知识图谱研究报告

知识图谱研究是自然语言处理，也是人工智能研究的一个重要领域，通过使用自然语言处理、语言表示和深度学习等技术，大规模知识图谱系统被广泛应用于信息检索、自动问答、推荐系统、金融风控等领域。本报告通过对知识图谱技术进行梳理，介绍如下内容：

知识图谱概念及核心要素。介绍知识图谱的相关概念，知识图谱和自然语言处理的关系，知识图谱的发展历程和未来方向。

知识图谱在各个领域的实践案例。介绍知识图谱在业界各个公司的重要应用，在搜索、广告、推荐、风控和自动问答等领域的不同应用。介绍知识图谱不同应用背后的关键技术，对不同技术进行分析和深入研究。

知识图谱的关键技术研究。介绍知识图谱构建的核心技术信息抽取，关注关系抽取的方法，并介绍知识图谱应用的核心技术知识表示和知识推理，同时介绍如何将不同的知识图谱结合进行知识融合。

知识图谱对银联的启示。介绍知识图谱如何对银联的各项业务进行推动，如何在公司内部使用知识图谱，如何开展后续的工作。

## 概念及核心要素

本部分介绍知识图谱的概念和发展历程，包括知识图谱的定义、分类和知识图谱核心要素，并对知识图的研究现状和未来发展挑战进行介绍。

### 知识图谱的定义

知识图谱，是为了表示知识，描述客观世界的概念、实体、事件等之间关系的一种表示形式。在维基百科的词条中，知识图谱是Google用于增强其搜索引擎功能的知识库。这一概念起源于语义网络——提出于20世纪五六十年代的一种知识表示形式，是一种基于图的数据结构，由节点(Point)和边(Edge)组成。节点表示的是概念或对象，边表示各个节点之间的关系。知识图谱在表现形式上与语义网络比较类似，不同的是，语义网络侧重于表示概念与概念之间的关系，而知识图谱更侧重于表述实体之间的关系。现在的知识网络被用来泛指大规模的知识图谱。知识图谱中包含下面几点概念：

**实体** 实体指独立存在且具有某种区别性的事物。如一个人、一种动物、一个国家、一种植物。具体的事物就是实体所代表的内容，实体是知识图谱中的基本元素，不同的实体间有不同的关系。

**语义类** 语义类具有同种特性的实体构成的集合，如人类、动物、国家、植物等。概念主要指集合、类别、对象类型、事物的种类，例如人物、地理等。

**内容** 内容通常是实体和语义类的名字、描述、解释等，展示形式一般有文本、图像、音视频等。

**属性值** 属性值指对象属性的值，不同的属性类型对应于不同类型属性的边。

**关系** 在知识图谱上，关系的表现形式是一个将节点（实体、语义类、属性值）映射到布尔值的函数。

**通用知识图谱** 通用知识图谱注重广度，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。

**领域知识图谱** 领域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。在领域知识图谱中，通常需要结合行业特点，行业知识进行构建，实体的属性与数据模式往往比较丰富，不同行业的知识图谱具有专业性。

如图1是关于银联的一种简单的知识图谱网络表示。而基于三元组的知识图谱是知识图谱的一种常用表示形式，三元组的基本形式主要包括实体关系型三元组如(头实体，关系，尾实体)，实体属性型三元组如(实体，属性，属性值)等,(中国银联，位于，上海)就是一种实体关系型三元组，而（中国银联，成立于，2002年）就是一种实体属性型三元组。

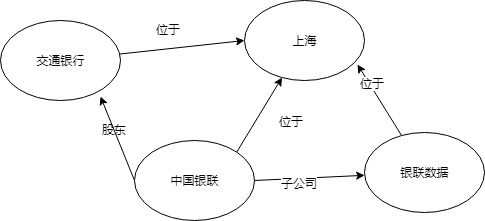


图 1 知识图谱网络表示示意图

知识图谱的数学定义表示为，其中,表示知识图谱中实体的集合，表示关系的集合，而，表示实体和关系组成的向量空间。



图 2常见的知识图谱系统

如图2展示了一些常用知识图谱的信息，并介绍这些知识图谱构建相关的公司和应用产品。知识图谱技术的发展起源于20世纪60年代的语义网络(Semantic Network)，随着人工智能对知识的需求，知识图谱的发展大概经历了语义网络、专家系统（Expert System）、语义网（Semantic Web）、新兴知识图谱构建等几个阶段。现代新兴的知识图谱的显著特点是规模巨大，无法单一依靠人工和专家构建。传统的知识库，如由Douglas Lenat从1984年开始创建的常识知识库Cyc仅包含700万条的事实描述。由著名人工智能专家Marvin Minsky于1999年起开始构建的ConceptNet常识知识库依靠了互联网众包、专家创建和游戏三种方法，但早期ConceptNet规模在百万级别，最新的ConceptNet 5.0也仅包含2800万RDF三元组关系描述。现代知识图谱如谷歌和百度的知识图谱都已经包含超过千亿级别的三元组，阿里巴巴于2017年8月份发布的仅包含核心商品数据的知识图谱也已经达到百亿级别。DBPedia已经包含约30亿RDF三元组，多语种的大百科语义网络BabelNet包含19亿的RDF三元组，Yago3.0包含1.3亿元组，Wikidata已经包含4265万条数据条目，元组数目也已经达到数十亿级别。截止目前，开放链接数据项目Linked Open Data 统计了其中有效的2973个数据集，总计包含大约1494亿三元组。

根据覆盖范围而言，知识图谱也可分为**开放域通用知识图谱**和**垂直行业知识图谱**。开放通用知识图谱注重广度，强调融合更多的实体，较垂直行业知识图谱而言，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。

### 知识图谱系统构建架构

知识图谱在逻辑上可分为模式层与数据层两个层次，数据层主要是由一系列的事实组成，而知识将以事实为单位进行存储。如果用(头实体，关系，尾实体)、(实体、属性，属性值)这样的三元组来表达事实，可选择图数据库作为存储介质，例如开源的Neo4j、Twitter的FlockDB、sones的GraphDB等。模式层构建在数据层之上，是知识图谱的核心，通常采用本体库来管理知识图谱的模式层。本体是结构化知识库的概念模板，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。

知识图谱的体系架构是其指构建模式结构，如图 3知识图谱构建系统架构图所示。其中虚线框内的部分为知识图谱的构建过程，也包含知识图谱的更新过程。知识图谱构建从最原始的数据（包括结构化、半结构化、非结构化数据）出发，采用一系列自动或者半自动的技术手段，从原始数据库和第三方数据库中提取知识事实，并将其存入知识库的数据层和模式层，这一过程包含：信息抽取、知识表示、知识融合、知识推理四个过程，每一次更新迭代均包含这四个阶段。知识图谱主要有自顶向下(top-down)与自底向上(bottom-up)两种构建方式。自顶向下指的是先为知识图谱定义好本体与数据模式，再将实体加入到知识库。该构建方式需要利用一些现有的结构化知识库作为其基础知识库，例如Freebase项目就是采用这种方式，它的绝大部分数据是从维基百科中得到的。自底向上指的是从一些开放链接数据中提取出实体，选择其中置信度较高的加入到知识库，再构建顶层的本体模式。目前，大多数知识图谱都采用自底向上的方式进行构建，其中最典型就是Google的Knowledge Vault和微软的Satori知识库。现在也符合互联网数据内容知识产生的特点。

