【刘知远】知识图谱——机器大脑中的知识库

墨白找 关注
0.1 2016.05.29 11:07* 字数 11724 阅读 4505 评论 3 喜欢 49

作者：刘知远（清华大学）；整理：林颖（RPI） 本文来自Big Data Intelligence
知识就是力量。——[英]弗兰西斯·培根
1 什么是知识图谱
在互联网时代，搜索引擎是人们在线获取信息和知识的重要工具。当用户输入一个查询词，搜索引擎会返回它认为与这个关键词最相关的网页。从诞生之日起，搜索引擎就是这样的模式，直到2012年5月，搜索引擎巨头谷歌在它的搜索页面中首次引入“知识图谱”：用户除了得到搜索网页链接外，还将看到与查询词有关的更加智能化的答案。如下图所示，当用户输入“Marie Curie”（玛丽·居里）这个查询词，谷歌会在右侧提供了居里夫人的详细信息，如个人简介、出生地点、生卒年月等，甚至还包括一些与居里夫人有关的历史人物，例如爱因斯坦、皮埃尔·居里（居里夫人的丈夫）等。

图1-1 谷歌搜索引擎知识图谱
谷歌知识图谱一出激起千层浪，美国的微软必应，中国的百度、搜狗等搜索引擎公司在短短的一年内纷纷宣布了各自的“知识图谱”产品，如百度“知心“、搜狗“知立方“等。为什么这些搜索引擎巨头纷纷跟进知识图谱，在这上面一掷千金，甚至把它视为搜索引擎的未来呢？这就需要从传统搜索引擎的原理讲起。以百度为例，在过去当我们想知道“泰山”的相关信息的时候，我们会在百度上搜索“泰山”，它会尝试将这个字符串与百度抓取的大规模网页做比对，根据网页与这个查询词的相关程度，以及网页本身的重要性，对网页进行排序，作为搜索结果返回给用户。而用户所需的与“泰山”相关的信息，就还要他们自己动手，去访问这些网页来找了。
当然，与搜索引擎出现之前相比，搜索引擎由于大大缩小了用户查找信息的范围，随着网络信息的爆炸式增长，日益成为人们遨游信息海洋的不可或缺的工具。但是，传统搜索引擎的工作方式表明，它只是机械地比对查询词和网页之间的匹配关系，并没有真正理解用户要查询的到底是什么，远远不够“聪明”，当然经常会被用户嫌弃了。
而知识图谱则会将“泰山”理解为一个“实体”（entity），也就是一个现实世界中的事物。这样，搜索引擎会在搜索结果的右侧显示它的基本资料，例如地理位置、海拔高度、别名，以及百科链接等等，此外甚至还会告诉你一些相关的“实体”，如嵩山、华山、衡山和恒山等其他三山五岳等。当然，用户输入的查询词并不见得只对应一个实体，例如当在谷歌中查询“apple”（苹果）时，谷歌不止展示IT巨头“Apple-Corporation”（苹果公司）的相关信息，还会在其下方列出“apple-plant”（苹果-植物）的另外一种实体的信息。
从杂乱的网页到结构化的实体知识，搜索引擎利用知识图谱能够为用户提供更具条理的信息，甚至

顺着知识图谱可以探索更深入、广泛和完整的知识体系，让用户发现他们意想不到的知识。谷歌高级副总裁艾米特·辛格博士一语道破知识图谱的重要意义所在：“构成这个世界的是实体，而非字符串（things, not strings）”。

很明显，以谷歌为代表的搜索引擎公司希望利用知识图谱为查询词赋予丰富的语义信息，建立与现实世界实体的关系，从而帮助用户更快找到所需的信息。谷歌知识图谱不仅从 Freebase和维基百科等知识库中获取专业信息，同时还通过分析大规模网页内容抽取知识。现在谷歌的这幅知识图谱已经将5亿个实体编织其中，建立了35 亿个属性和相互关系，并在不断高速扩充。

谷歌知识图谱正在不断融入其各大产品中服务广大用户。最近，谷歌在Google Play Store的Google Play Movies & TV应用中添加了一个新的功能，当用户使用安卓系统观看视频时，暂停播放，视频旁边就会自动弹出该屏幕上人物或者配乐的信息。这些信息就是来自谷歌知识图谱。谷歌会圈出播放器窗口所有人物的脸部，用户可以点击每一个人物的脸来查看相关信息。此前，Google Books 已经应用此功能。

