

**知识图谱技术综述**

**知识图谱A组**

**知识图谱技术综述**

刘影 李欣悦 张心阳 钟豪

北京理工大学 北京

**摘 要** 知识图谱技术作为一门新兴的技术，是人工智能技术的重要组成部分，其建立的具有语义处理能力与开放互联能力的知识库，可在智能搜索、智能问答、个性化推荐等智能信息服务中产生应用价值。该文在全面阐述知识图谱定义、架构的基础上，综述知识图谱中的知识抽取、知识表示、知识融合、知识推理四大核心技术的研究进展以及一些实际生活中的典型应用。该文还将评论当前研究存在的挑战。

关键词 知识图谱；知识抽取；知识融合；知识库；知识推理

**团队分工**

**刘影：知识图谱定义与架构**

**李欣悦：知识获取与知识表示**

**张心阳：知识融合与知识推理**

**钟豪：知识图谱的技术应用**

# 第一章 绪论

1.1 研究背景

随着互联网的飞速发展，互联网信息总量呈现爆炸性的增长趋势。同时大数据时代的互联网信息内容呈现出大规模、多元化、组织结构松散的特点，传统的搜索虽然网页检索效率较高，但是用户常常不能快速的获取准确的信息，通常还要经过人工对检索所得的大量的结果进行排查筛选，目前，更多新的搜索形态诸如语音搜索、拍照搜索、人机对话等不断产生并快速演化，用户同样希望搜索引擎能够结合着自身喜好、地理位置、时间点等给出更精确的回答。准确有效地定位相关知识成为网络用户的迫切需求，这种信息检索方式已经很难满足人们全面掌控信息资源的需求。因此，人们需要根据新环境下知识组织的原则［1］，从新的视角探索既符合网络信息资源发展变化新特点又适应用户认知需求的知识组织方法［2］，从更深层次上揭示人类知识的整体性和关联性［3］。知识图谱技术的出现为解决信息检索问题提供了新的思路。

在2006年，Berners-Lee提出了数据链接（linked data）的思想，呼吁推广和完善相关的技术标准如URI(uniform resource identifier)，RDF（resource description on framework）,OWL(web ontology language)[4]，为迎接语义网络时代的到来做了好准备，从而掀起了一场语义网研究的热潮。根据万维网联盟的解释，语义网络是一张数据构成的网络，语义网络技术向用户提供的是一个查询环境，其核心要义是以图形的方式向用户返回经过加工和推理的知识．基于语义网的相关研究， 2012年5月17日， Google正式推出其用于改善下一代搜索引擎功能的辅助知识库—知识图谱。知识图谱运用“搜索+知识库”的方式合理有序地组织海量的网络数据，以更直观的方式把知识展示给用户。它可以描述不同层次和粒度的抽象概念，是对客观世界的知识映射，可以用作互联网资源组织的基础，使更大范围内资源的知识整合与检索成为可能。

1.2 国内外研究进展

就覆盖范围而言，知识图谱也可分为通用知识图谱和行业知识图谱。通用知识图谱注重广度，强调融合更多的实体，较行业知识图谱而言，其准确度不够高，并且受概念范围的影响，很难借助本体 库对公理、规则以及约束条件的支持能力规范其实体、属性、实体间的关系等。通用知识图谱主要应用于智能搜索等领域。行业知识图谱通常需要依靠特定行业的数据来构建，具有特定的行业意义。行业知识图谱中，实体的属性与数据模式往往比较丰富，需要考虑到不同的业务场景与使用人员。我国对于中文知识图谱的研究已经起步，并取 得了许多有价值的研究成果．早期的中文知识库主要采用人工编辑的方式进行构建，例如中国科学院计算机语言信息中心董振东领导的知网（How Net）项目，其知识库特点是规模相对较小、知识质量高、但领域限定性较强。由于中文知识图谱的构建对中文信息处理和检索具有重要的研究和应用价值，近年来吸引了大量的研究。例如在业界，出现了百度知心、搜狗知立方等商业应用。在学术界，清华大学建成了第１个大规模中英文跨语言知识图谱XLore[5]、 中国科学院计算技术研究所基于开放知识网络(Open KN)建立了“人立方、事立方、知立方”原型系统、中国科学院数学与系统科学研究院陆汝钤院士提出知件(Know ware)的概念、上海交通大学构建并发布了中文知识图谱研究平台zhishi.me[6] 、复旦大学GDM实验室[7]推出的中文知识图谱项目等[8]， 这些项目的特点是知识库规模较大，涵盖的知识领域较广泛，并且能为用户提供一定的智能搜索及问答服务。

表1 知识图谱及相关类似产品

|  |  |  |  |
| --- | --- | --- | --- |
| 知识图谱库名称 | 机构 | 特点、构建手段 | 应用产品 |
| FreeBase | MetaWeb(2010年被Google收购) | * 实体、语义类、属性、关系； * 自动+人工：部分数据从维基百科等数据源抽取而得到；另一部分数据来自人工协同编辑 * https://developeers.google.com.freebase/ | Google,Search Engine,Google Now |
| Knowledge Vault（谷歌知识图谱） | Google | * 实体、语义类、属性、关系； * 超大规模数据库；源自维基百科、FreeBase、《世界各国纪实年鉴》 * <https://research.google.com>/pubs/pub45634 | Google,Search Engine,Google Now |
| DBpedia | 莱比锡大学、柏林自由大学、OpenLink Software | * 实体、语义类、属性、关系； * 从维基百科抽取 | DBpedia |
| 维基数据（Wikidata） | 维基媒体基金会（Wikimedia Foundation） | * 实体、语义类、属性、关系，与维基百科紧密结合； * 人工协同编辑 | Wikipedia |
| Wolfram Alpha | 沃尔夫勒姆公司(Wolfram Research) | * 实体、语义类、属性、关系，知识计算； * 部分知识来自于Mathematica；其他知识来自于各个垂直网站 | Apple Siri |
| Bing Satori | Microsoft | * 实体、语义类、属性、关系，知识计算； * 自动加人工 | Bing Search Engine Microsoft Contana |
| YAGO | 马克思 普朗克研究所 | * 从维基百科、WordNet、GeoNames抽取信息 | YAGO |
| Facebook Social Graph | Facebook | * Facebook社交网络数据 | Social Graph Search |
| 百度知识图谱 | 百度 | * 搜索结构化数据 | 百度搜索 |
| 搜狗知立方 | 搜狗 | * 搜索结构化数据 | 搜狗搜索 |
| ImageNet | 斯坦福大学 | * 搜索引擎 * 亚马逊AMT | 计算机视觉相关应用 |

知识图谱是为了适应新的网络信息环境而产生的一种语义知识组织和服务的方法，通过把用户查询的关键词映射到语义知识库的概念上，使计算机能够理解人类的语言交流模式，从而更加智能地反馈给用户需要的答案。知识库是知识图谱的核心，采用某种知识表示方式来存储管理互相关联的知识片集合[9]，它必须包含丰富的数据，数据来源于原有的关系型数据库、LOD中的部分关联数据集、领域本体、用户数据、从半结构化和非结构的数据内容中抽取出的理论知识、事实数据、启发式知识等。知识库是服从于本体控制的知识单元的载体，覆盖了各种概念、实例、属性、关系等要素，并保持高效率地更新，以便随时满足用户的知识需求。以谷歌知识图谱为例，它在2012年5月发布时已包含5亿多的对象实体和关于这些实体的超过35亿的事实关系，仅仅6个月后，实体数量增长到5.7亿，事实关系增长到180亿，到目前为止，还在不断地更新扩展。

1.3 本文章节安排

本文的第一部分将沿着前面叙述，进一步剖析知识图谱的定义与架构；第二部分将以开放链接知识库、垂直行业知识这两类主要的知识库类型为代表，简要介绍其中的几个知名度较高的大规模知识库；第三部分将以知识图谱中的关键技术为重点，详细阐述知识获取、知识表示、知识融合、知识推理技术中的相关研究以及若干技术细节；第四部分将介绍知识图谱在实际生活垂直行业中的典型应用；第五部分将介绍知识图谱所面临的一些困难与挑战；

# 第二章 知识图谱的定义与架构

2.1 知识图谱的定义

辛格博士在介绍知识图谱时提到的： “世界不是由字符串组成，而是由实体组成（The world is not made of strings , but is made of things）。” 知识图谱旨在描述真实世界中存在的各种实体或概念。本质上，知识图谱是一种揭示实体之间关系的结构化的语义知识库，可以对现实世界的事物及其相互关系进行形式化地描述。其构成一张巨大的语义网络图，节点表示实体或概念，边则由属性或关系构成。其中，每个实体或概念可以用一个全局唯一确定的标识符来标识；每个属性-值对（attribute-value pair,AVP）用来刻画实体的内在特性；关系则用来连接两个实体，刻画它们之间的关联。知识图谱亦可被看作是一张巨大的图，图中的节点表示实体或概念，图中的边由属性或关系构成。

如下图1的知识图谱例子所示，中国是一个实体，北京是一个实体，中国-首都-北京 是一个（实体-关系-实体）的三元组样例北京是一个实体 ，人口是一种属性2069.3万是属性值。北京-人口-2069.3万构成一个（实体-属性-属性值）的三元组样例。



图1 知识图谱示例

通过知识图谱，可以实现Web从网页链接向概念链接转变，支持用户按主题而不是字符串检索，从而真正实现语义检索。基于知识图谱的搜索引擎，能够以图形方式向用户反馈结构化的知识，用户不必浏览大量网页，就可以准确定位和深度获取知识。

知识图谱本身是一个具有属性的实体通过关系链接而成的网状知识库。从图的角度来看，知识图谱在本质上是一种概念网络，其中的节点表示物理世界的实体（或概念），而实体间的各种语义关系则构成网络中的边。由此，知识图谱是对物理世界的一种符号表达。

知识图谱的研究价值在于，它是构建在当前Web基础之上的一层覆盖网络（overlay network），借助知识图谱，能够在Web网页之上建立概念间的链接关系，从而以最小的代价将互联网中积累的信息组织起来，成为可以被利用的知识。

知识图谱的应用价值在于，它能够改变现有的信息检索方式，一方面通过推理实现概念检索（相对于现有的字符串模糊匹配方式而言）；另一方面以图形化方式向用户展示经过分类整理的结构化知识，从而使人们从人工过滤网页寻找答案的模式中解脱出来。通过知识图谱，可以实现Web从网页链接向概念链接转变，支持用户按主题而不是字符串检索，=从而真正实现语义检索。基于知识图谱的搜索引擎，能够以图形方式向用户反馈结构化的知识，用户不必浏览大量网页，就可以准确定位和深度获取知识。