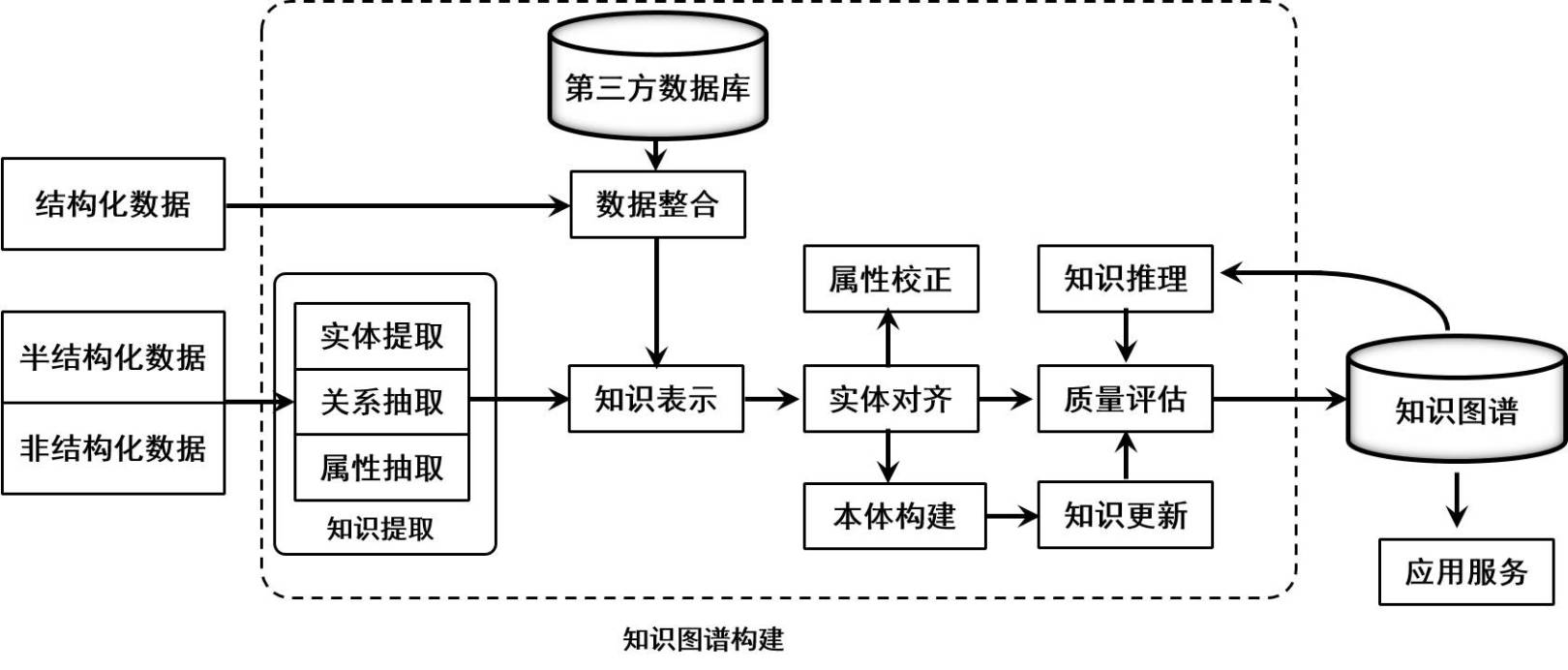


图 3知识图谱构建系统架构图

### 知识图谱核心技术要素

国内外知识图谱的应用领域大致可归纳为以下领域：搜索引擎、智能客服、风险防控、智能投顾以及其他医疗、金融、教育等垂直行业应用中。

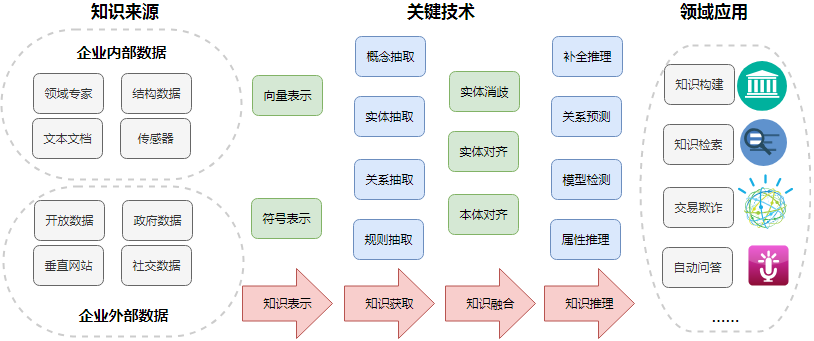


图 4知识图谱研究的核心技术

如图 4知识图谱研究的核心技术所示，知识图谱工程的核心流程包括：知识建模、知识抽取、知识融合、知识推理、知识检索、知识分析等核心环节。一般的技术流程包括：首先确定知识表示模型，然后根据数据来源选择不同的知识获取手段导入知识，接下来需要综合利用知识推理、知识融合、知识挖掘等技术对所构建的知识图谱进行质量提升，最后根据场景需求设计不同的知识访问与呈现方法，如：语义搜索、问答交互、图谱可视化分析等，下面简要概述这些技术流程的核心技术要素。

对于知识图谱构建来说，首先要基于企业内部和企业外部的数据构建知识库，并结合领域专家、结构数据、开放知识、社交数据等多源信息进行信息融合。对于多源信息融合构建的知识库要基于符号表示、向量表示等方法进行知识表示和获取，从而能进行知识推理、关系预测等模型构建，这个构建的知识推理模型被交易欺诈、自动问答、知识检索等。

### 知识图谱未来发展突破

知识图谱当前的研究方向主要分为三部分：（1）知识图谱的构建，不同构建方法需要结合，包括基于专家领域知识构建，基于自动或半自动的信息抽取构建等；（2）知识图谱推理，分为基于图的关系路径进行知识图谱推理和基于低维度向量表示的知识图谱表示推理；（3）知识图谱应用，包括问答系统、推荐系统和搜索系统中使用知识图谱提高系统可用性。

当前知识图谱在种类、规模和质量等方面都实现了跨越式发展，未来知识图谱将展现出新的发展空间和挑战。发展空间主要体现在以下几个方面：

#### 知识图谱非结构化数据处理

目前，非结构化数据占全球信息85%以上，并呈海量增长态势，这些科学知识信息的形式包括视频、音频文件和图片等，是知识图谱未来发展空间最大的资源。目前，已经有一些开放、共享、面向异构科学知识来源的数据库正在尝试构建。下一阶段的知识图谱构建将是面向网络环境下大规模知识库的，针对多样化的知识来源，开展语义资源的理解和知识总结，从海量异构数据所含有的丰富知识中完成知识图谱的构建将是目前以及未来面临的主要挑战。

未来大数据与云计算的结合使用，将有助于进一步提升科研相关数据存储、数据处理和数据分析的能力，使得知识图谱的构建基于更大量、更全面、更准确的科学知识库，并能将更多的隐性知识转变为显性知识用于知识图谱的构建。

#### 知识图谱的表示推理发展

机器自动推理技术是人工智能发展的核心问题之一，知识图谱基于图结构的数据能帮助机器理解现实世界中的常识知识、语义知识，让知识理解自然语言、人类的逻辑推理等高层次的知识。

知识图谱和语音、图片等技术任然存在鸿沟，使用知识图谱的智能推理技术，结合一阶逻辑推理、知识表示、随机游走等不同领域，通过构建知识图谱自动推理技术，能有效帮助解决机器进行问题理解、社区发现、用户分类等不同的问题。同时知识图谱能够突破单一领域的限制，直观揭示关联领域的关系，因此近年来越来越多的被应用于学科交叉的研究中。随着语义理解、智慧搜索和机器智脑的飞速发展，未来将会涌现出更多大规模、高质量的科学知识图谱，这些知识图谱将会交叉融合，构建出关系更为复杂、智能程度更高、覆盖更为广泛、信息更新更加实时的超大型知识图谱。

#### 知识图谱商业应用价值发现

目前已经有一些互联网公司推动知识图谱在智能搜索、自动问答、查询理解、商业广告、社交媒体等领域的落地应用，但是这些领域构建知识图谱的关键技术还不成熟，自然语言理解、智能推理等相关技术任然存在瓶颈，相关领域的专业人才不够，如果有对自然语言处理、知识推理、大数据和云计算等多方面复合型的人才，将会引领知识图谱走上新的发展高度。

知识图谱在商业上的应用也不成熟，当前仅仅关注于信息检索、知识问答等领域，在医疗、客服等领域的应用人才很稀缺，缺少跨领域的复合型人才。此外，在很多场景上，知识图谱都需要和现有的机器学习技术进行结合，将知识图谱和深度学习相互融合，才能发现更大的商业价值。

## 二、知识图谱实践案例



本部分介绍知识图谱如何在各个领域进行应用。知识图谱本质上是一个异构的网络图，描述了节点（实体）之间的关系，是一种基于图的数据结构。知识图谱可以作为数据处理的知识基础，根据不同的行业、不同的场景，通过现代化信息技术手段形成适应于各行各业的应用服务。国内知识图谱已经在部分领域中进行应用探索，如搜索引擎、智能客服、交易风控等。

### 搜索引擎

搜索引擎是从相关文档集合中查找用户所需信息的过程。先将信息按一定的方式组织和存储起来，然后根据用户的需求从已经存储的文档集合当中找出相关的信息。信息检索包括“存”与“取”两个方面，对信息进行收集、标引、描述、组织，进行有序的存放是“存”。按照查询机制从有序存放的信息集合（数据库）中找出用户所需信息或获取其线索的过程是“取”。信息检索的原理是将用户输入的检索关键词与数据库匹配获取，将可能的结果反馈给用户。

传统的搜索引擎在处理搜索结果时面临着以下几个挑战：

（1）搜索引擎无法理解一词多义的情况。例如“苹果”既可以是一种水果，也可以是一个手机品牌，也可能是一部电影。“射雕英雄传”可以指金庸的小说，也可以指各类改编的电影、电视剧等。搜索引擎自身只能够根据搜索词在页面内的相关程度将搜索结果排序而不能根据搜索词的各种含义进行分类处理，导致用户必须要指定领域关键词才能得到准确的结果。