图1-2 Google利用知识图谱标示视频中的人物和音乐信息

2 知识图谱的构建

最初知识图谱是谷歌推出的产品名称，与Facebook提出的社交图谱（Social Graph）异曲同工。由于其表意形象，现在知识图谱已经被用来泛指各种大规模知识库。

我们应当如何构建知识图谱呢？首先，我们先了解一下，知识图谱的数据来源都有哪些。知识图谱的最重要的数据来源之一是以维基百科、百度百科为代表的大规模知识库，在这些由网民协同编辑构建的知识库中，包含了大量结构化的知识，可以高效地转化到知识图谱中。此外，互联网的海量网页中也蕴藏了海量知识，虽然相对知识库而言这些知识更显杂乱，但通过自动化技术，也可以将其抽取出来构建知识图谱。接下来，我们分别详细介绍这些识图谱数据来源。

2.1 大规模知识库

大规模知识库以词条作为基本组织单位，每个词条对应现实世界的某个概念，由世界各地的编辑者义务协同编纂内容。随着互联网的普及和Web 2.0理念深入人心，这类协同构建的知识库，无论是数量、质量还是更新速度，都早已超越传统由专家编辑的百科全书，成为人们获取知识的主要来源之一。目前，维基百科已经收录了超过2200万词条，而仅英文版就收录了超过400万条，远超过英文百科全书中最权威的大英百科全书的50万条，是全球浏览人数排名第6的网站。值得一提的是，2012年大英百科全书宣布停止印刷版发行，全面转向电子化。这也从一个侧面说明在线大规模知识库的影响力。人们在知识库中贡献了大量结构化的知识。如下图所示，是维基百科关于“清华大学”的词条内容。可以看到，在右侧有一个列表，标注了与清华有关的各类重要信息，如校训、创建时间、校庆日、学校类型、校长，等等。在维基百科中，这个列表被称为信息框（infobox），是由编辑者们共同编辑而成。信息框中的结构化信息是知识图谱的直接数据来源。

除了维基百科等大规模在线百科外，各大搜索引擎公司和机构还维护和发布了其他各类大规模知识库，例如谷歌收购的Freebase，包含3900万个实体和18亿条实体关系；DBpedia是德国莱比锡大学等机构发起的项目，从维基百科中抽取实体关系，包括1千万个实体和14亿条实体关系；YAGO则是

德国马克斯·普朗克研究所发起的项目，也是从维基百科和WordNet等知识库中抽取实体，到2010年该项目已包含1千万个实体和1.2亿条实体关系。此外，在众多专门领域还有领域专家整理的领域知识库。

图2-1 维基百科词条“清华大学”部分内容

2.2 互联网链接数据

国际万维网组织W3C在2007年发起了开放互联数据项目（Linked Open Data，LOD）。该项目旨在将由互联文档组成的万维网（Web of documents）扩展成由互联数据组成的知识空间（Web of data）。LOD以RDF（Resource Description Framework）形式在Web上发布各种开放数据集，RDF是一种描述结构化知识的框架，它将实体间的关系表示为 (实体1, 关系, 实体2) 的三元组。LOD还允许在不同来源的数据项之间设置RDF链接，实现语义Web知识库。目前世界各机构已经基于LOD标准发布了数千个数据集，包含数千亿RDF三元组。随着LOD项目的推广和发展，互联网会有越来越多的信息以链接数据形式发布，然而各机构发布的链接数据之间存在严重的异构和冗余等问题，如何实现多数据源的知识融合，是LOD项目面临的重要问题。

图2-2 开放互联数据项目发布数据集示意图

2.3 互联网网页文本数据

与整个互联网相比，维基百科等知识库仍只能算沧海一粟。因此，人们还需要从海量互联网网页中直接抽取知识。与上述知识库的构建方式不同，很多研究者致力于直接从无结构的互联网网页中抽取结构化信息，如华盛顿大学Oren Etzioni教授主导的“开放信息抽取”（open information extraction，OpenIE）项目，以及卡耐基梅隆大学Tom Mitchell教授主导的“永不停止的语言学习”（never-ending language learning， NELL）项目。OpenIE项目所开发的演示系统TextRunner已经从1亿个网页中抽取出了5亿条事实，而NELL项目也抽取了超过5千万条事实。