2.2 知识图谱的架构

知识图谱的架构，包括知识图谱自身的逻辑架构以及构建知识图谱所采用的技术架构。

2.2.1 知识图谱的逻辑架构

从逻辑上将知识图谱划分为２个层次：数据层和模式层。知识图谱的数据层主要是由一系列的事实（fact）组成，知识以事实为单位存储在图数据库。例如谷歌的Graphd和微软的Trinity都是典型的图数据库。如果以“实体－关系－实体”或者“实体－属性－性值”三元组作为事实的基本表达方式，则存储在图数据库中的所有数据将构成庞大的实体关系网络，形成知识的“图谱”。模式层构建在数据层之上，是知识图谱的核心。在模式层存储的是经过提炼的知识，通常采用本体库来管理知识图谱的模式层，借助本体库对公理、规则和约束条件的支持能力来规范实体、关系以及实体的类型和属性等对象之间的联系。本体库在知识图谱中的地位相当于知识库的模具，通过本体库而形成的知识库不仅层次结构较强，并且冗余程度较小。

2.2.2 知识图谱的技术架构

知识图谱的技术架构是其指构建模式结构，图2给出了知识图谱技术的整体架构，其中虚线框内的部分为知识图谱的构建过程，同时也是知识图谱更新的过程。如图2所示，知识图谱的构建过程是从原始数据出发，采用一系列自动或半自动的技术手段，从原始数据中提取出知识要素（即事实），并将其存入知识库的数据层和模式层的过程。这是一个迭代更新的过程，根据知识获取的逻辑，每一轮迭代包含3个阶段：知识抽取、知识表示、知识融合以及知识推理。



图2 知识图谱的技术架构图

知识图谱有自顶向下和自底向上２种构建方式。所谓自顶向下构建是指借助百科类网站等结构化数据源，从高质量数据中提取本体和模式信息，加入到知识库中；所谓自底向上构建，则是借助一定的技术手段，从公开采集的数据中提取出资源模式，选择其中置信度较高的新模式，经人工审核之后，加入到知识库中。在知识图谱技术发展初期，多数参与企业和科研机构都是采用自顶向下的方式构建基础知识库，例如，Freebase项目就是采用维基百科作为主要数据来源。随着自动知识抽取与加工技术的不断成熟，目前的知识图谱大多采用自底向上的方式构建，其中最具影响力的例子包括谷歌的Knowledge Vault和微软的Bing Satori知识库，都是以公开采集的海量网页数据为数据源，通过自动抽取资源的方式来构建、丰富和完善现有的知识库。

因此，本文主要介绍自底向上的知识图谱构建技术，按照知识获取的过程分为4个层次：知识抽取、知识表示、知识融合以及知识推理。

第三章 知识图谱关键技术

采用自底向上的方式构建知识图谱需要不断进行知识抽取、知识表示、知识融合、知识推理等操作，本文将以知识抽取、知识表示、知识融合以及知识推理技术为重点，详细说明其中的相关研究。通过知识抽取技术，可以从一些公开的半结构化、非结构化的数据中提取出实体、关系、属性等知识要素。通过知识融合，可消除实体、关系、属性等指称项与事实对象之间的歧义，形成高质量的知识库。知识推理则是在已有的知识库基础上进一步挖掘隐含的知识，从而丰富、扩展知识库。分布式的知识表示形成的综合向量对知识库的构建、推理、融合以及应用均具有重要的意义。

3.1知识抽取

由于数据源可以是有结构、半结构、无结构的，因而知识抽取的关键问题是如何从这些异构数据源中自动抽取信息得到实体、属性、关系等知识要素，如下图3所示。进而通过知识表示技术将这些知识要素表达为语义空间上的向量，为上层模式层的构建奠定基础。由于要抽取的知识要素的不同，将知识抽取划分为实体抽取、属性抽取、关系抽取，以采用不同的技术手段提取不同的知识要素。

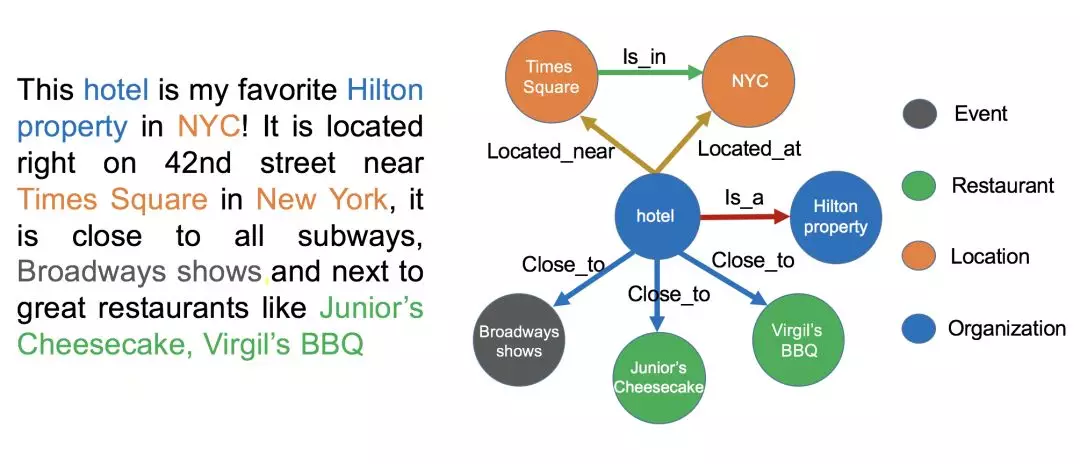


图3 知识图谱

3.1.1 实体抽取

实体抽取，也称为命名实体识别，是指从文本数据集中自动识别出命名实体。由于实体是知识图谱中的最基本元素，其抽取的完整性、准确率、召回率等将直接影响到知识库的质量。因此，实体抽取是知识抽取中最为基础的一步。

命名实体识别是自然语言处理的一项基本技术，它不但是信息抽取的基础，也是信息检索、机器翻译、组块分析、问答系统等多种自然语言处理技术的重要基础，因而被广泛研究。命名实体识别任务分为两步：(1)发现命名实体，即判断一个文本串是否代表一个命名实体；(2)标注命名实体，即将发现的命名实体标注为某一种具体类型。

文献[11]将实体抽取的方法分为3种：基于规则与词典的方法、基于统计机器学习的方法以及面向开放域的抽取方法。基于规则的方法通常需要为目标实体编写模板，然后在原始语料中进行匹配；基于统计机器学习的方法主要是通过机器学习的方法对原始语料进行训练，然后再利用训练好的模型去识别实体；面向开放域的抽取将是面向海量的Web语料。  
1) 基于规则与词典的实体抽取方法  
 早期的实体抽取方法研究对象是单一领域（如特定行业或特定业务），关注如何识别出文本中的人名、地名等专有名词和有意义的时间等实体信[12]。由于研究对象的特定性，因而普遍采用基于规则与词典的方法，在识别的过程中加入词法规则、语法规则甚至语义规则来提高命名实体识别的质量。一般都是由人工添加规则，或者在人工添加的基础上再从有限的训练语料库中得到规则。1991年Rau[采用启发式算法与人工编写规则相结合的方法，首次实现了从文本中自动抽取公司名称的实体抽取原型系统。然而，基于规则的方法具有明显的局限性，规则编写人员往往需要有丰富的语言知识和深厚的计算语言学背景，带来了极高的人力成本。而且系统缺乏适应性，由于规则的制订往往要依赖于具体的语言、领域和文本格式，系统若要进行移植，则需要做很大的改动。

2) 基于统计机器学习的实体抽取方法  
 随着统计机器学习技术的成熟，研究者尝试将机器学习中的监督学习算法用于命名实体的识别问题上，研究者将命名实体识别任务转化为序列标注的问题，利用人工标注的语料进行训练，标注语料时不需要广博的语言学知识，并且可以在较短时间内完成，因此这类系统在移植到新的领域时可以不做或少做改。具体来说可采用隐马尔可夫模型、最大熵、支持向量机、条件随机场等技术。

隐马尔可夫模型是一种重要的统计自然语言模型，广泛应用于语音识别、词性及NE识别等领域。Zhou[13]利用一个基于HMM的组块标注器进行NE识别，识别型包括名字、时间、数字短语。作者对HMM的传统公式做了改进使之便于融合更的信息，包括四种内部和外部的信息：1)词本身的简单特征；2)重要的触发词的内语义特征；3)内部词典特征；4)外部宏观上下文特征。但是它融合多种信息的能力不强，同时它不能利用下文的信息，对于复杂NE的识别将遇到困难。

最大熵模型在这方面就优于基于HMM的方法。Borthwich[14]最早将该方法引入到英文NE识别问题中，在他的系统中利用了二元特征、词汇特征、段落特征、字典信息等。最大熵方法可以结合更丰富的特征，同时体现出每个特征的重要性。但是它的计算比较复杂且系统开销比较大。

支持向量机是上世纪90年代中期发展起来的一种机器学习方法，它建立在统计学习理论的结构风险最小化原则上，其主要思想是针对两类分类问题，在高维空间中寻找一个超平面作为两类的分割，以保证最小的分类错误率。支持向量机的优点有：1)具有较好的推广能力，即能在训练样本较少的情况下得到较好的效果；2)高维处理能力，即支持向量机的学习误差不依赖于特性空间的维数，不会出现其它机器学习方法的“过学习”现象；3)强大的非线性处理能力。

决策树是一种基于实例的归纳学习算法。Bechet采用决策树的方法识别专有名词，系统可以自动从名词短语包括专有名词的集合中选择最容易区分的特征，然后利用决策树方法学习，最终可用来分类未知的专有名词。决策树方法的问题在于如何选择最好的特征，否则不好的特征将影响最终的结果。另外，决策的每个叶子上的事件太少，不能有效的估计每个NE的概率，且不易融合太多的信息。

条件随机域方法，它是一种无向图模型，可用于最大化条件概率。常用的特殊图结构是线性链，与一个有限状态机相关。一个基于CRF的识别系统是Zhou的基于层叠条件随机场(CCRFs)模型的中文机构名识别模型。 在CCRFs模型中，低层的条件随机场模型仅以观察值为条件，用于人名、地名等简单命名实体的识别，识别的结果再传递到高层模型，这样高层模型的输入变量将不仅包含观察值，而且包含了来自低层模型的识别结果，从而为高层条件随机场模型对复杂机构名的识别提供了决策支持，最后采用约束的前向后向算法对识别的结果进行可信度计算。