（2）搜索引擎无法提供精确的、结构化的知识。当用户搜索某个关键词时，通常他想要得到的是这个关键词的精确的介绍，如对某个人物、某些事件的介绍，而不是相关的信息汇集。同时用户需要在搜索结果中找到百科类网站的结果，并将这些百科知识进行结构化，转化为自己的知识也是目前搜索引擎无法提供的。

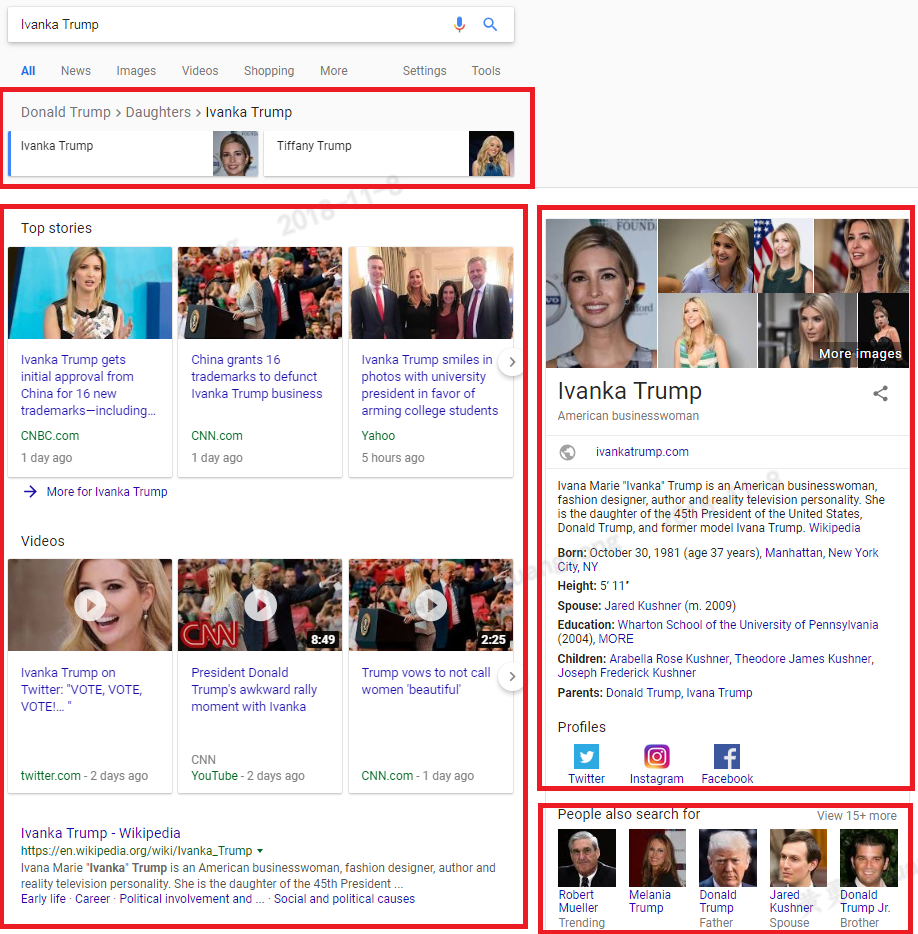


图 5 “相关性”检索和“精确”检索的知识卡片对比

（3）搜索引擎无法理解人类的逻辑和自动推理。例如我们在搜索“特朗普的女儿是谁”时，搜索引擎可能提供特朗普的女儿的新闻，特朗普的新闻，但是并没有提供特朗普女儿精确的信息；对于一些复杂的推理如“谢霆锋的前女友的老公是谁？”传统的搜索引擎很难直接检索到相关的信息。相关性的检索结果很难做出正确的逻辑推理，理解人的语义知识。

谷歌将知识图谱技术应用在以下几个方面：

（1）检索结果结构化

通过知识图谱网络，谷歌可以理解用户搜索词的具体指代，并以知识卡片的形式用户提供该关键词的简单摘要信息。如图 5 “相关性”检索和“精确”检索的知识卡片对比所示，对于“Ivanka Trump”的检索结果中，不仅是对于相关故事、新闻、视频的检索，更有涉及推理结果的知识如“Donald Trump -> Daughters ->Ivanka Trump”,也有涉及个人结构化知识介绍的知识如：“Ivanka Trump Born October,30,1981”等信息，还有涉及语义推荐的结果，推荐相关的人的链接如“Donald Trump”等信息。

（2）检索结果推荐

通过知识图谱网络，谷歌可以确定与用户搜索关键词同类的关键词，并以“人们也在搜索”的形式向用户提供建议。如上图在搜索“Ivanka Trump”的时候，谷歌通过知识图谱网络查找到在科学领域与“Ivanka Trump”相关的家人、朋友的姓名，并在搜索结果之下呈现。

由于谷歌通过搜索引擎向公众展示了知识图谱对大量跨领域知识的关联性整合的潜力，不仅其它的搜索引擎也开始采用精确搜索建议、知识卡片、“人们也在搜索”这三种应用来为用户提供更好的搜索体验，一些自身业务流程与搜索紧密关联的应用——如社交网络、购物网站等也使用了类似的技术。此外知识图谱的影响力也扩散到了其它需要大量整理以及关联性分析的领域。

（3）知识检索的推理

知识图谱的推理技术是人工智能需要解决的核心问题，通过知识图谱逻辑推理算法，可以实现机器的自动推理。对于检索问题“特朗普的女儿是谁？”，可以直接返回准确的检索答案“Ivanka Trump”和“Tiffany Trump”。对于一些复杂的逻辑推理问题如“谢霆锋的前女友的老公是谁？”，相关的知识检索系统也可以进行准确的检索，返回答案。除此之外，谷歌的知识图谱还用来自动挖掘新知识的准确程度、文本中命名实体的识别、纯文本搜索词条在知识图谱上的结构化搜索词条的转换等，效果都领先于其他公司，而且很多技术都实现了产品化。

谷歌的知识图谱相继融入了维基百科、CIA世界概览等公共资源以及从其他网站搜集、整理的大量语义数据，微软的BingSearch和Facebook、Twitter等大型社交服务站点达成了合作协议，在用户个性化内容的搜集、定制化方面具有显著的优势。 国内的主流搜索引擎公司，如百度、搜狗等在近两年来相继将知识图谱的相关研究从概念转向产品应用。搜狗的知立方是国内搜索引擎行业的第一款知识图谱产品，它通过整合互联网上的碎片化语义信息，对用户的搜索进行逻辑推荐与计算，并将最核心的知识反馈给用户。百度将知识图谱命名为知心，主要致力于构建一个庞大的通用型知识网络，以图文并茂的形式展现知识的方方面面。百度自然语言处理部是百度最早成立的部门之一，除了问句理解、答案抽取、观点分析与聚合等自然语言处理技术方案，百度也应用知识图谱基础技术实现深度问答、对话系统、语义计算、知识挖掘等，目前已经在搜索、度秘、咨询流（feed）等多个产品中实现应用。阿里巴巴的知识图谱技术也为其产品服务，在电商平台中构建知识图谱实现智能导购，同时进行全网用户兴趣挖掘，在客服场景中也运用自然语言处理技术打造机器人客服，例如蚂蚁金融智能小宝、淘宝卖家的辅助工具千牛插件等。

### 问答系统

问答系统是信息检索系统的一种高级形式，能够以准确简洁的自然语言为用户提供问题的解答。在问答系统中有查询式理解与知识检索这两个重要的过程，并且与智能搜索中相应过程中的相关细节是完全一致的。多数问答系统更倾向于将给定的问题分解为多个小的问题，然后逐一去知识库中抽取匹配的答案，并自动检测其在时间与空间上的吻合度等，最后将答案进行合并，以直观的方式展现给用户。目前，很多问答平台都引入了知识图谱，例如华盛顿大学的Paralex系统和苹果的智能语音助手Siri，都能够为用户提供回答、介绍等服务；亚马逊收购的自然语言助手Echo，它授权了Nuance的语音识别技术，采用True Knowledge引擎进行开发，也可提供类似Siri的服务。下图 6一种基于知识图谱的问答系统的模型结构图是问答系统的系统架构图。

对于金融类行业来说，客服的人力成本很高，而收益较低，而在客服领域使用基于知识图谱的问答技术，也是知识图谱应用的热点之一。传统人工客服在这一课题上面临以下几个挑战：