显而易见，与从维基百科中抽取的知识库相比，开放信息抽取从无结构网页中抽取的信息准确率还很低，其主要原因在于网页形式多样，噪音信息较多，信息可信度较低。因此，也有一些研究者尝试限制抽取的范围，例如只从网页表格等内容中抽取结构信息，并利用互联网的多个来源互相印证，从而大大提高抽取信息的可信度和准确率。当然这种做法也会大大降低抽取信息的覆盖面。天下没有免费的午餐，在大数据时代，我们需要在规模和质量之间寻找一个最佳的平衡点。

2.4 多数据源的知识融合

从以上数据来源进行知识图谱构建并非孤立进行。在商用知识图谱构建过程中，需要实现多数据源的知识融合。以谷歌最新发布的Knowledge Vault (Dong, et al. 2014)技术为例，其知识图谱的数据来源包括了文本、DOM Trees、HTML表格、RDF语义数据等多个来源。多来源数据的融合，能够更有效地判定抽取知识的可信性。

知识融合主要包括实体融合、关系融合和实例融合。对于实体，人名、地名、机构名往往有多个名称。例如“中国移动通信集团公司”有“中国移动”、“中移动”、“移动通信”等名称。我们需要将这些不同名称规约到同一个实体下。同一个实体在不同语言、不同国家和地区往往会有不同命名，例如著名足球明星Beckham在大陆汉语中称作“贝克汉姆”，在香港译作“碧咸”，而在台湾则被称为“贝克汉”。与此对应的，同一个名字在不同语境下可能会对应不同实体，这是典型的一词多义问题，例如“苹果”有时是指一种水果，有时则指的是一家著名IT公司。在这样复杂的多对多对应关系中，如何实现实体融合是非常复杂而重要的课题。如前面开放信息抽取所述，同一种关系可能会有不同的命名，这种现象在不同数据源中抽取出的关系中尤其显著。与实体融合类似，关系融合对于知识融合至关重要。在实现了实体和关系融合之后，我们就可以实现三元组实例的融合。不同数据源会抽取出相同的三元组，并给出不同的评分。根据这些评分，以及不同数据源的可信度，我们就可以实现三元组实例的融合与抽取。

知识融合既有重要的研究挑战，又需要丰富的工程经验。知识融合是实现大规模知识图谱的必由之路。知识融合的好坏，往往决定了知识图谱项目的成功与否，值得任何有志于大规模知识图谱构建与应用的人士高度重视。

3 知识图谱的典型应用

知识图谱将搜索引擎从字符串匹配推进到实体层面，可以极大地改进搜索效率和效果，为下一代搜索引擎的形态提供了巨大的想象空间。知识图谱的应用前景远不止于此，目前知识图谱已经被广泛应用于以下几个任务中。

3.1 查询理解（Query Understanding）

谷歌等搜索引擎巨头之所以致力于构建大规模知识图谱，其重要目标之一就是能够更好地理解用户输入的查询词。用户查询词是典型的短文本（short text），一个查询词往往仅由几个关键词构成。传统的关键词匹配技术没有理解查询词背后的语义信息，查询效果可能会很差。

例如，对于查询词“李娜 大满贯”，如果仅用关键词匹配的方式，搜索引擎根本不懂用户到底希望寻找哪个“李娜”，而只会机械地返回所有含有“李娜”这个关键词的网页。但通过利用知识图谱识别查询词中的实体及其属性，搜索引擎将能够更好地理解用户搜索意图。现在，我们到谷歌中查询“李娜 大满贯”，会发现，首先谷歌会利用知识图谱在页面右侧呈现中国网球运动员李娜的基本信息，我们可以知道这个李娜是指的中国网球女运动员。同时，谷歌不仅像传统搜索引擎那样返回匹配的网页，更会直接在页面最顶端返回李娜赢得大满贯的次数“2”。

图3-1 谷歌中对“李娜 大满贯”的查询结果

主流商用搜索引擎基本都支持这种直接返回查询结果而非网页的功能，这背后都离不开大规模知识图谱的支持。以百度为例，下图是百度中对“珠穆朗玛峰高度”的查询结果，百度直接告诉用户珠穆朗玛峰的高度是8844.43米。