基于条件随机域方法的命名实体识别需要人工提取针对不同实体的特征，因此对于不同的实体类型，研究者进行大量实验以找寻合适的特征。近年来，随着硬件能力的发展以及词的分布式表示（word embedding）的出现，神经网络成为可以有效处理许多NLP任务的模型。这类方法对于序列标注任务（如CWS、POS、NER）的处理方式是类似的，将token从离散one-hot表示映射到低维空间中成为稠密的embedding，随后将句子的embedding序列输入到RNN中，用神经网络自动提取特征，Softmax来预测每个token的标签。这种方法使得模型的训练成为一个端到端的整体过程，而非传统的pipeline，不依赖特征工程，是一种数据驱动的方法；但网络变种多、对参数设置依赖大，模型可解释性差。

此外，这种方法的一个缺点是对每个token打标签的过程中是独立的分类，不能直接利用上文已经预测的标签（只能靠隐状态传递上文信息），进而导致预测出的标签序列可能是非法的，例如标签B-PER后面是不可能紧跟着I-LOC的，但Softmax不会利用到这个信息。为了加入标注的特征信息，学术界提出了BiLstm-CRF，首先通过双向神经网络学习特征，最后一层加入CRF，使得人工干预标注序列来最大化最终得到标注序列的概率。如下图4所示。

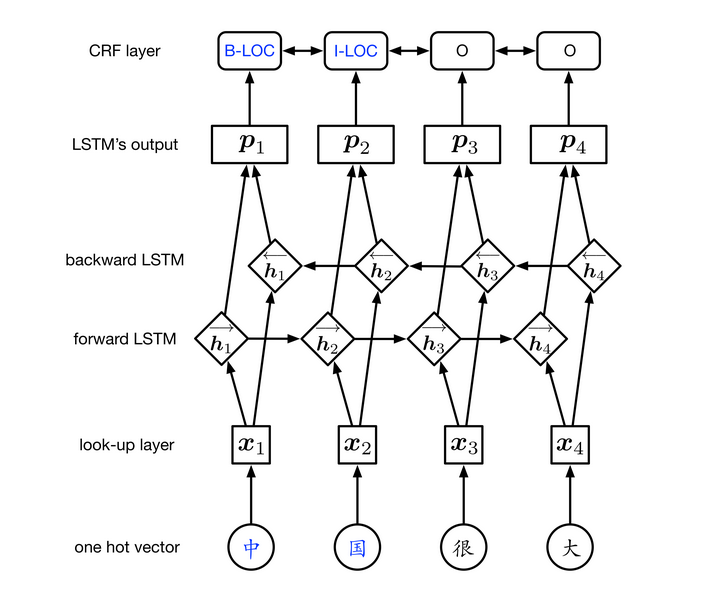


图4 BiLstm-CRF模型

(3) 面向开放域的抽取方法

针对如何从少量实体实例中自动发现具有区分力的模式，进而扩展到海量文本去给实体做分类与聚类的问题，文献[15]提出了一种通过迭代方式扩展实体语料库的解决方案，其基本思想是通过少量的实体实例建立特征模型，再通过该模型应用于新的数据集得到新的命名实体。文献[16]提出了一种基于无监督学习的开放域聚类算法，其基本思想是基于已知实体的语义特征去搜索日志中识别出命名的实体，然后进行聚类。

3.1.2 关系抽取

文本语料经过实体抽取，得到的是一系列离散的命名实体，为了得到语义信息，还需要从相关语料中提取出实体之间的关联关系，通过关系将实体联系起来，才能够形成网状的知识结构。  
 早期的关系抽取主要是通过人工构造语义规则以及模板的方法识别实体关系，这种方法有２点明显的不足：１）要求制定规则的人具有良好的语言学造诣，并且对特定领域有深入的理解和认知；２）规则制定工作量大，难以适应丰富的语言表达风格，且难以拓展到其他领域。为此学术界开始尝试采用统计机器学习方法，通过对实体间关系的模式进行建模，替代预定义的语法和语义规则。例如 kambhat[17]等人利用自然语言中的词法、句法以及语义特征进行实体关系建模，通过最大熵方法成功地实现了不借助规则硬编码的实体关系抽取。

面向开放域的关系抽取技术直接利用语料中的关系词汇对实体关系进行建模，因此不需要预先指定关系的分类，这是一个很大的进步，例如，Wu等人[18]在OIE的基础上，发布了面向开放域信息抽取的WOE系统，该系统能够利用维 基 百 科 网 页 信 息框提供的属性信息，自动构造实体关系训练集，性能优于早期的TextRunner系统，这项工作也为批量构造高质量的训练语料提供了新的思路。

由于当前的面向开放域的关系抽取方法在准确率和召回率等综合性能指标方面与面向封闭领域的传统方法相比仍有一定的差距，因此有部分学者开始尝试将两者的优势结合起来。例如 Banko等人[19]提出了一 种 基 于 条 件 随 机 场 的 关 系 抽 取 模 型 ，当目标数据集中拥有的关系数量不大，而且有预先定义好的实体关系分类模型可用的情况下，采用传统的机器学习算法进行关系抽取，而对于没有预先定义好的实体关系模型或者关系数量过多的情况，则采用开放域关系抽取方法。微软公司人立方项目所采用的 StaSnowball模型也是基于这种策略实现其关系抽取功能[20]。

当前流行的 OIE 系统在关系抽取方面存在 ２个主要问题，1)当前研究的重点是如何提高二元实体间关系（三元组模式）的抽取准确率和召回率，很少考虑到在现实生活中普遍存在的高阶多元实体关系；2)所采用的研究方法大多只关注发掘词汇或词组之间的关系模式，而无法实现对隐含语义关系的抽取。对此，学 术 界 有 着 清 醒 的 认 识，例如 Alan 等人[21]采用 Ｎ 元关系模型对OIE系统进行改进，提出了 KRAKEN 模型，能够有效提高 OIE 系统对多元实 体 关 系 的 识 别 能 力。在 隐 含 关 系 识 别 方 面，McCallum[22]提出采用后期关系推理的方法，提高OIE系统对隐含实体关系的发现能力。这些工作都是该领域值得重视的研究动向，然而在 OIE 关系抽取研究领域，要实现算法性能由量变到质变的飞跃，还需要一段时间的积累。

3.1.3 属性抽取

属性抽取主要是针对实体而言的，通过属性可形成对实体的完整勾画。由于实体的属性可以看成是实体与属性值之间的一种名称性关系，因此可以将实体属性的抽取问题转换为关系抽取问题。文献[23,24]提出的基于规则与启发式算法的属性抽取方法能够从Wikipedia及WordNet的半结构化网页中自动抽取相应的属性名称与属性值，还可扩展为一套本体知识库。实验表明：该算法的抽取准确率可达到95%。  
 大量的属性数据主要存在于半结构化、非结构化的大规模开放域数据集中。抽取这些属性的方法，一种是将上述从百科网站上抽取的结构化数据作为可用于属性抽取的训练集，然后再将该模型应用于开放域中的实体属性抽取[25]；另一种是根据实体属性与属性值之间的关系模式，直接从开放域数据集上抽取属性。但是由于属性值附近普遍存在一些限定属性值含义的属性名等，所以该抽取方法的准确率并不高[26]。

3.2 知识表示

在网络表示形式下,人们需要设计专门的图算法存储和利用知识库，存在费时费力的缺点，并受到数据稀疏问题的困扰。最近，以深度学习为代表的表示学习技术受到广泛关注。表示学习旨在将研究对象的语义信息表示为稠密低维实值向量,知识表示学习则面向知识库中的实体和关系进行表示学习。该技术可以在低维空间中高效计算实体和关系的语义联系有效解决数据稀疏问题，使知识获取、融合和推理的性能得到显著提升。

3.2.1表示学习的基本概念

表示学习的目标是，通过机器学习将研究对象的语义信息表示为稠密低维实值向量。本文用黑斜体表示研究对象所对应的向量。以知识库中的实体e和关系r为例，我们将表示学习得到的向量表示为Z和Z。在该向量空间中，我们可以通过欧氏距离或余弦距离等方式，计算任意2个对象之间的语义相似度。

实际上，在表示学习之外，有更简单的数据表示方案，即独热表示。该方案也将研究对象表示为向量，只是该向量只有某一维非零，其他维度上的值均为0。显而易见，为了将不同对象区分开，有多少个不同的对象，独热表示向量就有多长。独热表示是信息检索和搜索引擎中广泛使用的词袋模型的基础。以中文为例，假如网页中共有W个不同的词，词袋模型中的每个词都被表示为一个w维的独热表示向量。在此基础上，词袋模型将每个文档表示为一个w维向量，每一维表示对应的词在该文档中的重要性。与表示学习相比，独热表示无需学习过程，简单高效，在信息检索和自然语言处理中得到广泛应用。但是独热表示的缺点也非常明显。独热表示方案假设所有对象都是相互独立的。也就是说，在独热表示空间中，所有对象的向量都是相互正交的，通过余弦距离或欧氏距离计算的语义相似度均为O。这显然是不符合实际情况的，会丢失大量有用信息。与独热表示相比，表示学习的向量维度较低，有助于提高计算效率，同时能够充分利用对象间的语义信息，从而有效缓解数据稀疏问题。

3.2.2知识表示学习的主要方法

知识表示学习的几个代表模型，包括距离模型、单层神经网络模型、能量模型、双线性模型、张量神经网络模型、矩阵分解模型和翻译模型等。

1)距离模型  
 文献[27]提出了知识库中实体以及关系的结构化表示方法，其基本思想是：首先将实体用向量进行表示，然后通过关系矩阵将实体投影到与实体向量同一维度的向量空间中，最后通过计算投影向量之间的距离来判断实体间已存在的关系的置信度。由于距离模型中的关系矩阵是两个不同的矩阵，故实体间的协同性较差，这也是该模型本身的主要缺陷。

2)单层神经网络模型  
 文献[28]针对上述提到的距离模型中的缺陷，提出了采用单层神经网络的非线性模型，模型为知识库中每个三元组(,,)*hrt*定义了以下形式的评价函数：

  
 式中，为关系*r*的向量化表示；*g*()为tanh函数；、∈是通过关系*r*定义的两个矩阵。单层神经网络模型的非线性操作虽然能够进一步刻画实体在关系下的语义相关性，但在计算开销上却大大增加。  
3)双线性模型  
 双线性模型又叫隐变量模型，由文献[29-30]首先提出。模型为知识库中每个三元组(h,r,t)定义的评价函数具有如下形式：  
 