（1）客服人员难以查询内部业务知识，难以解决客户问题。当公司的业务日益复杂，尤其像银联系统这种涉及银行、商户和消费者业务，同时要解决云闪付APP内的使用问题和银行卡的消费者服务问题，这些对客服人员都有很高的业务知识要求，如何能使用内部知识库系统，提高客服对业务知识的查询和使用能力，是智能客服面临的问题之一。

（2）传统客服难以识别客户的意图。对于使用APP出现问题的客户，在使用电话网络咨询客服的时候，有着较强的客户标签，这需要客服能够在服务客户过程中能提前识别客户的意图，结合客户的内部标签，客户分类等信息进行分析，提供有效的服务方式。

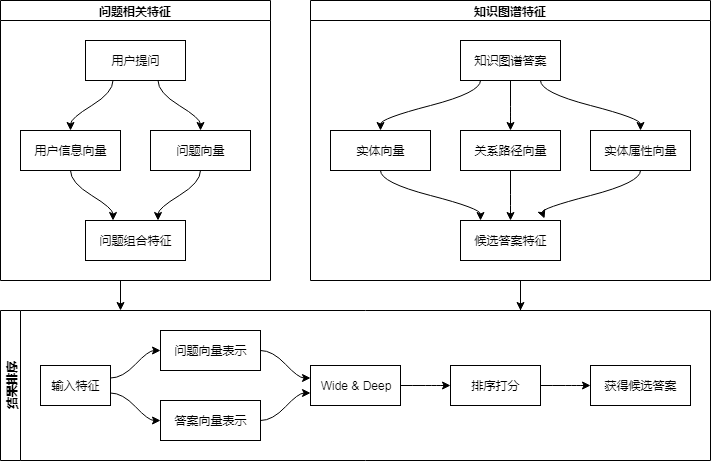


图 6一种基于知识图谱的问答系统的模型结构图

美国最大的第三方客户中心Amdocs就向其客户（AT&T、Verizon以及Sprint）提供基于知识图谱的智能客服服务，并知识图谱应用于以下两个方面：

（1）客服知识图谱的构建

应用知识图谱技术，一家公司可以将其自身业务整理成一张知识图谱网络，并能结合用户自身的档案信息匹配到关联性最高的潜在问题。

例如某位客户发现自己的手机号被流氓软件订购了一系列自己不想要的套餐，并拨打了运营商的客服电话。运营商的客服根据客户的档案信息（用户账单上出现了不寻常的订阅信息、该服务历史上受到很多投诉），结合知识图谱网络，最终推断出用户呼叫客服的目的（取消该订阅）。

（2）针对客户的个性化解决方案

不同的客户对公司的重要程度各不相同，公司永远想将有限的资源向优质客户倾斜，从而能最大化每一个客户的价值。Amdocs的智能客服系统能够将客户的账户信息结构化，并与公司的业务知识图谱匹配，从而能给出每一个用户的评分。除此之外，结合前述的目的推测系统，Amdocs的智能客服系统还能找到匹配度最高的解决方案，并向客服提示方案以及相关的话术指引。还是以上文退订套餐的客户为例，Amdocs通过用户的历史账单信息推断出该用户为优质客户，从而在各种可能的解决方案中选出最能挽留优质客户的方案（退订+返还已产生的费用），客服可以直接根据Amdocs系统展示的信息与客户交流。

总的来说，知识图谱技术通过对客户现有资料的匹配达到提前感知用户需求的目标，并结合公司业务规则的知识图谱网络自动给出合适的解决方案，从而达到了“想用户所想”的理想目标。不仅人工客服能够应用该工具大大提升效率，如聊天机器人等自动化的对话平台也能通过接入知识图谱的方式来提供更智能化的客服服务。

### 交易欺诈检测

如何识别交易中的欺诈用户和异常交易，发现在营销活动中的黄牛用户，无论对于公司还是对于正常的交易用户，都有着非常重要的作用。

由于知识图谱技术能跨领域的各类结构化及非结构化数据，构建实体和实体之间的逻辑关系，并基于表示学习、逻辑推理等技术进行推理表示，从而能辅助金融风控，尤其辅助。因此有不少金融领域公司试图使用该技术来为金融相关性分析提供助力，如下图提供一种基于知识图谱推理的风控模型示意图，图中借助图表示学习和交易行为、用户特征构建统一的机器学习模型，提高风控的推理能力和可解释性。

一些用户为了获利，短时间内大量使用其他人的卡进行交易，造成交易欺诈越来越普遍。同时由于现代金融体系的不断发展，互联网金融的快速发展，金融风控也面临一些新的挑战，主要包括以下几点问题：

（1）依赖风控规则引擎，但是风控业务快速变化。基于历史的业务经验和历史风控事件进行分析，根据业务不同、对风险的偏好和业务的处理不同，构建业务相关的风控规则，并发现交易欺诈，但当业务逻辑发送变化时，风控引擎灵活性不足，导致风控效果不佳；

（2）基于机器学习的异常检测方法，关注数据相关性关系，而非实体之间的因果关系。当前风控基于聚类或半监督的算法，对正常的交易和异常的交易进行聚类分析；通过交易特征因子、用户特征因子构建特征工程，利用神经网络算法、支持向量机、贝叶斯网络等训练分类模型对不同类型的交易进行分类识别。但通常这类机器学习算法是基于数据之间的相关性发现，而非数据之间的因果逻辑关系，这会导致风控的逻辑不清晰，风控规则可解释性差。

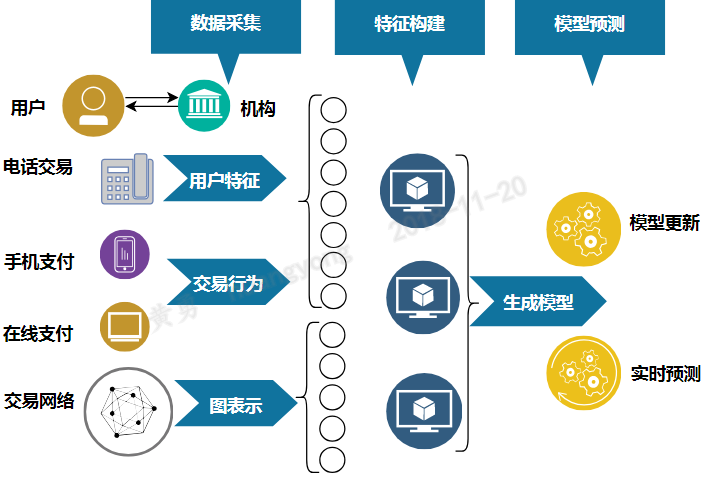


图 7 基于知识图谱推理的风控模型示意图

如图 7 基于知识图谱推理的风控模型示意图所示将图表示用于交易欺诈检测。除此之外知识图谱技术可以应用在交易欺诈检测和风控领域以下几个方面：

（1）结构化的金融知识图谱

利用知识图谱技术将各种结构化以及非结构化金融产品信息、政策信息、行业信息、新闻信息、公司信息等金融领域的核心数据综合形成一张知识图谱，并能向用户提供各种角度的数据筛查及统计服务。从而解决了金融数据无结构化、无法合理展示分析的问题。

除了宏观的金融知识数据，为了对交易欺诈检测，也可以结合用户互相转账数据、用户和商户之间的交易数据结构化，构建用户-商户-银行之间的交易网络进行交易分析，并基于知识图谱的交易行为进行建模分析，发现交易欺诈。

（2）金融领域知识图谱推理

知识图谱技术对金融领域的助力在于为金融分析提供了结构清晰的情报数据库，但仅有图数据和实体关系是不够的，基于图数据的知识图谱推理也是金融领域关注的热点之一。一种常见的知识图谱推理算法是知识图谱表示学习，这类算法将图中的节点和边转化为矩阵，对矩阵进行降维获取图中的节点和边的向量表示，这类算法能很好的获取图的全局结构关系。除此之外，很多知识图谱表示推理也结合深度学习，学习获取图中的节点和边的非线性结构特征，基于自动编码（autoencoder）等方法获取节点和边的向量表示。

在金融行业中，反欺诈是一个重要的环节。它的难点在于如何将不同部门子系统中的数据整合在一起。通过知识图谱，一方面有利于组织相关的知识碎片，通过深入的语义分析与推理，可对信息内容的一致性充分验证，从而识别或提前发现欺诈行为；另一方面，知识图谱本身就是一种基于图结构的关系网络，基于这种图结构能够帮助人们更有效地分析复杂税务关系中存在的潜在风险。

### 电商知识图谱

随着新零售、多语言、线上线下结合的复杂购物场景的发展，电子商务所用到的数据也往往超出了以往的文本范围，构建并合理使用知识图谱成为电商领域一个重要的挑战。电商面临的问题包括：