图3-2 百度中对“珠穆朗玛峰高度”的查询结果

基于知识图谱，搜索引擎还能获得简单的推理能力。例如，下图是百度中对“梁启超的儿子的妻子”的查询结果，百度能够利用知识图谱知道梁启超的儿子是梁思成，梁思成的妻子是林徽因等人。

图3-3 百度中对“梁启超的儿子的妻子”的查询结果

采用知识图谱理解查询意图，不仅可以返回更符合用户需求的查询结果，还能更好地匹配商业广告信息，提高广告点击率，增加搜索引擎受益。因此，知识图谱对搜索引擎公司而言，是一举多得的重要资源和技术。

3.2 自动问答（Question Answering）

人们一直在探索比关键词查询更高效的互联网搜索方式。很多学者预测，下一代搜索引擎将能够直接回答人们提出的问题，这种形式被称为自动问答。例如著名计算机学者、美国华盛顿大学计算机科学与工程系教授、图灵中心主任Oren Etzioni于2011年就在Nature杂志上发表文章“搜索需要一场变革“（Search Needs a Shake-Up）。该文指出，一个可以理解用户问题，从网络信息中抽取事实，并最终选出一个合适答案的搜索引擎，才能将我们带到信息获取的制高点。如上节所述，目前搜索引擎已经支持对很多查询直接返回精确答案而非海量网页而已。

关于自动问答，我们将有专门的章节介绍。这里，我们需要着重指出的是，知识图谱的重要应用之一就是作为自动问答的知识库。在搜狗推出中文知识图谱服务”知立方“的时候，曾经以回答”梁启超的儿子的太太的情人的父亲是谁？“这种近似脑筋急转弯似的问题作为案例，来展示其知识图谱的强大推理能力。虽然大部分用户不会这样拐弯抹角的提问，但人们会经常需要寻找诸如”刘德华的妻子是谁？“、”侏罗纪公园的主演是谁？“、“姚明的身高？”以及”北京有几个区？“等问题的答案。而这些问题都需要利用知识图谱中实体的复杂关系推理得到。无论是理解用户查询意图，还是探索新的搜索形式，都毫无例外需要进行语义理解和知识推理，而这都需要大规模、结构化的知识图谱的有力支持，因此知识图谱成为各大互联网公司的必争之地。

最近，微软联合创始人Paul Allen投资创建了艾伦人工智能研究院（Allen Institute for Artificial Intelligence），致力于建立具有学习、推理和阅读能力的智能系统。2013年底，Paul Allen任命Oren Etzioni教授担任艾伦人工智能研究院的执行主任，该任命所释放的信号颇值得我们思考。

3.3 文档表示（Document Representation）

经典的文档表示方案是空间向量模型（Vector Space Model），该模型将文档表示为词汇的向量，而且采用了词袋（Bag-of-Words，BOW）假设，不考虑文档中词汇的顺序信息。这种文档表示方案与上述的基于关键词匹配的搜索方案相匹配，由于其表示简单，效率较高，是目前主流搜索引擎所采用的技术。文档表示是自然语言处理很多任务的基础，如文档分类、文档摘要、关键词抽取，等等。

经典文档表示方案已经在实际应用中暴露出很多固有的严重缺陷，例如无法考虑词汇之间的复杂语义关系，无法处理对短文本（如查询词）的稀疏问题。人们一直在尝试解决这些问题，而知识图谱的出现和发展，为文档表示带来新的希望，那就是基于知识的文档表示方案。一篇文章不再只是由一组代表词汇的字符串来表示，而是由文章中的实体及其复杂语义关系来表示(Schuhmacher, et al.

2014)。该文档表示方案实现了对文档的深度语义表示，为文档深度理解打下基础。一种最简单的基于知识图谱的文档表示方案，可以将文档表示为知识图谱的一个子图（sub-graph），即用该文档中出现或涉及的实体及其关系所构成的图表示该文档。这种知识图谱的子图比词汇向量拥有更丰富的表示空间，也为文档分类、文档摘要和关键词抽取等应用提供了更丰富的可供计算和比较的信息。

知识图谱为计算机智能信息处理提供了巨大的知识储备和支持，将让现在的技术从基于字符串匹配的层次提升至知识理解层次。以上介绍的几个应用可以说只能窥豹一斑。知识图谱的构建与应用是一个庞大的系统工程，其所蕴藏的潜力和可能的应用，将伴随着相关技术的日渐成熟而不断涌现。