式中，是通过关系*r*定义的双线性变换矩阵；、∈是三元组中头实体与尾实体的向量化表示。  
 双线性模型主要是通过基于实体间关系的双线性变换来刻画实体在关系下的语义相关性。模型不仅形式简单、易于计算，而且还能够有效刻画实体间的协同性[31]。基于上述工作，文献[32]尝试将双线性变换矩阵***M***变换为对角矩阵，提出了DISTMULT模型，不仅简化了计算的复杂度，并且实验效果得到了显著提升。

4)矩阵分解模型  
通过矩阵分解的方式可得到低维的向量表示，故不少研究者提出可采用该方式进行知识表示学习，其中的典型代表是文献[33]提出的RESACL模型。在RESCAL模型中，知识库中的三元组(h,r,t)集合被表示为一个三阶张量，如果该三元组存在，张量中对应位置的元素被置1，否则置为0。通过张量分解算法，可将张量中每个三元组(h,r,t)对应的张量值分解为双线性模型中的知识表示形式，并使尽量小。

5)翻译模型  
文献[34]受到平移不变现象的启发，提出了TransE模型，即将知识库中实体之间的关系看成是从实体间的某种平移，并用向量表示。关系可以看作是从头实体向量到尾实体向量的翻译。对于知识库中的每个三元组(h,r,t)，TransE都希望满足以下关系：，其损失函数为：，即向量和的或距离。该模型的参数较少，计算的复杂度显著降低。与此同时，TransE模型在大规模稀疏知识库上也同样具有较好的性能与可扩展性。  
3.2.3复杂关系模型  
知识库中的实体关系类型也可分为1-to-1、1-to-*N*、*N*-to-1、*N*-to-*N*4种类型，而复杂关系主要指的是1-to-*N*、*N*-to-1、*N*-to-*N*的3种关系类型。由于TransE模型不能用在处理复杂关系上[35]，一系列基于它的扩展模型纷纷被提出，下面将着重介绍其中的几项代表性工作。  
1)TransH模型  
文献[36]提出的TransH模型尝试通过不同的形式表示不同关系中的实体结构，对于同一个实体而言，它在不同的关系下也扮演着不同的角色。模型首先通过关系向量与其正交的法向量选取某一个超平面*F*，然后将头实体向量和尾实体向量沿法向量的方向投影到*F*，最后计算损失函数。TransH使不同的实体在不同的关系下拥有了不同的表示形式，但由于实体向量被投影到了关系的语义空间中，故它们具有相同的维度。  
2)TransR模型  
由于实体、关系是不同的对象，不同的关系所关注的实体的属性也不尽相同，将它们映射到同一个语义空间，在一定程度上就限制了模型的表达能力。所以，文献[37]提出了TransR模型。模型首先将知识库中的每个三元组(h,r,t)的头实体与尾实体向关系空间中投影，然后希望满足的关系，最后计算损失函数。  
文献[37]提出的CTransR模型认为关系还可做更细致的划分，这将有利于提高实体与关系的语义联系。在CTransR模型中，通过对关系*r*对应的头实体、尾实体向量的差值进行聚类，可将*r*划分为若干个子关系。  
3)TransD模型  
考虑到在知识库的三元组中，头实体和尾实体表示的含义、类型以及属性可能有较大差异，之前的TransR模型使它们被同一个投影矩阵进行映射，在一定程度上就限制了模型的表达能力。除此之外，将实体映射到关系空间体现的是从实体到关系的语义联系，而TransR模型中提出的投影矩阵仅考虑了不同的关系类型，而忽视了实体与关系之间的交互。因此，文献[38]提出了TransD模型，模型分别定义了头实体与尾实体在关系空间上的投影矩阵。  
4)TransG模型  
文献[39]提出的TransG模型认为一种关系可能对应多种语义，而每一种语义都可以用一个高斯分布表示。TransG模型考虑到了关系*r*的不同语义，使用高斯混合模型来描述知识库中每个三元组(h,r,t)的头实体与尾实体之间的关系，具有较高的实体区分度。

3.3 知识融合

由于知识图谱中的知识来源广泛，存在知识质量良莠不齐、来自不同数据源的知识重复、知识间的关联不够明确等问题，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[[40](http://www.xml-data.org/dzkj-nature/html/201645589.htm#b72)]，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[[41](http://www.xml-data.org/dzkj-nature/html/201645589.htm#b73)]，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。

3.3.1实体对齐

实体对齐(entity alignment)也称为实体匹配(entity matching)或实体解析(entity resolution)，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。

在大数据的环境下，受知识库规模的影响，在进行知识库实体对齐时，主要会面临以下3个方面的挑战[[42](http://www.xml-data.org/dzkj-nature/html/201645589.htm#b74)]：1) 计算复杂度。匹配算法的计算复杂度会随知识库的规模呈二次增长，难以接受；2) 数据质量。由于不同知识库的构建目的与方式有所不同，可能存在知识质量良莠不齐、相似重复数据、孤立数据、数据时间粒度不一致等问题[[43](http://www.xml-data.org/dzkj-nature/html/201645589.htm#b75)]；3) 先验训练数据。在大规模知识库中想要获得这种先验数据却非常困难。通常情况下，需要研究者手工构造先验训练数据。

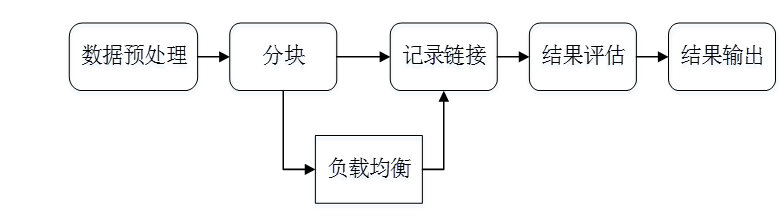


图5 知识融合技术流程

图5所示为知识融合的基本技术流程。其中数据预处理也就是前面所说的知识抽取与知识表示阶段，这里主要介绍一下记录链接步骤。

假设两个实体的记录x 和y， x和y在第i个属性上的值是xi,yixi,yi, 那么通过如下两步进行记录连接：

* 属性相似度： 综合单个属性相似度得到属性相似度向量:   
  [sim(x1,y1),sim(x2,y2),…,sim(xN,yN)][sim(x1,y1),sim(x2,y2),…,sim(xN,yN)]
* 实体相似度： 根据属性相似度向量得到一个实体的相似度。

属性相似度的计算有多种方法，常用的有编辑距离、集合相似度计算、基于向量的相似度计算等。

1）编辑距离法

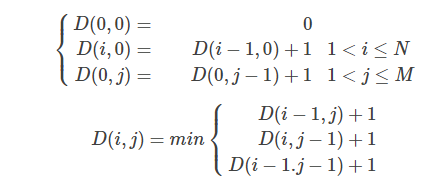
a. Levenshtein Distance

Levenshtein 距离，即最小编辑距离，目的是用最少的编辑操作将一个字符串转换成另一个.举个例子,计算Lvensshtain 与 Levenshtein 间的编辑距离:

Lvensshtain→insert"e"→LevensshtainLvensshtain→insert"e"→Levensshtain   
Levenshtain→delete"s"→LevenshtainLevenshtain→delete"s"→Levenshtain   
Levenshtain→sub"a"to"e"→LevenshteinLevenshtain→sub"a"to"e"→Levenshtein

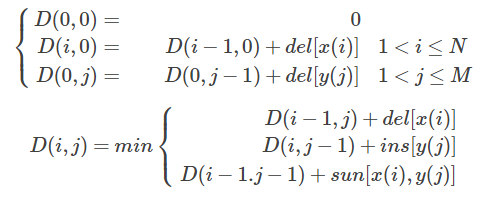
上述将Lvensshtain转换为Levenshtein ，总共操作3次，编辑距离也就是3。

Levenstein Distance 是典型的动态规划问题，可以通过动态规划算法计算，具体公式如下：



b. Wagner and Fisher Distance

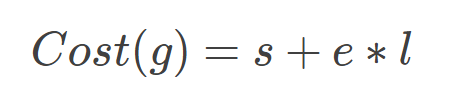
它是Levenshtein距离的一个扩展，将这个模型中的编辑操作的代价赋予了不同的权重，如下：



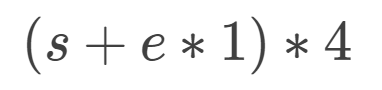
其中del、ins和sub分别是删除、插入和替换的代价。

c. Edit Distance with affine gaps

在上面的两种算法基础上，引入了gap的概念，将上述的插入、删除和替换操作用gap opening 和gap extension代替，编辑操作的代价也就表示为：



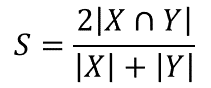
其中s 是open extension的代价， e是extend gap的代价，l是gap的长度。如计算 Lvensshtain 与 Levenshtein间的距离，首先将两个单词首尾对齐，将对应缺少的部分视为gap，如下图中上面和下面单词相比少了第一个e和倒数第三个的e，这是两个gap。下面的单词与上面的比则少了一个s和a，这又是两个gap。加一起一共4个gap，每个长度为1.因此编辑距离为:



2）集合相似度

a. Dice系数

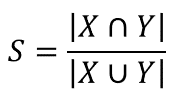
Dice系数用于度量两个集合的相似性，因为可以把字符串理解为一种集合，因此Dice距离也会用于度量字符串的相似性，Dice系数定义如下：



以Lvensshtain 和 Levenshtein为例，两者的相似度为 2 \* 9 / (11+11) = 0.82。

b. Jaccard系数

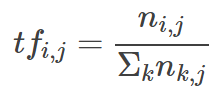
Jaccard 系数适合处理短文本的相似度，定义如下：

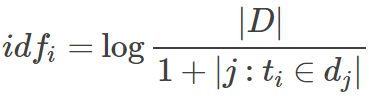


可以看出与Dice系数的定义比较相似。两种方法,将文本转换为集合,除了可以用符号分格单词外,还可以考虑用n-gram分割单词,用n-gram分割句子等来构建集合,计算相似度。

c. TF-IDF 基于向量的相似度

TF-IDF主要用来评估某个字或者用某个词对一个文档的重要程度。其中:





举个例子，比如某个语料库中有5万篇文章,含有“健康”的有2万篇,现有一篇文章,共1000个词,‘健康’出现30次,则sim TF-IDF = 30/1000 \* log(50000/(20000+1)) = 0.012。