（1）数据分散，无法互联，非结构化数据都是分散在各个来源，而且多是非结构化文本表示。电子商务的商品类目体体系复杂而低效。传统的文本分析，使用的数据如query、title、评论、攻略等，由于用户习惯和商家诉求，存在不同于普通文本的语法结构，难以利用。各个电商部门由于业务快速发展，往往需要维护自己的一套商品体系，这是后期做商品管理和搜索的非常关键的一环，但是由于应用场景的行业属性不一样，比如闲鱼的"包配饰"由于业务场景高频会是一个需要再细分的类目，但在淘宝系由于交易搜索低频，"鞋包配饰"仅仅是二手闲置下的一个小类目，这造成各个部门需要费力地维护在自己的商品体系上的查询和搜索。

（2）底层类目体系的重复、或过于粗糙、或过于细分等问题对推荐结果体验带来诸多问题，同时，由于没有类目体系，对用户和商品所在场景的认知深度不足，导致推荐结果效果很差。例如用户购买木炭后，基于传统的机器学习算法会推荐相关的木质产品、牛肉干等产品，但基于知识图谱的语义逻辑推理可以发现用户更想购买烧烤相关的产品，因此推荐羊肉串、烧烤炉等产品可以获取更好的语义推荐效果。

基于上述问题，阿里构建了公司统一知识图谱通用平台。如图 8阿里巴巴电商知识图谱系统架构所示。阿里将电商知识图谱应用于如下几个方面：

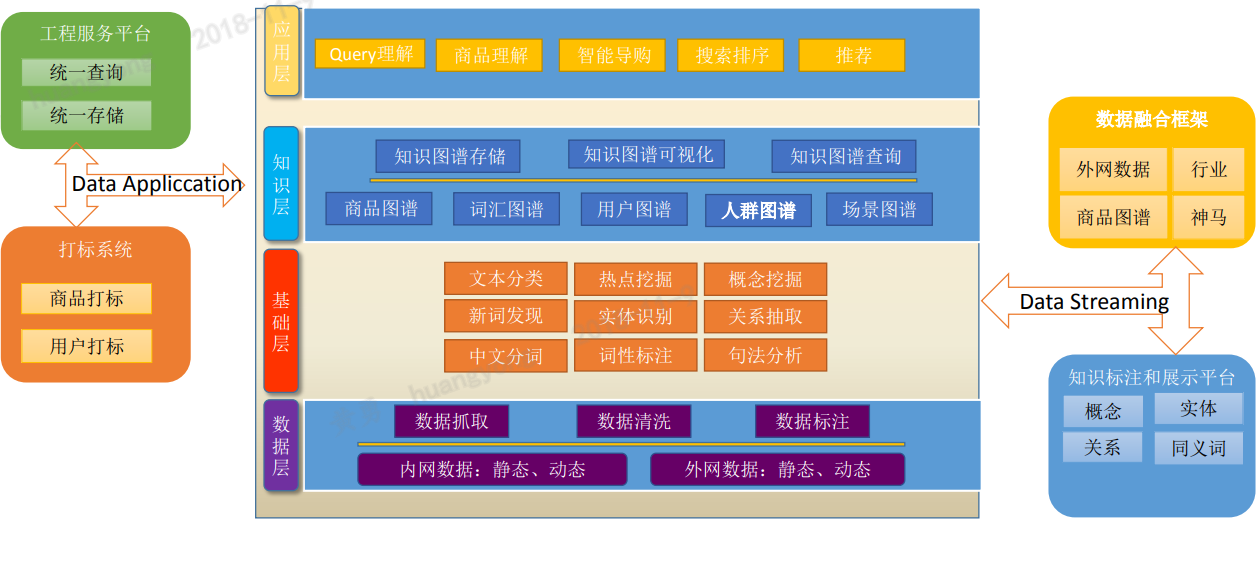


图 8阿里巴巴电商知识图谱系统架构

（1）阿里巴巴电商知识图谱通过数据层、基础层、知识层和应用层的构建，将复杂的数据和场景结构化，将分散的数据统一模式（schema），对数据进行深度认识，通过精准认知用户需求，实现电商场景下关联人-货-场的联动，赋能电商业务方和相关的行业。

（2）商品图谱构建，并进行语义推荐。例如如商品打标：商品打标是将知识和商品建立关联的关键技术，数据通过打标建立与item的联系，在商品打标完成后我们就可以实现从query到商品的整个语义认知闭环。如图 9 电商基于知识图谱的语义推荐的推荐场景，基于知识图谱可以实现更高效的商品语义推荐。



图 9 电商基于知识图谱的语义推荐

（3）数据标注、审核和展示。集成了所有知识标注和审核流程，并且提供了数据查询和可视化，后期算法的概念挖掘服务和商品打标服务也会通过相关系统对外提供。同时数据审核在不断试错过程中我们已经建立了一套比较完善的从初审到终审的流程，对数据进行结构化、关系标注和知识拓展。同时，阿里也基于知识图谱构建了一系列的业务选择和广告投放平台。

## 三、知识图谱关键技术研究



### 关系抽取（Relation Extraction）

关系抽取是构建知识图谱的核心技术之一，也是知识抽取中最重要的任务之一。关系抽取任务被广泛应用于数据简化和构建知识图谱中。给定用户输入的一段自然语言，在正确识别实体的基础上，抽取它们之间的关系就是亟待解决的重要问题。在关系提取的过程主要挑战有：（1）如何确定领域知识？（2）如何提取三元组？（3）如何评估提取出的三元组的可靠性？

目前解决这些问题的方法分为串联抽取和联合抽取两类。一般传统的串联抽取方法是在实体抽取的基础上进行实体之间关系的识别。在这种方法中，先期实体识别的结果会影响到关系抽取的结果，前后容易产生误差累积。针对这一问题，基于传统机器学习的联合模型（Joint model）被提出并逐步用于对这一类的NLP任务进行联合学习。

知识图谱中信息一般以三元组的方式进行组织，一般来说有（实体，关系，实体）和（实体，属性，属性值）两种形式。关系抽取方法也可以基于模型类型分为：基于特征模板的方法，基于核函数的监督学习方法，基于远程监督的方法，和基于深度学习的监督或远程监督方法。



#### 人工模式匹配

通过人工的模式匹配发现实体和实体之间的关系，并抽取关系。由于需要语言学家对需要抽取的领域做深入的了解和分析，穷举所有关系模板使得效率低下，一些学者开始尝试研究抽取模式的自动生成方法。郑家恒等人计算模式实例间的相似度，采用单链法聚类抽取模式，并将模式一一分类，归并同一类型的模式实例并获得最终的关系抽取模式。聚类方法在不同领域的抽取效率还没有具体的定量标准，姜吉发等人提出了基于自举的关系与关系模式获取方法。

对关系抽取来说，可以通过 “such as”可以判断这是一种 “IS-A” 的关系，例如“Y including X”可以获取一个“X IS-A Y”这种关系。基于人工模式匹配的的方法优点是有较高的准确率（Precision），同时，针对特定的领域，手写规则能定制不同的模式匹配模板。但手写的模式匹配通常也有较大的缺点，这种方法需要为每条关系来定义模式，并且很难考虑所有的模式，通常这种方法的召回率比较低。

#### 监督学习算法

监督的关系抽取通常包括有监督学习算法和半监督学习算法。半监督学习主要是利用少量的标注信息进行学习，这方面的工作主要是基于 Bootstrap 的方法。利用少量的实例作为初始种子(seed tuples)的集合，然后利用模式学习方法进行学习，通过不断的迭代，从非结构化数据中抽取实例，然后从新学到的实例中学习新的模式并扩充模式集合。

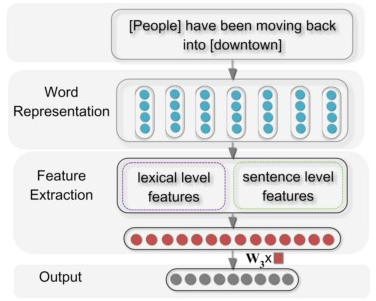


图 10 基于CNN的有监督关系抽取算法示意图

有监督的机器学习将关系抽取看作二元分类问题，人们使用人工标注语料得到正例和反例，如图 10 基于CNN的有监督关系抽取算法示意图。通过语义分析、句法分析等选取特征集合，构造合适的分类器并训练得到分类模型，常用的如支持向量机和最大熵分类器。然后根据分类模型构造抽取器从无结构文本中抽取出实体关系，通常的抽取模型有条件随机场模型等。有监督学习的关系抽取方法也分为流水线方法和联合抽取方法，因此需要大量的训练语料，尤其是对基于神经网络的方法，需要大量的语料进行模型训练，因此这些方法都不适用于构建大规模的知识库。