4 知识图谱的主要技术

大规模知识图谱的构建与应用需要多种智能信息处理技术的支持，以下简单介绍其中若干主要技术。

4.1 实体链指（Entity Linking）

互联网网页，如新闻、博客等内容里涉及大量实体。大部分网页本身并没有关于这些实体的相关说明和背景介绍。为了帮助人们更好地了解网页内容，很多网站或作者会把网页中出现的实体链接到相应的知识库词条上，为读者提供更详尽的背景材料。这种做法实际上将互联网网页与实体之间建立了链接关系，因此被称为实体链指。

手工建立实体链接关系非常费力，因此如何让计算机自动实现实体链指，成为知识图谱得到大规模应用的重要技术前提。例如，谷歌等在搜索引擎结果页面呈现知识图谱时，需要该技术自动识别用户输入查询词中的实体并链接到知识图谱的相应节点上。

实体链指的主要任务有两个，实体识别（Entity Recognition）与实体消歧（Entity Disambiguation），都是自然语言处理领域的经典问题。

实体识别旨在从文本中发现命名实体，最典型的包括人名、地名、机构名等三类实体。近年来，人们开始尝试识别更丰富的实体类型，如电影名、产品名，等等。此外，由于知识图谱不仅涉及实体，还有大量概念（concept），因此也有研究者提出对这些概念进行识别。

不同环境下的同一个实体名称可能会对应不同实体，例如“苹果”可能指某种水果，某个著名IT公司，也可能是一部电影。这种一词多义或者歧义问题普遍存在于自然语言中。将文档中出现的名字链接到特定实体上，就是一个消歧的过程。消歧的基本思想是充分利用名字出现的上下文，分析不同实体可能出现在该处的概率。例如某个文档如果出现了iphone，那么”苹果“就有更高的概率指向知识图谱中的叫”苹果“的IT公司。

实体链指并不局限于文本与实体之间，如下图所示，还可以包括图像、社交媒体等数据与实体之间的关联。可以看到，实体链指是知识图谱构建与应用的基础核心技术。

图4-1 实体链指实现实体与文本、图像、社交媒体等数据的关联

4.2 关系抽取（Relation Extraction）

构建知识图谱的重要来源之一是从互联网网页文本中抽取实体关系。关系抽取是一种典型的信息抽取任务。

典型的开放信息抽取方法采用自举（bootstrapping）的思想，按照“模板生成实例抽取”的流程不断迭

代直至收敛。例如，最初可以通过“X是Y的首都”模板抽取出(中国，首都，北京)、(美国，首都，华盛顿)等三元组实例；然后根据这些三元组中的实体对“中国-北京”和“美国-华盛顿”可以发现更多的匹配模板，如“Y的首都是X”、“X是Y的政治中心”等等；进而用新发现的模板抽取更多新的三元组实例，通过反复迭代不断抽取新的实例与模板。这种方法直观有效，但也面临很多挑战性问题，如在扩展过程中很容易引入噪音实例与模板，出现语义漂移现象，降低抽取准确率。研究者针对这一问题提出了很多解决方案：提出同时扩展多个互斥类别的知识，例如同时扩展人物、地点和机构，要求一个实体只能属于一个类别；也有研究提出引入负实例来限制语义漂移。

我们还可以通过识别表达语义关系的短语来抽取实体间关系。例如，我们通过句法分析，可以从文本中发现“华为”与“深圳”的如下关系：(华为，总部位于，深圳)、(华为，总部设置于，深圳)、以及(华为，将其总部建于，深圳)。通过这种方法抽取出的实体间关系非常丰富而自由，一般是一个以动词为核心的短语。该方法的优点是，我们无需预先人工定义关系的种类，但这种自由度带来的代价是，关系语义没有归一化，同一种关系可能会有多种不同的表示。例如，上述发现的“总部位于”、“总部设置于”以及“将其总部建于”等三个关系实际上是同一种关系。如何对这些自动发现的关系进行聚类规约是一个挑战性问题。

我们还可以将所有关系看做分类标签，把关系抽取转换为对实体对的关系分类问题。这种关系抽取方案的主要挑战在于缺乏标注语料。2009年斯坦福大学研究者提出远程监督（Distant Supervision）思想，使用知识图谱中已有的三元组实例启发式地标注训练语料。远程监督思想的假设是，每个同时包含两个实体的句子，都表述了这两个实体在知识库中的对应关系。例如，根据知识图谱中的三元组实例(苹果，创始人，乔布斯)和(苹果，CEO，库克)，我们可以将以下四个包含对应实体对的句子分别标注为包含“创始人”和“CEO”关系：