而计算实体相似度可从三大方面入手，即聚合、聚类和表示学习。其中：

* 聚合：加权平均、手动制定规则、分类器
* 聚类：层次聚类、相关性聚类、Canopy + K-means
* 表示学习

1）聚合

a. 加权平均法

对相似度得分向量的各个分量进行加权求和，得到最终的实体相似度：



b. 手动制定规则法

给每一个相似度向量的分量设置一个阈值，若超过该阈值则将两实体相连:



c. 分类器方法

对于分类器等机器学习方法，最大的问题是如何生成训练集合，对于此可采用无监督/半监督训练，如EM、生成模型等。或主动学习如众包等方案。

2）聚类

聚类又可分为层次聚类、Canopy + K-means等。

1. 层次聚类

层次聚类 (Hierarchical Clustering) 通过计算不同类别数据点之间的相似度对在不同的层次的数据进行划分,最终形成树状的聚类结构。

底层的原始数据可以通过相似度函数计算，类之间的相似度有如下三种算法：

SL(Single Linkage)算法： SL算法又称为最邻近算法 (nearest-neighbor),是用两个类数据点中距离最近的两个数据点间的相似度作为这两个类的距离。

CL (Complete Linkage)算法: 与SL不同的是取两个类中距离最远的两个点的相似度作为两个类的相似度。

AL (Average Linkage) 算法: 用两个类中所有点之间相似度的均值作为类间相似度。

1. Canopy + K-means

与K-means不同,Canopy聚类最大的特点是不需要事先指定k值 (即clustering的个数),因此具有很大的实际应用价值,经常将Canopy和K-means配合使用。

用图形表达流程如图6所示：

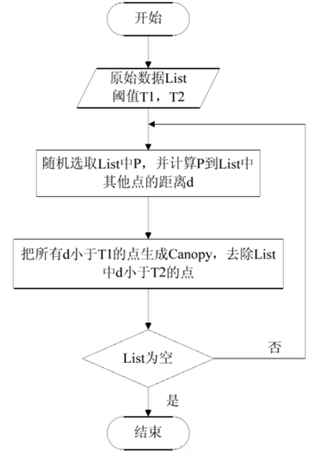


图6 Canopy流程

文字表述为：初始时有一个大的list，其中list中每个点都是一个canopy，设置阈值T1，T2。随机玄奇List中的点P，并计算list中其他的点到点P的距离d，把所有距离d小于T1的点生成Canopy，去除list中d小于T2的点。如此往复这个过程就得到了聚类结果。生成Canopy的过程就像以T2为中心扣下来一块，然后剩下的环就是Canopy。这样一块一块的扣就知道最终list为空。如图7所示。

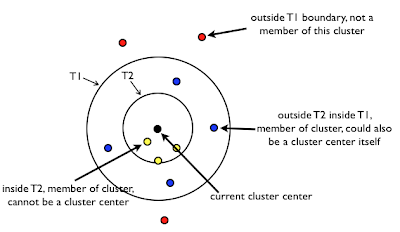


图7 Canopy图解

3.3.2 知识加工

通过实体对齐，可以得到一系列的基本事实表达或初步的本体雏形，然而事实并不等于知识，它只是知识的基本单位。要形成高质量的知识，还需要经过知识加工的过程，从层次上形成一个大规模的知识体系，统一对知识进行管理。知识加工主要包括本体构建与质量评估两方面的内容。

1) 本体构建

本体是同一领域内不同主体之间进行交流、连通的语义基础[[44](http://www.xml-data.org/dzkj-nature/html/201645589.htm#b95)]，其主要呈现树状结构，相邻的层次节点或概念之间具有严格的“IsA”关系，有利于进行约束、推理等，却不利于表达概念的多样性。本体在知识图谱中的地位相当于知识库的模具，通过本体库而形成的知识库不仅层次结构较强，并且冗余程度较小[[45](http://www.xml-data.org/dzkj-nature/html/201645589.htm#b96)]。

本体可通过人工编辑的方式手动构建，也可通过数据驱动自动构建，然后再经质量评估方法与人工审核相结合的方式加以修正与确认[[10](http://www.xml-data.org/dzkj-nature/html/201645589.htm#b12)]。在海量的实体数据面前，人工编辑构建的方式工作量极其巨大，故当前主流的本体库产品，都是面向特定领域，采用自动构建技术而逐步扩展形成的。例如Microsoft的Probase本体库就是采用数据驱动的方法，利用机器学习算法从网页文本中抽取概念间的“IsA”关系，然后合并形成概念层次结构。目前，Probase所包含的概念总数已达到千万级别，准确率高达92.8%，是目前为止包含概念数量最多，同时也是概念可信程度最高的知识库[[46](http://www.xml-data.org/dzkj-nature/html/201645589.htm#b97)]。

数据驱动的本体自动构建过程主要可分为以下3个阶段[[98](http://www.xml-data.org/dzkj-nature/html/201645589.htm#b98)]：① 纵向概念间的并列关系计算。通过计算任意2个实体间并列关系的相似度，可辨析它们在语义层面是否属于同一个概念。计算方法主要包括模式匹配与分布相似度两种[[10](http://www.xml-data.org/dzkj-nature/html/201645589.htm#b12), [47](http://www.xml-data.org/dzkj-nature/html/201645589.htm#b99)]。② 实体上下位关系抽取。上下位关系抽取方法包括基于语法的抽取与基于语义的抽取两种方式，例如目前主流的信息抽取系统KnowltAll[[48](http://www.xml-data.org/dzkj-nature/html/201645589.htm#b38)]、TextRunner[[49](http://www.xml-data.org/dzkj-nature/html/201645589.htm#b37)]、NELL[[50](http://www.xml-data.org/dzkj-nature/html/201645589.htm#b100)]等，都可以在语法层面抽取实体的上下位关系，而Probase则是采用基于语义的抽取模式[[51](http://www.xml-data.org/dzkj-nature/html/201645589.htm#b101)]。③ 本体生成。对各层次得到的概念进行聚类，并为每一类的实体指定1个或多个公共上位词。文献[[52](http://www.xml-data.org/dzkj-nature/html/201645589.htm#b102)]基于主题层次聚类的方法构建了本体结构。与此同时，为了解决主题模型不适用于短文本的问题，提出了基于单词共现网络的主题聚类与上下位词抽取模型。

2) 质量评估

对知识库的质量评估任务通常是与实体对齐任务一起进行的，其意义在于，可以对知识的可信度进行量化，保留置信度较高的，舍弃置信度较低的，有效确保知识的质量。

文献[[53](http://www.xml-data.org/dzkj-nature/html/201645589.htm#b103)]基于LDIF框架，提出了一种新的知识质量评估方法，用户可根据业务需求来定义质量评估函数，或者通过对多种评估方法的综合考评来确定知识的最终质量评分。例如在对REVERRB系统的信息抽取质量进行评估时，文献[[54](http://www.xml-data.org/dzkj-nature/html/201645589.htm#b104)]采用人工标注的方式对1 000个句子中的实体关系三元组进行了标注，并以此作为训练集，使用logistic回归模型计算抽取结果的置信度。例如Google的Knowledge Vault项目则根据指定数据信息的抽取频率对信息的可信度进行评分，然后利用从可信知识库中得到的先验知识对可信度进行修正。实验结果表明：该方法可以有效地降低对数据信息正误判断的不确定性，提高知识的质量[[55](http://www.xml-data.org/dzkj-nature/html/201645589.htm#b105)]。

3.3.3 知识更新

人类的认知能力、知识储备以及业务需求都会随时间而不断递增。因此，知识图谱的内容也需要与时俱进，不论是通用知识图谱，还是行业知识图谱，它们都需要不断地迭代更新，扩展现有的知识，增加新的知识。

根据知识图谱的逻辑结构，其更新主要包括模式层的更新与数据层的更新。模式层的更新是指本体中元素的更新，包括概念的增加、修改、删除，概念属性的更新以及概念之间上下位关系的更新等。其中，概念属性的更新操作将直接影响到所有直接或间接属性的子概念和实体[[56](http://www.xml-data.org/dzkj-nature/html/201645589.htm#b106)]。通常来说，模式层的增量更新方式消耗资源较少，但是多数情况下是在人工干预的情况下完成的，例如需要人工定义规则，人工处理冲突等。因此，实施起来并不容易[[57](http://www.xml-data.org/dzkj-nature/html/201645589.htm#b107)]。数据层的更新指的是实体元素的更新，包括实体的增加、修改、删除，以及实体的基本信息和属性值。由于数据层的更新一般影响面较小，因此通常以自动的方式完成。

3.3.4 知识融合工具简介

Falcon-AO是一个自动的本体匹配系统,已经成为RDF(S)和OWL所表达的Web本体相匹配的一种实用和流行的选择。编程语言为Java。其结构如下图8所示：

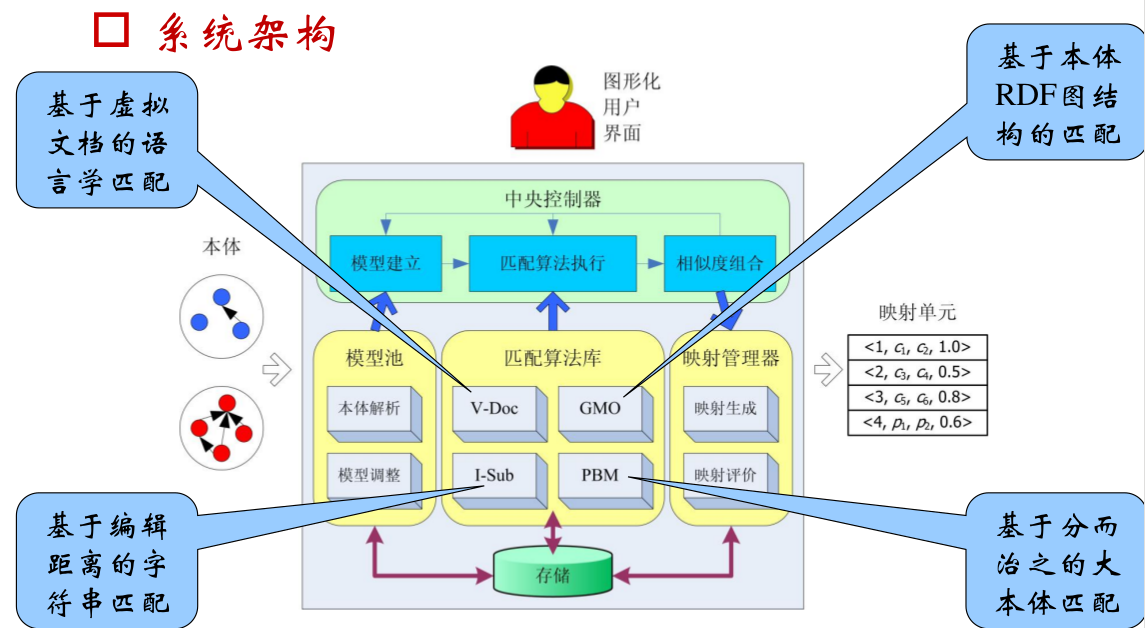


图8 Falcon-AO系统架构

此处主要介绍它的匹配算法库，其余部分可查看官方文档。

匹配算法库包含V-Doc、I-sub、GMO、PBM四个算法。其中V-Doc即基于虚拟文档的语言学匹配，它是将实体及其周围的实体、名词、文本等信息作一个集合形成虚拟文档的形式。这样我们就可以用TD-IDF等算法进行操作。I-Sub是基于编辑距离的字符串匹配，这个前面我们有详细介绍。可以看出，I-Sub和V-Doc都是基于字符串或文本级别的处理。更进一步的就有了GMO，它是对RDF本体的图结构上做的匹配。PBM则基于分而治之的思想做。