#### 开放域关系抽取

关系抽取研究已经从传统的限定类别、限定领域信息抽取任务发展到开放类别、开放领域关系抽取。技术手段也从基于人工标注语料库的统计方法发展为有效地挖掘和集成多源异构网络知识并与统计方法结合的开放式关系抽取。

在开放式实体关系抽取方面，华盛顿大学的人工智能研究组在这方面做了大量代表性的工作，并且开发了一系列原型系统：TextRunner、WOE、ReVerb等。对于关系名称的抽取，TextRunner把动词作为关系名称，抽取过程类似于语义角色标注，通过动词链接两个论元，从而挖掘论元之间的关系。WOE是以 Wikipedia 为目标，从中抽取实体关系类型，从而构建实体的属性描述框架。在 Wikipedia 中，在每个概念条目中，通常都会有人工标注的Infobox信息，其中包含了大量实体关系类别，依据Infobox中蕴含的大量实体关系对，WOE 对于概念条目正文进行回标，以此来自动产生关系值抽取的训练语料，从而解决了开放式关系抽取训练语料不足的问题。

除了关系抽取外，知识抽取按任务可以分为概念抽取、实体识别、关系抽取、事件抽取和规则抽取等。传统专家系统时代的知识主要依靠专家手工录入，难以扩大规模。现代知识图谱的构建通常大多依靠已有的结构化数据资源进行转化形成基础数据集，再依靠自动化知识抽取和知识图谱补全技术从多种数据来源进一步扩展知识图谱，并通过人工众包来进一步提升知识图谱的质量。

### 知识融合（Knowledge Fusion）

在构建知识图谱时，可以从第三方知识库产品或已有结构化数据获取知识输入。例如，关联开放数据项目（Linked Open Data）会定期发布其经过积累和整理的语义知识数据，其中既包括前文介绍过的通用知识库 DBpedia和 Yago，也包括面向特定领域的知识库产品，如 MusicBrainz和DrugBank等。当多个知识图谱进行融合，或者将外部关系数据库合并到本体知识库时需要处理两个层面的问题：（1）通过模式层的融合，将新得到的本体融入已有的本体库中，以及新旧本体的融合；（2）数据层的融合，包括实体的指称、属性、关系以及所属类别等，主要的问题是如何避免实例以及关系的冲突问题，造成不必要的冗余。

数据层的融合是指实体和关系（包括属性）元组的融合，主要是实体匹配（或者对齐），由于知识库中有些实体含义相同但是具有不同的标识符，因此需要对这些实体进行合并处理。此外，还需要对新增实体和关系进行验证和评估，以确保知识图谱的内容一致性和准确性，通常采用的方法是在评估过程中为新加入的知识赋予可信度值，据此进行知识的过滤和融合。实体对齐的任务包括实体消歧和共指消解，即判断知识库中的同名实体是否代表不同的含义以及知识库中是否存在其他命名实体与之表示相同的含义。实体消歧（entity disambiguation）专门用于解决同名实体产生歧义问题的，通常采用聚类法，其关键问题是如何定义实体对象与指称项之间的相似度，常用方法有：空间向量模型（词袋模型）、语义模型、社会网络模型、百科知识模型和增量证据模型等。一些最新的工作利用知识图谱嵌入方法进行实体对齐，并引入人机协作方式提升实体对齐的质量。

本体是针对特定领域中Schema定义、概念模型和公理定义而言的，目的是弥合词汇异构性和语义歧义的间隙，使沟通达成共识。这种共识往往是通过一个反复的过程达到，每次迭代都是一次共识的修改。因此，本体对齐通常带来的是共识模式的演化和变化，本体对齐的主要问题之一也可以转化为怎样管理这种演化。常见的本体演化管理框架有KAON、Conto-diff、OntoView等。

#### 实体链接

实体链接（entity linking）是指对于从文本中抽取得到的实体对象，将其链接到知识库中对应的正确实体对象的操作。实体消歧是专门用于解决同名实体产生歧义问题的技术，通过实体消歧，就可以根据当前的语境，准确建立实体链接，实体消歧主要采用聚类法。其实也可以看做基于上下文的分类问题，类似于词性消歧和词义消歧。基本思想是首先根据给定的实体指称项，从知识库中选出一组候选实体对象，然后通过相似度计算将指称项链接到正确的实体对象。实体消岐的研究难点包括：（1）仅关注如何将从文本中抽取到的实体链接到知识库中，忽视了位于同一文档的实体间存在的语义联系。（2）如何发现实体之间的共现关系，如何将多个实体链接到知识库中对应的实体。

实体链接的流程通常包括：（1）基于实体抽取，获得文档、查询（query）中的实体指称项（mention）；（2）基于文档上下文、用户信息等进行实体消岐和共指消解，判断知识库中的实体（entity）和实体指称项（mention）是否具有相同的含义；，例如对于“天龙八部刘亦菲版”中的实体指称项“天龙八部”，将其链接到实体“天龙八部（2003年胡军版电视剧）”，而非“天龙八部（金庸小说）”，从而能在知识图谱中获取准确的检索结果。

#### 知识合并

在获得不同来源的新知识之后，需要对其进行整合，以消除矛盾和歧义，比如某些实体可能有多种表达，某个特定称谓也许对应于多个不同的实体等。知识合并包括将不同来源的知识库进行合并、将结构化数据库中的数据和知识库进行合并。目前知识库合并主要分为两种：（1）合并外部知识库，主要处理数据层和模式层的冲突，（2）合并关系数据库，包括RDB2RDF等方法。

### 知识推理（Knowledge Reasoning）

知识图谱推理（知识推理）是人工智能系统的重要组成部分，包括归纳推理、演绎推理、确定性推理和不确定性推理等不同种类。智能系统的知识推理过程是通过推理机来完成的，推理机是智能系统中用来实现推理的程序。推理机的基本任务就是在一定控制策略指导下，搜索知识库中可用的知识，与数据库匹配，产生或论证新的事实，搜索和匹配是推理机的两大基本任务。

知识图谱相关的推理算法目前主要分为单步推理（直接关系，没有考虑路径特征）和多步推理（间接关系，考虑路径特征），在单步推理和多步推理中主要包括4个研究方向：基于传统的规则推理、基于分布式表示推理、基于神经网络的推理、基于上述方法的混合推理。在每个研究方向又延伸了很多小的方向和方法，已存在的算法和论文比较多，具体可以参考如下的分类图。

#### 一阶逻辑推理

一阶逻辑推理。一阶逻辑是一种形式系统（Formal System)，即形式符号推理系统，也叫一阶谓词演算、低阶谓词演算、限量词（Quantifier）理论，也有人称其为“谓词逻辑”。

一阶逻辑推理是一种一阶规则学习算法，Foil是著名的一阶规则学习算法，它遵循序贯覆盖框架采用自顶向下的规则归纳策略。其中序贯覆盖通过规则学习的目标是产生一个能覆盖尽可能多的样例的规则集。最直接的做法是逐条归纳：在训练集上每学到一条规则，就将该规则覆盖的训练样例去除，然后以剩下的训练样例组成训练集重复上述过程。由于每次只处理一部分数据，因此也称为分治（separate-and-conquer）策略。自顶向下策略从比较一般的规则开始，逐渐增加新文字以缩小规则覆盖范围，直到满足预定条件为止，也称为生成-测试法，是规则逐渐特化的过程，是从一般到特殊的过程。

除了一阶逻辑推理之外，基于路径排序算法和基于关系规则发现的知识推理算法也很重要，关系规则发现和随机游走算法等。

#### 规则挖掘算法

一个知识库包含一组事实是孩子c有一个妈妈m，这样可以推理得出孩子妈妈的丈夫f很可能是孩子的父亲。该逻辑规则形式化的描述如下：

MotherOf(m,c)∧MarriedTo(m,f)→FatherOf(f,c)

规则挖掘算法通过知识图谱中频繁存在的规则，对知识图谱中存在的关系进行规则挖掘。通过一个或多个频繁存在的规则进行统计计算，计算出频繁出现的规则的概率，从而预测知识图谱中频繁出现的规则。

挖掘这种规则可帮助做一下四种事情，辅助机器进行知识推理：（1）利用这种规则来推理出新的事实，而这些被挖掘出的新的事实可以使知识库更完整。 （2）这些规则可以检测出知识库潜在的错误例如一个陈述是一个与一个男孩无关的人是这个男孩的父亲，这样的陈述很可能是错误的。（3）有很多推理工具依赖其他工具提供规则，所以这些被挖掘出来的规则可以用于推理。（4）这些规则描述一个普遍的规律，这些规律可以帮我我们理解分析知识库中的数据，如找到一些国家通常与说同一种语言的国家交易。或结婚是一个对称关系，或使用同一个乐器的音乐家通常互相影响等等。 AMIE的目标是从RDF格式的知识库中挖掘如上所述的逻辑规则，可以用于知识图谱如YAGO、Freebase和DBpedia等不同的知识库中，对这些知识库进行知识库补全。