样例句子关系/分类标签

苹果-乔布斯苹果公司的创始人是乔布斯。创始人

苹果-乔布斯乔布斯创立了苹果公司。创始人

苹果-库克苹果公司的CEO是库克。CEO

苹果-库克库克现在是苹果公司的CEO。CEO

我们将知识图谱三元组中每个实体对看做待分类样例，将知识图谱中实体对关系看做分类标签。通过从出现该实体对的所有句子中抽取特征，我们可以利用机器学习分类模型（如最大熵分类器、SVM等）构建信息抽取系统。对于任何新的实体对，根据所出现该实体对的句子中抽取的特征，我们就可以利用该信息抽取系统自动判断其关系。远程监督能够根据知识图谱自动构建大规模标注语料库，因此取得了瞩目的信息抽取效果。

与自举思想面临的挑战类似，远程监督方法会引入大量噪音训练样例，严重损害模型准确率。例如，对于(苹果，创始人，乔布斯)我们可以从文本中匹配以下四个句子：

句子关系/分类标签是否正确

苹果公司的创始人是乔布斯。创始人正确

乔布斯创立了苹果公司。创始人正确

乔布斯回到了苹果公司。创始人错误

乔布斯曾担任苹果的CEO。创始人错误

在这四个句子中，前两个句子的确表明苹果与乔布斯之间的创始人关系；但是，后两个句子则并没

有表达这样的关系。很明显，由于远程监督只能机械地匹配出现实体对的句子，因此会大量引入错误训练样例。为了解决这个问题，人们提出很多去除噪音实例的办法，来提升远程监督性能。例如，研究发现，一个正确训练实例往往位于语义一致的区域，也就是其周边的实例应当拥有相同的关系；也有研究提出利用因子图、矩阵分解等方法，建立数据内部的关联关系，有效实现降低噪音的目标。

关系抽取是知识图谱构建的核心技术，它决定了知识图谱中知识的规模和质量。关系抽取是知识图谱研究的热点问题，还有很多挑战性问题需要解决，包括提升从高噪音的互联网数据中抽取关系的鲁棒性，扩大抽取关系的类型与抽取知识的覆盖面，等等。

4.3 知识推理（Knowledge Reasoning）

推理能力是人类智能的重要特征，能够从已有知识中发现隐含知识。推理往往需要相关规则的支持，例如从“配偶”+“男性”推理出“丈夫”，从“妻子的父亲”推理出“岳父”，从出生日期和当前时间推理出年龄，等等。

这些规则可以通过人们手动总结构建，但往往费时费力，人们也很难穷举复杂关系图谱中的所有推理规则。因此，很多人研究如何自动挖掘相关推理规则或模式。目前主要依赖关系之间的同现情况，利用关联挖掘技术来自动发现推理规则。

实体关系之间存在丰富的同现信息。如下图，在康熙、雍正和乾隆三个人物之间，我们有(康熙，父亲，雍正)、(雍正，父亲，乾隆)以及(康熙，祖父，乾隆)三个实例。根据大量类似的实体X、Y、Z间出现的(X，父亲，Y)、(Y，父亲，Z)以及(X，祖父，Z)实例，我们可以统计出“父亲+父亲=>祖父”的推理规则。类似的，我们还可以根据大量(X，首都，Y)和(X，位于，Y)实例统计出“首都=>位于”的推理规则，根据大量(X，总统，美国)和(X，是，美国人)统计出“美国总统=>是美国人”的推理规则。

图4-2 知识推理举例

知识推理可以用于发现实体间新的关系。例如，根据“父亲+父亲=>祖父”的推理规则，如果两实体间存在“父亲+父亲”的关系路径，我们就可以推理它们之间存在“祖父”的关系。利用推理规则实现关系抽取的经典方法是Path Ranking Algorithm (Lao & Cohen 2010)，该方法将每种不同的关系路径作为一维特征，通过在知识图谱中统计大量的关系路径构建关系分类的特征向量，建立关系分类器进行关系抽取，取得不错的抽取效果，成为近年来的关系抽取的代表方法之一。但这种基于关系的同现统计的方法，面临严重的数据稀疏问题。