计算相似度的组合策略如图9所示:

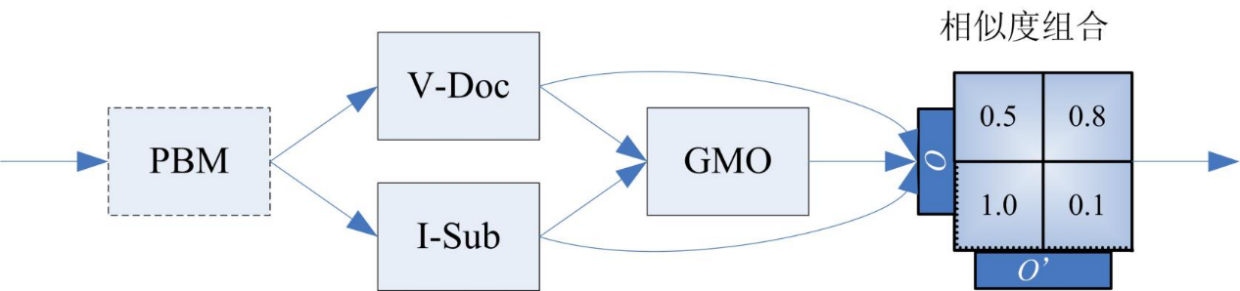


图9 计算相似度的组合策略

首先经由PBM进行分而治之，后进入到V-Doc和 I-Sub ，GMO接收两者的输出做进一步处理，GMO的输出连同V-Doc和I-Sub的输出经由最终的贪心算法进行选取。

3.4 知识推理

3.4.1 基于逻辑的推理

所谓推理就是通过各种方法获取新的知识或者结论，这些知识和结论满足语义。其具体任务可分为可满足性(satisfiability)、分类(classification)、实例化(materialization)。

可满足性可体现在本体上或概念上，在本体上即本体可满足性是检查一个本体是否可满足，即检查该本体是否有模型。如果本体不满足，说明存在不一致。概念可满足性即检查某一概念的可满足性，即检查是否具有模型，使得针对该概念的解释不是空集。

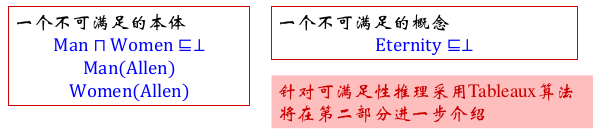


图10 不可满足的本体与概念

图10是两个不可满足的例子，第一个本体表示，Man 和 Women 的交集是空集，那么就不存在同一个本体Allen 既是Man 又是Women。 第二个概念是说概念Eternity是一个空集，那么他不具有模型，即不可满足。

分类，针对Tbox的推理，计算新的概念包含关系。如:

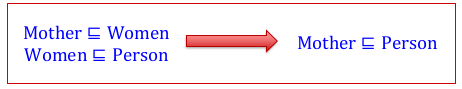


图11

即若Mother 是 Women的子集，Women是 Person的子集，那么我们就可以得出 Mother是 Person的子集这个新类别关系。

实例化即计算属于某个概念或关系的所有实例的集合。如图12所示:

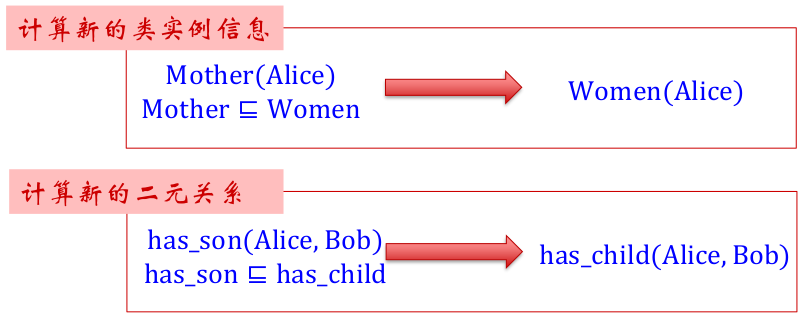


图12

第一个是计算新的类实例信息，首先已知Alice 是Mother，Mother 是 Women的子集，那么可知Alice 是一个Women。即为Women增加了一个新的实例。下面那个是计算新的二元关系，已知Alice 和Bob 有儿子，同时has\_son 是has\_child的子类，那么可知Alice 和Bob has\_child。

通过本体的概念层次进行推理时，其中概念主要是通过OWL(Web ontology language)本体语义进行描述的。OWL文档可以表示为一个具有树形结构的状态空间，这样一些对接结点的推理算法就能够较好地应用起来，例如文献[[58](http://www.xml-data.org/dzkj-nature/html/201645589.htm#b111)]提出了基于RDF和PD\*语义的正向推理算法，该算法以RDF蕴涵规则为前提，结合了sesame算法以及PD\*的语义，是一个典型的迭代算法，它主要考虑结点与推理规则的前提是否有匹配，由于该算法的触发条件导致推理的时间复杂度较高，文献[[59](http://www.xml-data.org/dzkj-nature/html/201645589.htm#b112)]提出了ORBO算法，该算法从结点出发考虑，判断推理规则中第一条推理关系的前提是否满足，不仅节约了时间，还降低了算法的时间复杂度。

3.4.2 基于图的推理

在基于图的推理方法中，文献[[60](http://www.xml-data.org/dzkj-nature/html/201645589.htm#b113)]提出的path- constraint random walk，path ranking等算法较为典型，主要是利用了关系路径中的蕴涵信息，通过图中两个实体间的多步路径来预测它们之间的语义关系。即从源节点开始，在图上根据路径建模算法进行游走，如果能够到达目标节点，则推测源节点和目标节点间存在联系。关系路径的建模方法研究工作尚处于初期，其中在关系路径的可靠性计算、关系路径的语义组合操作等方面，仍有很多工作需进一步探索并完成。

除上述两种类别的知识推理方法外，部分研究人员将研究重点转向跨知识库的推理方法研究，例如文献[[61](http://www.xml-data.org/dzkj-nature/html/201645589.htm#b75)]提出的基于组合描述逻辑的Tableau算法，该方法主要利用概念间的相似性对不同知识库中的概念进行关联、合并，通过已有的知识完成跨知识库的推理。

下面简单介绍一下Tableaux算法。

基于Tableaux运算适用于检查某一本体的可满足性，以及实例检测。其基本思想是通过一系列规则构建Abox,以检测可满足性,或者检测某一实例是否存在于某概念。这种思想类似于一阶逻辑的归结反驳。

Tableaux运算规则(以主要DL算子举例)如图13所示：

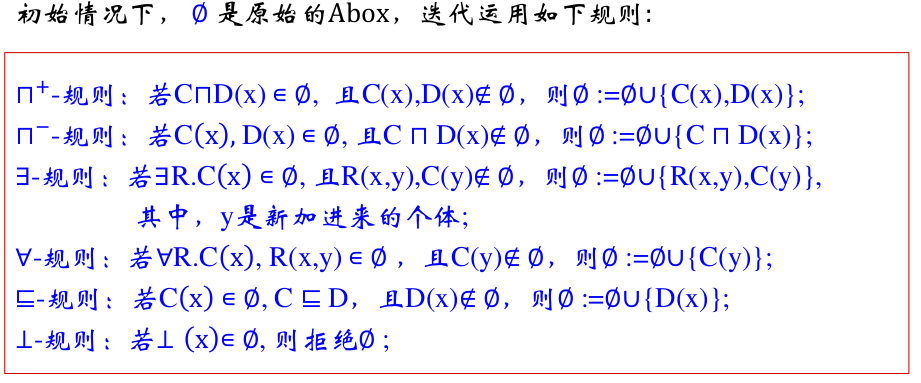


图13

Tableaux运算的基于Herbrand模型，Herbrand模型可以简单的理解为所有可满足模型的最小模型。

第四章 知识图谱技术应用

知识图谱在于对各对象实体关系、属性的链接，在互联网时代，几乎任何实体都能通过一系列的关系属性去连接到一起，那么我们就能通过知识图谱的搭建，去寻找之间的关系，进行一系列的推理，去预测某些知识。因此，在现代社会中， 知识图谱在许多领域都能很好地发挥作用。

4.1金融领域

在P2P网贷环境下，最核心的问题是风控，也就是怎么去评估一个借款人的风险。在线上的环境下，欺诈风险尤其为严重，而且很多这种风险隐藏在复杂的关系网络之中，而且知识图谱正好是为这类问题所设计的，所以我们“有可能”期待它能在欺诈，这个问题上带来一些价值。 等我们构建好知识图谱后，就可以使用它来解决具体问题。对于风控知识图谱来说，首要任务就是挖掘关系网络中隐藏的欺诈风险。从算法的角度来讲，有两种不同的场景：一种是基于规则的、另一种是基于概率的。基于目前AI的技术现状，基于规则的方法还是在垂直领域的应用中占据主导地位，但随着数据量的增加以及方法论的提升，基于概率的模型也将逐步带来更大的价值。

4.1.1基于规则的方法论

·不一致性验证

为了判断关系网络中存在的风险，一种简单的方法就是做不一致性验证，也就是通过一些规则去找出潜在的矛盾点。这些规则是以人为的方式提前定义好的，所以在设计规则这个事情上需要一些业务的知识。知识图谱可以进行信息的不一致性检查，来确定是不是存在可能的借款人欺诈的风险。

·组团欺诈

知识图谱可以帮助我们通过一些模式来找到有可能存在风险的团体或者子图，然后对这部分子图做进一步的分析。比如如果我们发现不同的三个实体共享了很多其他的信息，我们可以看做是一个团体，并对其做进一步的分析。

4.1.2基于概率的方法

除了基于规则的方法，也可以使用概率统计的方法。 比如社区挖掘、标签传播、聚类等技术都属于这个范畴。

社区挖掘算法的目的在于从图中找出一些社区。对于社区，我们可以有多种定义，但直观上可以理解为社区内节点之间关系的密度要明显大于社区之间的关系密度。由于社区挖掘是基于概率的方法论，好处在于不需要人为地去定义规则，特别是对于一个庞大的关系网络来说，定义规则这事情本身是一件很复杂的事情。