#### 路径排序算法

如图 11 路径排序算法计算路径示意图所示，对于某个给定的关系，如“位于”这个关系来说，判断实体和实体之间是否具有“位于”关系，是知识推理中的一类重要算法。路径排序的基本想法是用连接两个实体的路径作为特征，来预测两个实体之间的关系。学习阶段分为特征抽取、特征计算和构造分类器三个部分。预测阶段是使用分类器对任务进行预测。

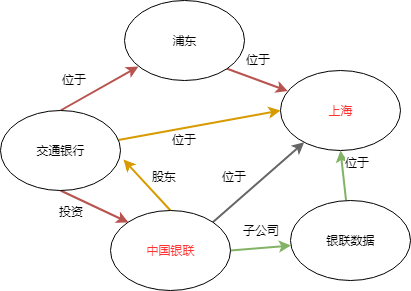


图 11 路径排序算法计算路径示意图

路径排序算法的想法是将路径作为特征来进行预测。对于给定的关系如“位于”，获取知识图谱中所有和“位于”相关的头尾实体对，如：（上海，中国），（中国银联，上海），（交通银行，上海）基于随机游走算法获取这些头尾实体对在知识图谱中相互连接的关系路径，如对于（中国银联，上海）这个实体对，可以基于广度优先搜索算法或深度优先搜索算法，计算抽取相关路径如：“中国银联->股东->交通银行->位于->上海”、“中国银联->股东->交通银行->位于->浦东->位于->上海”，“中国银联<-投资<-交通银行->位于->浦东->位于->上海”。对于上述“位于”关系下所有实体对的关系路径获取关系路径类型: “-股东->位于-”、“-股东->位于->位于-”、“-投资<-位于->位于-”等，欺诈“->”和“<-”表示实体和实体之间的关系路径的方向。基于上述抽取的关系路径类型计算获取实体和实体之间具有某种关系的概率，从而进行预测。

### 知识表示（Knowledge Representation）

知识表示将知识图谱中的实体关系表示为低维度向量。基于三元组的知识表示形式比较直观，但是其在计算效率、数据稀疏性等方面面临着诸多问题。如图 12 知识表示应用示意图所示，以深度学习为代表的表示学习技术可以将实体的语义信息表示为稠密低维实值向量，进而在低维空间中高效计算实体、关系及其之间的复杂语义关联，对知识库的构建、推理、融合及应用均具有重要的意义。知识表示学习的代表模型有距离模型、单层神经网络模型、双线性模型、神经张量模型、矩阵分解模型、翻译模型等。对于大规模知识图谱来说，三元组是稀疏、不完整、维度很高的数据，只有将知识图谱嵌入到低维度向量空间中，才能更好的实现语义和语法分析，并且实现知识库补全、知识融合等技术。

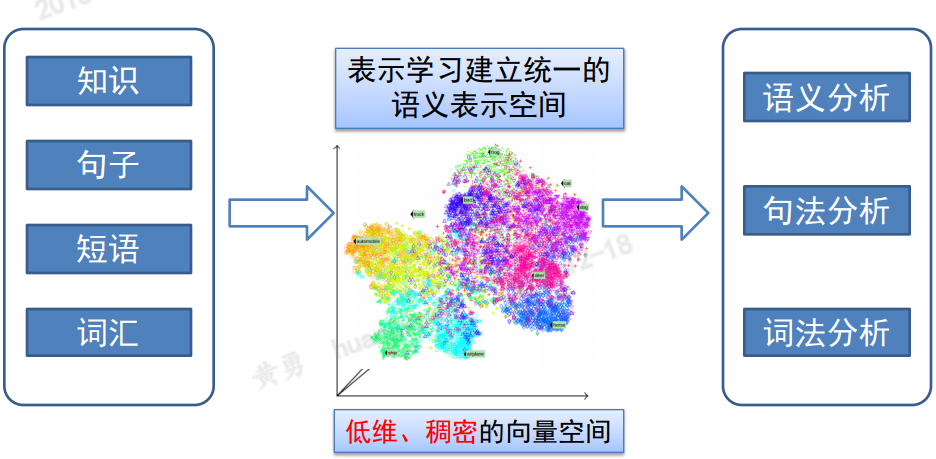


图 12 知识表示应用示意图

#### 协同过滤学习模型

基于协同学习的知识库补全算法借鉴了推荐系统中的协同过滤模型，将基于“用户-商品”的协同矩阵分解技术进行拓展，构建了基于“实体-关系-实体”的三维张量模型，进行矩阵分解和实体关系学习预测。RESCAL是一种基于表示学习的潜藏特征模型，基于矩阵分解、协同过滤进行知识库补全，RESCAL通过构建三维张量分解技术进行知识库补全，将“实体-关系-实体”构建三维的张量矩阵，提供了一种有效的表示学习方法。

#### 神经网络翻译模型

基于神经网络翻译模型的知识库补全算法近年来获得了极大的关注，很多基于神经网络翻译模型的知识库补全算法被提出如TransH、TransR等。TransE是一种典型的基于神经网络翻译模型的机器学习模型，研究考虑如何将知识库中的多种不同的关系、实体，学习获得它们的低维向量表示，期望能获得“h + r ≈ t”效果，其中h和t分别是头实体和尾实体学习到的低维向量，r是头实体和尾实体之间的关系向量表示。图 13知识图谱表示学习算法示意图展示了将知识图谱中的节点转化为二维空间中连续的向量，从而描述实体之间的语义关系。

一种常见的知识图谱表示学习算法是距离模型，这种算法将知识库中实体以及关系的结构化表示方法(structured embedding)。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关系对的向量空间中，并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。很多知识图谱表示算法也结合关系路径进行学习，如图是一种基于随机游走的关系路径表示学习算法，通过学习实体之间的关系路径，获取实体的低维度向量表示，从而能进行推理、分类等不同的任务。

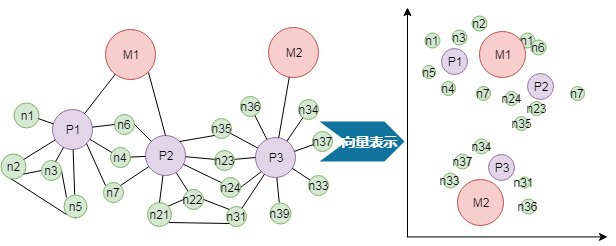


图 13知识图谱表示学习算法示意图

## 四、对银联的启示



知识图谱是人工智能学习逻辑推理的重要基础，知识图谱在搜索、问答、风控等领域，已经引起了业内的广泛关注，并获取了巨大的成功。由于其强大的知识描述能力和关系表达能力，在金融支付行业中具有广泛的应用前景。目前,银联已经在智能客服、风险防控这两个领域开展了与知识图谱相关联的基础性研究工作，并取得了一些基础的研究成果。

### 银联智能客服应用

智能客服能提高客服的服务效率，帮助客服快速检索到相关的结果。如图 14一种基于业务知识的问答系统示意图所示，展示了基于业务需求构建的知识图谱智能客服的观点系统框架。通过在业务需求层明确需求，对这些需求进行建模构建知识图谱系统，并进行存储和知识推理，最后将构建的知识图谱系统进行业务建模，构建知识查询，自动问答等服务。

#### 自动问答

在智能问答这一细分领域，研究院也开展了关于智能机器人的研究，主要为在银行卡办理业务场景下通过智能机器人向用户提供银行卡信息查询、转账和办理，以及优惠活动查询等功能。通过意图识别、模板匹配等技术，完成对话、问答、命令等多个语义理解引擎的构建，实现银行卡办理业务场景的对话过程模拟。

该智能机器人也未使用知识图谱作为知识存储的方式，而是采用了传统的关系型数据库的数据结构进行数据存储。但是，在人机对话的过程中，无论是使用知识图谱的图数据库还是使用关系型数据库存储实体、关系、属性等数据，均会遇到两类关键问题：一是问题入口确认，即对于用户的提问，如何确定对应于后台系统中的哪一类问题；二是关联参数确认，即确认用户问题后，如何确定该问题所关联的具体参数。针对第一个问题，需要根据具体场景做定制化的关键词匹配，如在转账场景下，用户的表述中可能完全没有“转账”这个词，而是会出现“打钱”、“汇款”等与转账同义的词，或者是“打给某人多少钱”，“汇给某人多少钱”这种表述，因此，需要事先在后台录入多个与转账问题相关联的关键词及相关提问方式，然后对于用户的具体问题采用命名实体识别技术匹配到对应的问题入口中。针对第二个问题，即在确定问题入口后，如确定为“转账”，则还需确定对方的帐户名、帐号、金额等属性才能发起转账流程，这些属性已事先在后台中设置在转账这一问题下， 然后需要通过多轮对话的形式向用户获取。