在知识推理方面还有很多的探索工作，例如采用谓词逻辑（Predicate Logic）等形式化方法和马尔科夫逻辑网络（Markov Logic Network）等建模工具进行知识推理研究。目前来看，这方面研究仍处于百家争鸣阶段，大家在推理表示等诸多方面仍为达成共识，未来路径有待进一步探索。

4.4 知识表示（Knowledge Representation）

在计算机中如何对知识图谱进行表示与存储，是知识图谱构建与应用的重要课题。

如“知识图谱”字面所表示的含义，人们往往将知识图谱作为复杂网络进行存储，这个网络的每个节点带有实体标签，而每条边带有关系标签。基于这种网络的表示方案，知识图谱的相关应用任务往往需要借助于图算法来完成。例如，当我们尝试计算两实体之间的语义相关度时，我们可以通过它们在网络中的最短路径长度来衡量，两个实体距离越近，则越相关。而面向“梁启超的儿子的妻子”这样的推理查询问题时，则可以从“梁启超”节点出发，通过寻找特定的关系路径“梁启超->儿子->妻子->?”，来找到答案。

然而，这种基于网络的表示方法面临很多困难。首先，该表示方法面临严重的数据稀疏问题，对于那些对外连接较少的实体，一些图方法可能束手无策或效果不佳。此外，图算法往往计算复杂度较高，无法适应大规模知识图谱的应用需求。

最近，伴随着深度学习和表示学习的革命性发展，研究者也开始探索面向知识图谱的表示学习方案。其基本思想是，将知识图谱中的实体和关系的语义信息用低维向量表示，这种分布式表示（Distributed Representation）方案能够极大地帮助基于网络的表示方案。其中，最简单有效的模型是最近提出的TransE(Bordes, et al. 2013)。TransE基于实体和关系的分布式向量表示，将每个三元组实例（head，relation，tail）中的关系relation看做从实体head到实体tail的翻译，通过不断调整h、r和t（head、relation和tail的向量），使（h + r）尽可能与 t 相等，即 h + r = t。该优化目标如下图所示。

通过TransE等模型学习得到的实体和关系向量，能够很大程度上缓解基于网络表示方案的稀疏性问题，应用于很多重要任务中。

首先，利用分布式向量，我们可以通过欧氏距离或余弦距离等方式，很容易地计算实体间、关系间的语义相关度。这将极大的改进开放信息抽取中实体融合和关系融合的性能。通过寻找给定实体的相似实体，还可用于查询扩展和查询理解等应用。

其次，知识表示向量可以用于关系抽取。以TransE为例，由于我们的优化目标是让 h + r = t，因此，当给定两个实体 h 和 t 的时候，我们可以通过寻找与 t – h 最相似的 r，来寻找两实体间的关系。(Bordes, et al. 2013)中的实验证明，该方法的抽取性能较高。而且我们可以发现，该方法仅需要知识图谱作为训练数据，不需要外部的文本数据，因此这又称为知识图谱补全（Knowledge Graph Completion），与复杂网络中的链接预测（Link Prediction）类似，但是要复杂得多，因为在知识图谱中每个节点和连边上都有标签（标记实体名和关系名）。

最后，知识表示向量还可以用于发现关系间的推理规则。例如，对于大量X、Y、Z间出现的(X，父亲，Y)、(Y，父亲，Z)以及(X，祖父，Z)实例，我们在TransE中会学习X+父亲=Y，Y+父亲=Z，以及X+祖父=Z等目标。根据前两个等式，我们很容易得到X+父亲+父亲=Z，与第三个公式相比，就能够得到“父亲+父亲=>祖父”的推理规则。前面我们介绍过，基于关系的同现统计学习推理规则的思想，存在严重的数据稀疏问题。如果利用关系向量表示提供辅助，可以显著缓解稀疏问题。

5 前景与挑战

如果未来的智能机器拥有一个大脑，知识图谱就是这个大脑中的知识库，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。

现在以商业搜索引擎公司为首的互联网巨头已经意识到知识图谱的战略意义，纷纷投入重兵布局知识图谱，并对搜索引擎形态日益产生重要的影响。同时，我们也强烈地感受到，知识图谱还处于发展初期，大多数商业知识图谱的应用场景非常有限，例如搜狗知立方更多聚焦在娱乐和健康等领域。根据各搜索引擎公司提供的报告来看，为了保证知识图谱的准确率，仍然需要在知识图谱构建过程中采用较多的人工干预。

