的表，每张表能保存在谓词（或属性）上相同的三元组。 SW-Store 称这种方法为垂直分割。

**全索引策略：**

简单的三列表存储的缺点在于自连接次数较多。为了提高简单三列表存储的查询效率，目前一种普遍被认可的方法是“全索引（exhaustive indexing）”策略。

**（2）基于图模型的 RDF 数据存储和查询**

RDF 数据的图模型可以最大限度的保持 RDF 数据的语义信息，也有利于对语义信息的查询。

在这种情况下， SPARQL 查询就可以视为在 RDF 数据图上进行子图匹配运算。子图匹配运算是图数据库中一个比较经典的问题：其问题定义在于给定一个数据图和一个查询图，找出数据上所有与查询图子图同态的位置。这个问题已被证明是一个 NP 难问题。

针对RDF 数据的 SPARQL 查询已经有一些基于图模型的查询处理系统，如 **gStore**、和**TurboHOM++**。它们都是利用 RDF 数据图的特点来构建索引。

## 知识应用

### 典型应用

知识应用能够将知识图谱特有的应用形态与领域数据与业务场景相结合并助力领域业务转型。

知识图谱的典型应用包括语义搜索、智能问答以及可视化决策支持三种。

如何**针对业务需求设计实现知识图谱应用，并基于数据特点进行优化调整**，是知识图谱应用的关键研究内容。

**（1） 语义搜索**

知识图谱是对客观世界认识的形式化表示，将字符串映射为客观事件的事务。

当前基于关键词的搜索技术在知识图谱的知识支持下可以上升到**基于实体和关系的检索**，称之为**语义搜索**。

语义搜索利用知识图谱可以准确地捕捉用户搜索意图，进而基于知识图谱中的知识解决传统搜索中遇到的关键字语义多样性及语义消歧的难题，通过实体链接实现知识与文档的混合检索。

语义搜索需要考虑**如何解决自然语言输入带来的表达多样性问题**，同时需要解决语言中**实体的歧义性**问题。同时借助于知识图谱，直接给出满足用户搜索意图的答案，而不是包含关键词的相关网页的链接。

**（2） 智能问答**

问答系统（Question Answering，QA）是能够让计算机自动回答用户所提出的问题。

不同于现有的搜索引擎，问答系统返回用户的不再是基于关键词匹配的相关文档排序，而是精准的自然语言形式的答案。

“以直接而准确的方式回答用户自然语言提问的自动问答系统将构成下一代搜索引擎的基本形态”。

智能问答需要针对用户输入的自然语言进行理解，从知识图谱中或目标数据中给出用户问题的答案，其关键技术及难点包括**准确的语义解析、正确理解用户的真实意图、以及对返回答案的评分评定以确定优先级顺序**。

**（3） 可视化决策支持**

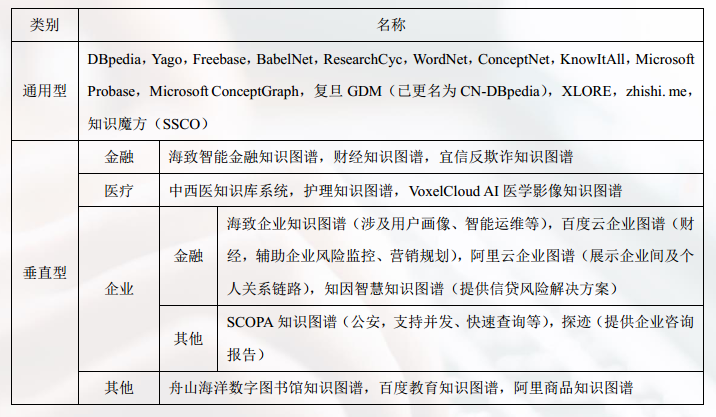
可视化决策支持是指通过**提供统一的图形接口，结合可视化、推理、检索等，为用户提供信息获取的入口**。

例如，决策支持可以通过图谱可视化技术对创投图谱中的初创公司发展情况、投资机构投资偏好等信息进行解读，通过节点探索、路径发现、关联探寻等可视化分析技术展示公司的全方位信息，通过知识地图、时序图谱等形态对地理分布、发展趋势等进行解读，为投融资决策提供支持。

可视化决策支持需要考虑的关键问题包括通过可视化方式辅助用户快速发现业务模式、提升可视化组件的交互友好程度、以及大规模图环境下底层算法的效率等。

### 通用和领域知识图谱

知识图谱分为通用知识图谱与领域知识图谱两类，两类图谱本质相同，其区别主要体现在覆盖范围与使用方式上。



**通用知识图谱：**

主要强调知识的广度，可以形象地看成一个面向通用领域的结构化的百科知识库，其中包含了大量的现实世界中的常识性知识，覆盖面广，通常运用百科数据进行**自底向上（Top-Down）的方法进行构建**。