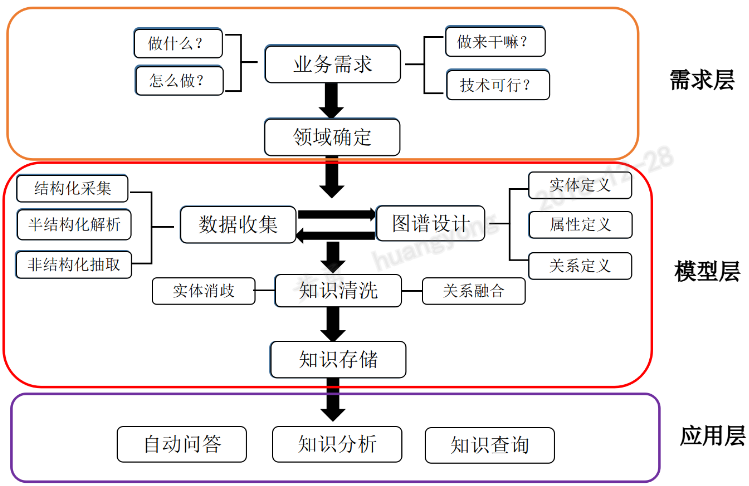


图 14一种基于业务知识的问答系统示意图

在智能机器人方面的研究是知识图谱相关技术的研究基础。在智能问答之一细分领域，建立一定规模以上的动态更新知识图谱是一个实现智能问答的较好解决方案。鉴于该领域用户的问题往往多种多样，涉及公司不同的业务场景、不同的产品、不同的优惠营销活动等，涉及多个客观事物的实体以及实体与实体之间的关系，将面临更加复杂的问题入口确认与更加多样的关联参数确认，而运用知识图谱及相关技术是解决负责问题入口确认和关联参数确认的途径之一。同时，知识图谱在实体、关系、属性值对的提取上，已经形成了一系列的理论研究成果，通过相应的算法和模型工具，结合对应场景进行一定优化和人为监督后，可以实现数据的自动提取，进而减少数据获取的人力成本，极大提升了知识图谱的数据规模。并且，通过相应的图数据库进行数据存储，可以优化数据查询效率，尤其在大规模数据处理上，提升数据的处理速度。

#### 问答推理

在智能客服领域，研究院也研究了基于知识图谱推理的问答系统方法。给定一个提出的问题，结合提问的上下文，提问人的基本信息等特征构建特征工程，基于向量空间模型将这些特征信息转化为向量。对于构建好的知识图谱作为候选答案集合，预先学习知识图谱中的实体向量，关系路径向量和实体属性向量。将问题特征和知识图谱中的候选答案特征结合，训练神经网络排序模型，预测给定提问在知识图谱中的答案。

对于用户提问的问题，并结合此问题前提出的两个问题作为问题上下文，组成问题特征，使用中文分词，并去除停用词，使用查询重写算法，对部分错误问题重构。将这些字符串使用向量空间模型进行向量化表示。对于用户的基本信息，如用户的职业、用户年龄、用户行为、用户兴趣爱好等信息构建成一个用户向量空间模型，和问题特征类似，将上述特征转化为用户信息向量。同时构建问题和用户信息交叉项特征，结合用户的基本信息和提问信息作为新的特征向量，构建排序模型进行训练。

对于知识图谱中的实体，利用实体和实体之间的关系，基于矩阵分解算法将“实体-关系-实体”组成的三维矩阵降维获取实体的低维度向量表示，如计算得到实体“中国银联”的五维向量为[0.1,0.8,0.3,0.4,0.6]。对于知识图谱中实体和实体之间的关系路径，也进行低维度嵌入，计算关系路径的算法基于知识图谱中的随机游走算法，获取关系路径的低维度向量表示。对于知识图谱中的实体属性，抽取实体的基本属性特征值，构建实体属性特征矩阵。如对于实体“中国银联”，抽取相关的实体属性信息如：成立日期、员工人数等实体属性值，并对于“实体-关系-实体”类型的三元组，计算头实体和尾实体之间的差值，将头实体属性、尾实体属性、头尾实体差值属性组合，构建实体属性特征矩阵。同时考虑实体属性值单位不同，将不同类型的实体属性特征值进行归一化，并学习实体属性矩阵的低维度向量表示。

对上述实体向量、关系路径向量、实体属性向量组成的候选答案特征，问题向量、用户信息向量组成的问题特征作为输入向量，构建Wide & Deep神经网络模型进行打分，预测给定问题下最优的答案。通过将实体属性向量、关系路径特征向量、实体向量嵌入，并将用户信息向量和问题向量组合，作为输入层，训练神经网络模型，预测问题相关的答案。

### 营销活动欺诈检测

#### 异常交易欺诈检测

在风险防控领域，银联研究院基于银行卡交易信息，构建了银行卡交易图谱，从图谱中挖掘异常交易结构。如图 15基于知识表示的异常交易欺诈检测模型图所示，可以基于用户交易网络构建知识图谱，基于卡在A地查询，在B地取现；多张卡同时在多地取现；多张卡互相转账，转完以后取现等信息，将这些异常的交易节点构建知识图谱的图特征表示系统。基于知识图谱中的随机游走算法进行建模，并结合传统的交易风控规则进行联合建模发现异常交易，从模型效果来看，已经能够较好地侦测出异常卡。

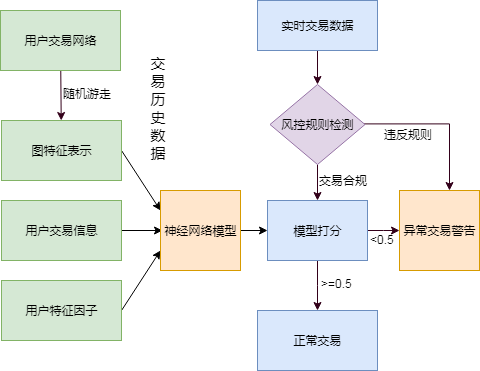


图 15基于知识表示的异常交易欺诈检测模型图

对于银行卡交易图谱，它是一种较为简单的知识图谱，主体和关系相对简单，仅包括银行卡号、地址两类主体，相互之间的关系也只有查询、转账、取现、消费四类。由于银联的主营业务中包含了银行卡查询、交易数据，因此对于银行卡交易知识图谱的构建具备了天然优势，无需从外部知识中提取主体和关系，只需从内部数据库中抽取数据并作相应的数据预处理即可。

#### 营销羊毛党识别

在银联每年的62营销等活动中，经常出现羊毛党的异常交易，他们占用太多银联的营销活动资源，使得营销受众变小，影响活动效果，对于羊毛党的识别上，银联研究院也开展了相关研究，通过卡号、手机号、设备号三个数据的绑定关系，用户和商户的交易关系构建黄牛图谱，从而在识别异常卡或者异常设备中提供辅助作用。对于黄牛图谱，实体和关系相对简单，其涉及的主体包括卡号、手机号、设备号三类，关系有绑定关系、用户和商户的交易等。

考虑到用户和商户之间通常会串通作弊，异常用户经常在相同的商户上进行交易，并且交易的笔数很多，交易金额相同，交易行为很相似。我们可以对用户节点和商户节点进行表示学习构建，获取实体和关系的低维度向量表示，从而可以建立机器学习异常检测分类模型，发现异常节点。

### 本章小结

知识图谱本身可以看做是一种新型的信息系统基础设施可以在银联的智能客服、问答推理、欺诈检测等业务中进行使用，银联对知识图谱使用有着很深的潜力。从数据维度来看，知识图谱要求用更加规范的语义来提升银联内部数据的质量，用链接数据（Linked Data）的思想提升数据之间的关联度，终极目标是将非结构、无显示关联的粗糙数据逐步萃取提炼为结构化、高度关联的高质量知识。每个企业都应该将知识图谱作为一种面向数据的信息系统基础设施，并进行持续性建设和投入才能获取知识图谱的收益。

从技术维度来看，知识图谱的构建涉及知识表示、关系抽取、图数据存储、数据融合、推理补全等多方面的技术，而知识图谱的利用涉及语义搜索、知识问答、自动推理、欺诈检测、描述性数据分析等多个方面。要构建并利用好知识图谱也要求系统性的综合利用好来自于知识表示、自然语言处理、机器学习、图数据库、多媒体处理等多个相关领域的技术，而非单个领域的单一技术。因此，未来一个发展趋势是，知识图谱的构建和利用都应注重系统思维，只有结合各自自然语言处理和深度学习算法，才能更好的构建和应用知识图谱。