可以看到，在未来的一段时间内，知识图谱将是大数据智能的前沿研究问题，有很多重要的开放性

问题亟待学术界和产业界协力解决。我们认为，未来知识图谱研究有以下几个重要挑战。

知识类型与表示。知识图谱主要采用(实体1,关系,实体2)三元组的形式来表示知识，这种方法可以较好的表示很多事实性知识。然而，人类知识类型多样，面对很多复杂知识，三元组就束手无策了。例如，人们的购物记录信息，新闻事件等，包含大量实体及其之间的复杂关系，更不用说人类大量的涉及主观感受、主观情感和模糊的知识了。有很多学者针对不同场景设计不同的知识表示方法。知识表示是知识图谱构建与应用的基础，如何合理设计表示方案，更好地涵盖人类不同类型的知识，是知识图谱的重要研究问题。最近认知领域关于人类知识类型的探索(Tenenbaum, et al. 2011)也许会对知识表示研究有一定启发作用。

知识获取。如何从互联网大数据萃取知识，是构建知识图谱的重要问题。目前已经提出各种知识获取方案，并已经成功抽取大量有用的知识。但在抽取知识的准确率、覆盖率和效率等方面，都仍不如人意，有极大的提升空间。

知识融合。来自不同数据的抽取知识可能存在大量噪音和冗余，或者使用了不同的语言。如何将这些知识有机融合起来，建立更大规模的知识图谱，是实现大数据智能的必由之路。

知识应用。目前大规模知识图谱的应用场景和方式还比较有限，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。这既需要丰富的知识图谱技术积累，也需要对人类需求的敏锐感知，找到合适的应用之道。

6 内容回顾与推荐阅读

本章系统地介绍了知识图谱的产生背景、数据来源、应用场景和主要技术。通过本章我们主要有以下结论：

知识图谱是下一代搜索引擎、自动问答等智能应用的基础设施。

互联网大数据是知识图谱的重要数据来源。

知识表示是知识图谱构建与应用的基础技术。

实体链指、关系抽取和知识推理是知识图谱构建与应用的核心技术。

知识图谱与本体（Ontology）和语义网（Semantic Web）等密切相关，有兴趣的读者可以搜索与之相关的文献阅读。知识表示（Knowledge Representation）是人工智能的重要课题，读者可以通过人工智能专著(Russell & Norvig 2009)了解其发展历程。在关系抽取方面，读者可以阅读(Nauseates, et al. 2013)、(Nickel, et al. 2015)详细了解相关技术。

参考文献

(Bordes, et al. 2013) Bordes, A., Usunier, N., Garcia-Duran, A., Weston, J., & Yakhnenko, O. (2013). Translating embeddings for modeling multi-relational data. In Proceedings of NIPS.

(Dong, et al. 2014) Dong, X., Gabrilovich, E., Heitz, G., Horn, W., et al. Knowledge Vault A web-scale approach to probabilistic knowledge fusion. In Proceedings of KDD.

(Lao & Cohen 2010) Lao, N., & Cohen, W. W. (2010). Relational retrieval using a combination of path-constrained random walks. Machine learning, 81(1), 53-67.

(Nauseates, et al. 2013) Nastase, V., Nakov, P., Seaghdha, D. O., & Szpakowicz, S. (2013). Semantic relations between nominals. Synthesis Lectures on Human Language Technologies, 6(1), 1-119.

(Nickel, et al. 2015) Nickel, M., Murphy, K., Tresp, V., & Gabrilovich, E. A Review of Relational Machine Learning for Knowledge Graphs.
(Russell & Norvig 2009) Russell, S., & Norvig, P. (2009). Artificial Intelligence: A Modern Approach, 3rd Edition. Pearson Press.（中文译名：人工智能——一种现代方法）.
(Schuhmacher, et al. 2014) Schuhmacher, M., & Ponzetto, S. P. Knowledge-based graph document modeling. In Proceedings of the 7th ACM international conference on Web search and data mining. In Proceedings of WSDM.
(Tenenbaum, et al. 2011) Tenenbaum, J. B., Kemp, C., Griffiths, T. L., & Goodman, N. D. (2011). How to grow a mind: Statistics, structure, and abstraction. science, 331(6022), 1279-1285