标签传播算法的核心思想在于节点之间信息的传递。这就类似于，跟优秀的人在一起自己也会逐渐地变优秀是一个道理。因为通过这种关系会不断地吸取高质量的信息，最后使得自己也会不知不觉中变得更加优秀。

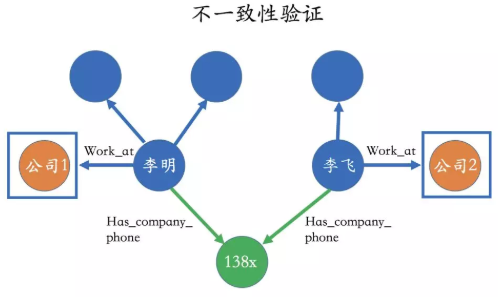


图14

相比规则的方法论，基于概率的方法的缺点在于：需要足够多的数据。如果数据量很少，而且整个图谱比较稀疏，基于规则的方法可以成为我们的首选。尤其是对于金融领域来说，数据标签会比较少，这也是为什么基于规则的方法论还是更普遍地应用在金融领域中的主要原因。

4.2教育领域

在教育行业中，知识图谱可以通过实体间的关系，来对学习路径做一个很好的规划。我们经常谈论个性化教育、因材施教的理念。其核心在于理解学生当前的知识体系，而且这种知识体系依赖于我们所获取到的数据比如交互数据、评测数据、互动数据等等。为了分析学习路径以及知识结构，我们则需要针对于一个领域的概念知识图谱，简单来讲就是概念拓扑结构。比如为了学习逻辑回归则需要先理解线性回归；为了学习CNN，得对神经网络有所理解等等。所有对学生的评测、互动分析都离不开概念图谱这个底层的数据。

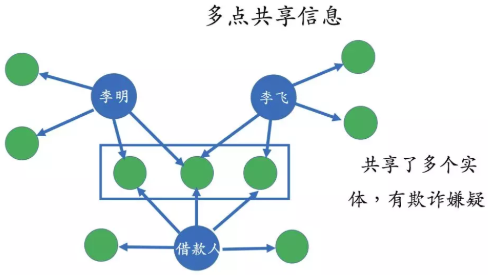


图15

4.3证券领域

在证券领域，我们经常会关心比如“一个事件发生了，对哪些公司产生什么样的影响？” 比如有一个负面消息是关于公司1的高管，而且我们知道公司1和公司2有种很密切的合作关系，公司2有个主营产品是由公司3提供的原料基础上做出来的。其实有了这样的一个知识图谱，我们很容易回答哪些公司有可能会被这次的负面事件所影响。当然，仅仅是“有可能”，具体会不会有强相关性必须由数据来验证。所以在这里，知识图谱的好处就是把我们所需要关注的范围很快给我们圈定。接下来的问题会更复杂一些，比如既然我们知道公司3有可能被这次事件所影响，那具体影响程度有多大？ 对于这个问题，光靠知识图谱是很难回答的，必须要有一个影响模型、以及需要一些历史数据才能在知识图谱中做进一步推理以及计算。

4.4推荐系统

·场景化推荐

比如用户购买了沙滩鞋，存在用户可能要去海边度假这样的场景，基于这样的场景可以继续给他推荐游泳衣、防晒霜或者其它的海岛旅游度假的产品。

·任务型推荐

比如用户买了牛肉卷或者羊肉卷，假设他实际上是要为了做一顿火锅。这时候系统可以给他推荐火锅底料或者是电磁炉。

·冷启动问题

推荐系统的冷启动一直是比较难以处理的问题，通常的做法是根据新用户的设备类型，或者他当前的时间位置等等，或者外面的关联数据来做推荐。可以基于知识图谱的语义关联标签进行推荐，比如旅游和摄影实际上是语义相近的两个标签，再比如相同的导演或者相同演员的电影在语义上也是比较相近的。

·个性化推荐

比如根据游戏来推荐游戏的道具。对于小白用户和骨灰级的用户，推荐的东西显然是不一样的

·跨领域的推荐问题

微博的信息流里会推荐淘宝的商品，然而微博和淘宝是两个不同的领域，它是怎么做到的呢？新浪微博有些用户会经常去晒黄山、九寨沟、泰山等这些照片，这个时候我们就知道他有可能是一位登山的爱好者，这个时候淘宝就会可以给他推荐登山的装备，登山杖、登山鞋等等这些装备，利用这些背景知识，能够打通不同的平台之间的语义鸿沟。

·知识形的推荐

是基于知识的。比如清华大学、北京大学都是顶级名校，复旦大学也同样是，这个时候是可以推荐复旦大学，再比如百度、阿里和腾讯都属于BAT级互联网公司，基于百度、阿里就可以推荐腾讯。

4.5知识图谱实际应用中的建议

1.知识图谱是一个比较新的工具，它的主要作用还是在于分析关系，尤其是深度的关系。所以在业务上，首先要确保它的必要性，其实很多问题可以用非知识图谱的方式来解决。

2.知识图谱领域一个最重要的话题是知识的推理。 而且知识的推理是走向强人工智能的必经之路。但很遗憾的，目前很多语义网络的角度讨论的推理技术（比如基于深度学习，概率统计）很难在实际的垂直应用中落地。其实目前最有效的方式还是基于一些规则的方法论，除非我们有非常庞大的数据集。

3.知识图谱工程本身还是业务为重心，以数据为中心。不要低估业务和数据的重要性。

4.图谱的设计也是一门艺术，不仅要对业务有很深的理解、也需要对未来业务可能的变化有一定预估，从而设计出最贴近现状并且性能高效的系统。在知识图谱设计的问题上，我们肯定会面临以下几个常见的问题：1. 需要哪些实体、关系和属性？ 2. 哪些属性可以做为实体，哪些实体可以作为属性？ 3. 哪些信息不需要放在知识图谱中？

第五章 知识图谱所面临的问题与挑战

在算法大时代，大部分人都会把目光投向算法的设计上，却偏偏忽略了知识库这一底层技术。因此有文提到“大家都在关注算法，这无可厚非，但是知识库这件事不仅仅是算法。”[62]知识图谱是谷歌对其所推出的大规模知识库产品的称呼，在此之前已有大量关于知识库研的研究，其中有很多代表性的工作，也相应提出了很多挑战性的问题。

5.1知识得自动获取本身就很难

对于结构化数据，需要处理复杂的表数据，而从链接数据中获取知识需要解决数据对齐问题，从半结构化（例如网站）数据中获取知识，需要对包装器进行定义、生成、更新与维护，而从文本中获取知识，需要考虑信息抽取中的准确率与覆盖率。

5.2不一定要选择多源获取

数据质量的挑战。包括命名模糊，数据输入错误，数据丢失，以及数据格式不一致等；数据规模的挑战。包括数据量大，数据种类多样，多种关系以及多种链接等。

5.3知识的表示学习不光看个体

面向知识的表示学习，需要是为知识建立统一的语义空间，使得语义可计算，从而实现预测、推理以及推荐。虽然由TransE为代表的知识表示方法已经得到了应用和扩展，但仍然具有很大的挑战，包括需要设计融合更多本体特征的知识图谱表示学习算法，分析知识图谱表示学习与本体推理之间的等价性分析，以及进一步研究神经符号系统等。

5.4知识推理其实是从无到有

目前知识推理还存在很多问题，比如说在本体推理与规则推理中，如何在大数据量下进行快速推理，以及对于增量知识和规则的快速加载，都需要研究解决。

5.5算法

算法到工具还有很长的一段路要走。不管是大数据，还是深度学习，如果它只是有几个算法的话，不会像现在这么火。大数据毕竟是有了hadoop和spark，深度学习毕竟是有了tensorflow，一些工具和生态圈的形成，让门槛变低了，大家才能更专注于业务层面的事情。目前，知识图谱还停留在解决技术问题层面，还没有开始去解决业务问题。因此，算法之后需要解决的事情，一是工具，二是高质量知识库的形成和积淀。

参考文献

1. 王知津, 王璇, 马婧. 论知识组织的十大原则[J]. 国家图 书馆学刊, 2012, 21(4): 3-11. WANG Zhi-jin, WANG Xuan, MA Jing. The ten principles of knowledge organization[J]. Journal of The National Library of China, 2012, 21(4): 3-11.
2. 索传军. 网络信息资源组织研究的新视角[J]. 图书馆情 报工作, 2013, 57(7): 5-12. SUO Chuan-jun. A new perspective for web resource organization research[J]. Library and Information Service, 2013, 57(7): 5-12.
3. 钟翠娇. 网络信息语义组织及检索研究[J]. 图书馆学研究, 2010, 75(17): 68-71. ZHONG Cui-jiao. Research on semantic organization of web information and retrieval[J]. Research on Library Science, 2010, 75(17): 68-71.
4. BERNERS-LEE T, HENDLER J, LASSILA O. The semantic Web[J]. Scientific American Magazine, 2008, 23(1): 1-4.
5. 王元卓，贾岩涛，刘大伟等．基于开放网络知识的信息检索与数据挖掘[J].计算机研究与发展，2014,52(2):456-474.
6. Cowie J, Lehnert W. Information extraction[J]. Communications of the ACM,1996,39(1):80-91
7. Chinchor N,Marsh E.Muc-7information extraction task definition [C]//Proc of the 7th Message Understanding Conf .Philadelphia:Linguistic Data Consortium,1998:359-367
8. 程学旗，靳小龙，王元卓等．大数据系统和分析技术综述[J]．软件学报，2014,25(9):1889-1908)
9. 钟秀琴，刘忠，丁盘苹. 基于混合推理的知识库的构建及其应用研究[J] ． 计算机学报，2012，35 ( 4) : 761766.
10. 刘峤, 李杨, 杨段宏, 等. 知识图谱构建技术综述[J]. 计 算机研究与发展, 2016, 53(3): 582-600.
11. 孙镇,王惠临. 命名实体识别研究进展综述[J]. 现代图书情报技术, 2010(6): 42-47.
12. CHINCHORN,MARSHE.Muc-7 information Extraction Task definition[C] //Proc of the 7th Message Understanding Conf. Philadelphia:Linguistic Data

Consortium, 1998: 359-367.