**通用知识库：**

**YAGO知识库**主要集成了Wikipedia、WordNet和GeoNames三个来源的数据，拥有千万级实体知识，包含超过1.2亿条三元组知识，能够将WordNet的词汇定义与Wikipedia的分类体系进行了融合集成。YAGO还考虑了时间和空间知识，为很多知识条目增加了时间和空间维度的属性描述，具有更加丰富的实体分类体系，经过人工评估证实确认，准确度达到 95%。

**NELL系统**目标是能够开发用自然语言回答用户提出的问题的方法，而不需要人为干预。

**Zhishi.m**从开放的百科数据中抽取结构化数据，当前已融合了包括百度百科、

互动百科、中文维基三大百科的数据，拥有 1000 万个实体数据、一亿两千万个 RDF 三元组。

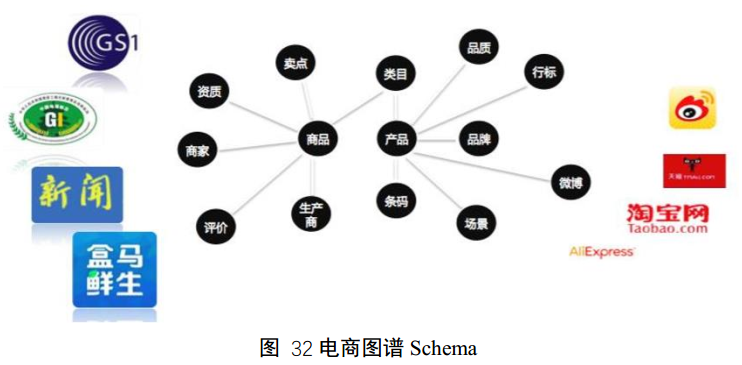
**XLore**则是基于中文维基百科、英文维基百科、百度百科、互动百科构建的大规模中英文知识平衡知识图谱。

**领域知识图谱：**

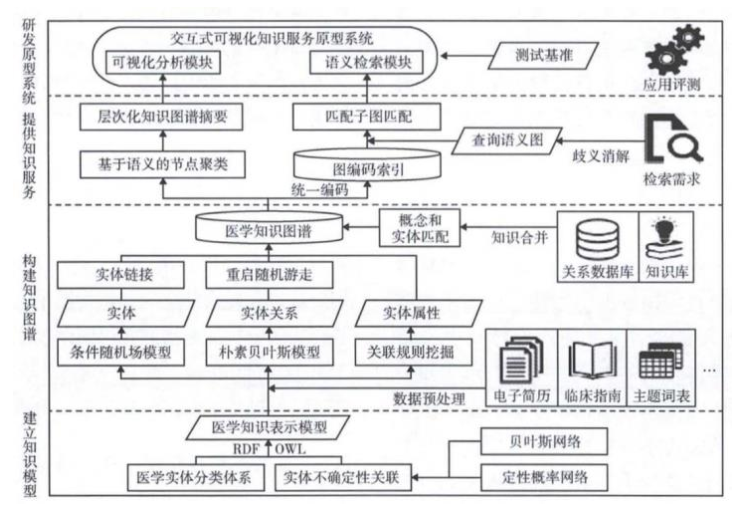
又被称为行业知识图谱或垂直知识图谱，可看成是一个面向某一特定领域的**基于语义技术**的行业知识库，有着严格而丰富的数据模式，应用需求各不相同，因此没有一套通用的标准和规范来指导构建，需要基于特定行业通过工程师与业务专家的不断交互沟通与定制来实现，所以对该领域知识的深度、知识准确性有着更高的要求。

**电商知识图谱：**

电商知识图谱以商品为核心，以人、货、场为主要框架。目前共涉及 9大类一级本体和 27 大类二级本体。一级本体分别为：人、货、场、百科知识、行业竞对、品质、类目、资质和舆情。人、货、场构成了商品信息流通的闭环，其他本体主要给予商品更丰富的信息描述。



**医疗知识图谱的构建与应用：**



## 趋势与瓶颈

**瓶颈：**

**知识类型与表示：**

知识图谱主要采用的三元组形式可以较好的表示事实性知识，但对各种类型的复杂知识就不适用了。

**知识获取：**

知识抽取的准确率和效率还有待提高。

**知识融合：**

不同来源的展示存在大量噪声和冗余、不同语言等，关键在于如何对其进行融合。

**知识应用：**

如何有效实现知识图谱的应用，利用知识图谱实现深度知识推理，提高大规模知识图谱计算效率，需要人们不断锐意发掘用户需求，探索更重要的应用场景，提出新的应用算法。

**趋势：**

* 特色化
* 开放化：开放知识图谱社区（Open KG）
* 智能化