1. ZhouGD,SuJ.Named Entity Recognition using allHMM-based ChunkTagger.

ACL.Philadelpha, USA, 2002: 473-48

1. BorthwickA.Maximum Entropy Approach to Named Entity Recognition.PhD  
   Dissertion，NewYorkUniversity, 1999: 18-25
2. WHITELAWC,KEHLENBECKA,PETROVICN,et al. Web-scale named entity recognition[C]//Procof the 17th ACM Confon Information and Knowledge Management. NewYork: ACM, 2008
3. JAINA,PENNAC CHIOTTIM. Open entity extraction from web search query logs[C]//Proc of the 23rd Int Confon Computational Linguistics. Stroudsburg, PA:ACL,2010:510-518.
4. Kambhatla N.Combing lexical, syntactic, and semantic features with maximum entroopy models for extracting relation [C] //Proc of the 42nd Association for Computational Linguistics. Stroudsburg, PA: ACL, 2004: 510-518
5. Wu Fei,Weld DS. Open informatjion extraction using Wikipedia [C] //Proc of the 48th Annual Meeting of the Association for Computational Linguistics. Stroudsburg, PA: ACL, 2010: 118-127
6. Banko M,Cafarella MJ,Soderland S,et al.Open information extraction for the Web [C] //Proc of the 20th Int Joint Conf on Artificial Intelligence. New York: ACM, 2007: 2670-2676
7. Zhu Jun,Nie Zaijiang,Liu xiaojiang,et al. StatSonwball:A statistical approach to extracting entity relationships [C] //Proc of the 18th Int Conf World Wide Web. New York: ACM, 2009: 101-110
8. Alan A, Alexander L. KrakeN: N-ary facts in open information extraction [C] //Proc of the Joint Workshop on Automatic Knowldege Base Construction and Web-scale Knowledge Extraction. Stroudsburg, PA: ACL, 2012: 52-56
9. McCallum A. Joint inference for natural language processing [C] //Proc of the 13th Conf onn Computational Natural Language Learnging. Stroudsburg, PA: ACL, 2009: 1
10. 杨博,蔡东风,杨华.开放式信息抽取研究进展[J].中文信息学报,2014,4:1-11
11. ETZIONIO,CAFARELLAM,DOWNEYD,etal. Unsupervised named-entity extraction from the Web :an experimental study[J]. ArtificialIntelligence, 2005,165(1):91-134.
12. WU Fei,WELDDS .Antonomously Semantifying wikipedia[C] //Proc of the 16th ACM Confon Information and Knowledge Management. NewYork: ACM,2007:41-50.
13. 王宇,谭松波,廖祥文,等.基于扩展领域模型的有名属性抽取[J].计算机研究与发展,2010,47(9):1567-1573.
14. BORDES A, WESTON J, COLLOBERT R, et al.Learnting structured embeddings for knowledgebases[C] //Proc of AAAI. Menlo Park, CA: AAAI, 2011:301-306.
15. SOCHER R, CHEN D, MANNING C D, et al. Reasoningwith neural tensor networks for knowledge base completion [C]//Proc of NIPS. Cambridge, MA: MIT Press,2013: 926-934.
16. JENATTON R, ROUX N L, BORDES A, et al. A latentfactor model for highly multi-relational data[C]//Proc of NIPS. Cambridge, MA: MIT Press, 2012: 3167-3175.
17. SUTSKEVER I, TENENBAUM J B, SALAKHUTDINOVR. Modelling relational data using Bayesian clusteredtensor factorization[C]//Proc of NIPS. Cambridge, MA:MIT Press, 2009: 1821-1828.
18. YANG B, YIH W, HE X, et al. Embedding enties and relations for learning and inference in knowledge bases[C]//Proc of Int Conf on Learning Representations  
    (ICLR). France: ICLR Press, 2015.
19. SOCHER R, CHEN D, MANNING C D, et al. Reasoning with neural tensor networks for knowledge base completion[C]//Proc of NIPS. Cambridge, MA: MIT Press, 2013: 926-934.
20. NICKEL M, TRESP V, KRIEGEL H. A three-way model for collective learning on multi-relational data[C] //Proc of ICML. New York: ACM, 2011: 809-816.
21. BORDES A, USUNIER N, GARCIA-DURAN A, et al.Translating embeddings for modeling multi-relational data[C]//Proc of NIPS. Cambridge, MA: MIT Press, 2013:2787-2795.
22. 刘知远, 孙茂松, 林衍凯等. 知识表示学习研究进展[J]. 计算机研究与发展, 2016, 53(2): 1-16.
23. WANG Z, ZHANG J, FENG J, et al. Knowledge graph embedding by translating on hyperplanes[C]//Proc of AAAI. Menlo Park, CA: AAAI, 2014: 1112-1119.
24. LIN Y, LIU Z, SUN M, et al. Learning entity and relation embedding for knowledge graph completion[C]//Proc of AAAI. Menlo Park, CA: AAAI, 2015.
25. JI G, HE S, XU L, et al. Knowledge graph embedding via dynamic mapping matrix[C]//Proc of ACL. Stroudsburg, PA: ACL, 2015: 687-696.
26. XIAO H, HUANG M, HAO Y, et al. TransG: a generative mixture model for knowledge graph embedding[J]. Arxiv Preprint ArXiv, 2015, 1509: 05488
27. XU Xu-kan, FANG Dao-wei, JIANG Xun, et al. Research on knowledge granularity representation and standardization during knowledge organization[J]. [Document, Information & Knowledge, 2014(6): 101–106, 90](http://www.cnki.com.cn/Article/CJFDTOTAL-TSQC201406014.htm)
28. ZHANG Kun. The search technology oriented knowledge graph(Sogou)[EB/OL].(2015-02-18). <http://www.cipsc.org.cn/kg1/>.<http://www.cipsc.org.cn/kg1/>
29. ZHUANG Yan, LI Guo-liang, FENG Jian-hua. A survey on entity alignment of knowledge base[J]. [Journal of Computer Research and Development, 2016, 1: 165–192.](http://www.cnki.com.cn/Article/CJFDTOTAL-JFYZ201601017.htm)
30. JIANG Xun, XU Xu-kan. Knowledge service-oriented model of knowledge base logical structure research[J]. [Library and Information, 2013(6): 23–31.](http://www.cnki.com.cn/Article/CJFDTOTAL-BOOK201306005.htm)
31. STUDER R, BENJAMINS V R, FENSEL D. Knowledge engineering:Principles and methods[J]. [Data & Knowledge Engineering, 1998, 25(1): 161–197.](http://www.sciencedirect.com/science/article/pii/S0169023X97000566)
32. WONG W, LIU Wei, BENNAMOUN M. Ontology learning from text:a look back and into the future[J]. [ACM Computing Surveys, 2012, 44(4): 18–24.](http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.606.5947)
33. WU Wen-tao, LI Hong-song, WANG Hai-xun, et al. Probase:a probabilistic taxonomy for text understanding[C]//Proc of the 31st ACM SIGMOD Int Conf on Management of Data. New York:ACM, 2012.
34. HARRIS Z S. Distributional structure[J]. Word, 1954, 10(23): 146–162.
35. ETZIONI O, CAFARELLA M, DOWNEY D, et al. Unsupervised named-entity extraction from the Web:an experimental study[J]. Artificial Intelligence, 2005, 165(1): 91–134. [DOI:10.1016/j.artint.2005.03.001](http://dx.doi.org/10.1016/j.artint.2005.03.001)
36. 杨博, 蔡东风, 杨华. 开放式信息抽取研究进展[J]. 中文信息学报, 2014, 4: 1–11.   
    YANG Bo, CAI Dong-feng, YANG Hua. Progress in open information extraction[J]. [Journal of Chinese Information Processing, 2014, 4: 1–11.](http://www.cnki.com.cn/Article/CJFDTOTAL-MESS201404001.htm)
37. Carnegie Mellon University. NELL[EB/OL].[2016-06-08]. <http://rtw.ml.cmu.edu/rtw/>. <http://rtw.ml.cmu.edu/rtw/>
38. ZENG Yi, WANG Dong-sheng, ZHANG Tie-lin, et al. CASIA-KB:a multi-source Chinese semantic knowledge base built from structured and unstructured Web data[C]//Semantic Technology. Berlin:Springer, 2014:75-88.
39. WANG C, DANILEVSKY M, DESAI N, et al. A phrase mining framework for recursive construction of a topical hierarchy[C]//Proc of the 19th ACM SIGKDD Int Conf on Knowledge Discovery and Data Mining. New York:ACM, 2013:437-445.
40. FADER A, SODERLAND S, ETZIONI O. Identifying relations for open information extraction[C]//Proc of the Conf on Empirical Methods in Natual Language Processing. Stroudsburg, PA:ACL, 2011:1535-1545.
41. MENDES P N, MUHLEISEN H, BIZER C. Sieve:Linked data quality assessment and fusion[C]//Proc of the 2nd Int Workshop on Linked Web Data Management at Extending Database Technology. New York:ACM, 2012:116-123.
42. DONG Xin, GABRILOVICH E, HEITZ G, et al. Knowledge vault:a Web-scale approach to probabilistic knowledge fusion[C]//Proc of the 20th Int Conf on Knowledge Discovery and Data Mining. New York:ACM, 2014:601-610.
43. TAN C H, AGICHTEIN E, IPEIROTIS P, et al. Trust, but verify:Predicting contribution quality for knowledge base construction and curation[C]//Proc of the 7th ACM Int Conf on Web Search and Data Mining. New York:ACM, 2014:553-562.
44. GENG Xia, ZHANG Ji-jun, LI Wei-yan. Knowledge graph construction techniques[J]. [Computer Science, 2014, 41(7): 148–152.](http://www.cnki.com.cn/Article/CJFDTOTAL-JFYZ201603009.htm)
45. 龚资. 基于OWL描述的本体推理研究[D]. 长春:吉林大学, 2007.   
    GONG Zi. Research on ontology reasoning based on OWL[D]. Changchun:Jilin University, 2007.
46. LIU Shao-yuan, HSU K H, KUO Li-jing. A semantic service match approach based on wordnet and SWRL rules[C]//Proc of the 10th IEEE Int Conf on E-Business Engineering. Piscataway, NJ:IEEE, 2013:419-422.
47. LAO N, MITCHELL T, COHEN W W. Random walk inference and learning in a large scale knowledge base[C]//Proc of EMNLP. Stroudsburg, PA:ACL, 2011:529-539.
48. 蒋勋, 徐绪堪. 面向知识服务的知识库逻辑结构模型[J]. 图书与情报, 2013(6): 23–31.   
    JIANG Xun, XU Xu-kan. Knowledge service-oriented model of knowledge base logical structure research[J]. [Library and Information, 2013(6): 23–31.](http://www.cnki.com.cn/Article/CJFDTOTAL-BOOK201306005.htm)
49. Gowild AI Lab . Adversarial Learning for Chinese NER from Crowd Annotations