《智能信息处理》课程考试

多模态知识图谱的研究与构建

齐海霞

考核	到课[10]	作业[20]	考试[70]	课程成绩[100]
得分				

2021 年 12 月 1 日

多模态知识图谱的研究与构建

摘要：为了总结前人工作，给相关研究者提供思路，首先讨论了当前多模态知识图谱的基本概念，然后从图数据库和知识图谱这两个角度介绍了多模态知识图谱的构建工作，并总结了两种主要方法的思路。还分析了多模态知识图谱的构建和应用中的关键技术和相关工作，如多模态信息提取、表示学习和实体链接。此外，列举了多模态知识图谱在四种场景中的应用，包括推荐系统、跨模态检索、人机交互。

关键词：多模态；知识图谱；构建

Abstract: In order to summarize the previous work and provide ideas for relevant researchers, this paper first discusses the basic concepts of current multimodal knowledge atlas, then introduces the construction of multimodal knowledge atlas from the perspectives of graph database and knowledge atlas, and summarizes the ideas of the two main methods. The key technologies and related work in the construction and application of multimodal knowledge map are also analyzed, such as multimodal information extraction, representation learning and entity link. In addition, the applications of multimodal knowledge atlas in four scenarios are listed, including recommendation system, cross modal retrieval, human-computer interaction and cross modal data management.

Keywords: Multimodal; Knowledge atlas; Structure

1 知识图谱

知识图谱(Knowledge Graph)，在图书情报界称为知识域可视化或知识领域映射地图，是显示知识发展进程与结构关系的一系列各种不同的图形，用可视化技术描述知识资源及其载体，挖掘、分析、构建、绘制和显示知识及它们之间的相互联系。

知识图谱是通过将应用数学、图形学、信息可视化技术、信息科学等学科的理论与方法与计量学引文分析、共现分析等方法结合，并利用可视化的图谱形象地展示学科的核心结构、发展历史、前沿领域以及整体知识架构达到多学科融合目的的现代理论。它能为学科研究提供切实的、有价值的参考。

知识图谱从语义角度出发，以事实三元组的形式描述客观世界中概念、实体及其关系，将实体和概念抽象为节点，将关系抽象为边，通过结构化的形式对知识进行建模。三元组由头实体、尾实体和描述它们之间的关系组成，如〈张三，国籍，中国〉构建一个三元组实例。知识图谱用本体[1](ontology)对概念和关系进行形式化描述，知识的本体框架和三元组实例共同构成完整的知识图谱，并用资源描述框架(resource description frame-work， RDF)进行统一表示。

文献指出，知识图谱构建一般采用自动或半自动技术从结构化、半结构化以及非结构化数据资源中抽取知识，并存入基于逻辑划分的数据层和模式层，是一个迭代更新的过程，主要包含信息抽取、知识融合、知识加工三个阶段。利用自然语言处理、机器学习等技术从多源异构的数据资源中自动构建知识图谱的技术取得长足进展，例如华盛顿大学的 TestRunner[2]、OLLIE[3]以及卡耐基梅隆大学的 NELL[4]。

当前知识图谱被广泛用于处理文本数据，对于半结构化或非结构化的图像、音频、视频等多模态数据的关注度则较低。作为一种知识表示、存储的手段，其可推理、可解释性在图像识别、图像分类中有较好表现。一方面，知识图谱可以提高视觉识别未知类的性能; 另一方面，视觉信息可以用来扩展知识图谱，两者相辅相成。多模态数据的涌现使跨模态语义理解与知识表示需求变得更加迫切，作为承载底层海量知识并支持上层智能应用的重要载体，知识图谱也急需多模态化。

2 多模态学习

每一种信息的来源或者形式，都可以称为一种模态。例如，人有触觉，听觉，视觉，嗅觉；信息的媒介，有语音、视频、文字等；多种多样的传感器，如雷达、红外、加速度计等。以上的每一种都可以称为一种模态。

同时，模态也可以有非常广泛的定义，比如我们可以把两种不同的语言当做是两种模态，甚至在两种不同情况下采集到的数据集，亦可认为是两种模态。

因此，多模态机器学习，英文全称 MultiModal Machine Learning (MMML)[5]，旨在通过机器学习的方法实现处理和理解多源模态信息的能力。目前比较热门的研究方向是图像、视频、音频、语义之间的多模态学习。多模态学习从 1970 年代起步，经历了几个发展阶段，在 2010 后全面步入 Deep Learning 阶段。人其实是一个多模态学习的总和，所以也有专家说了，多模态学习才是真正的人工智能发展方向。

从语义感知角度理解，一个客观实体可以被视觉、听觉、触觉等不同模态感知; 从数据层面理解，同一实体可以有图片、文本、语音等数据记录。多模态起源于计算机人机交互领域信息表示方式的研究，模态信息存储在多模态数据中。本文讨论的多模态数据可理解为描述同一对象的多媒体数据，多模态数据虽然在底层表征上是异构的，但是相同实体的不同模态数据在高层语义上是一致的。让人工智能更贴近人类对客观世界的认知，实现对多模态数据环境的理解，需要其具备解释多模态数据的能力。多模态数据之间由于其本身结构特点，其技术研究主要面临两大挑战: a) 语义鸿沟(semantic gap) ，指计算机表示系统与人类认知系统对同一个概念形成不同描述的差异，例如对于图像的像素信息、颜色、形状等人类认知中直观的语义表现在计算机视觉的表达中就需要借助复杂的数学形式化方法，利用参数组合对颜色、形状的概念进行编码表示; b)异构鸿沟(heterogeneity gap) ，指图像、文本等不同媒体的数据具有不同的特征表示形式，它们的相似性难以直接度量。

对多源异构数据的挖掘分析可被理解为多模态学习，其任务是通过学习多个模态数据中的信息，实现各个模态信息的转换和交流。不同模态之间的数据在进行综合建模时就会面临语义鸿沟和异构鸿沟带来的问题。随着深度学习的方法在获取自然语言、视觉、听觉等单模态表示上已经取得了较优的效果，多模态学习也已经进入多模态深度学习阶段。把不同媒体的数据从各自独立的空间映射到一个第三方的公共空间中进行相似性度量是一种直观的方法。现有的研究大多仅考虑两个模态数据，当同时面临三个或更多模态时，对于公共空间的寻找将面临一定的困难。

结合文献中对多模态深度学习的综述，多模态深度表示学习是多模态深度学习的一个重要研究方向，主要作用是用深度学习的方法将多模态数据在同一高层语义表示空间进行对齐，以便进行对齐、比较和融合。如图 1 为多模态深度表示学习一般框架，通常用合适的神经网络学习文本、图像、音频、视频等多模态数据在相应特征空间的表示，随后将各模态的表示作为输入，继续构建更深层的神经网络结构，利用深度跨模态表示学习构建的神经网络融合各模态的语义信息得到在共同表示空间中各模态的高层语义表示。

多模态数据还有数据量大、数据分布稀疏等特点。相较于单模态的研究，多模态数据集的构建更为困难，指代同一实体的不同模态数据通常需要昂贵的人工标注，大规模的多模态研究面临着训练数据缺失的难题，已有研究通过跨模态知识迁移和预训练来避免对标注数据的过高依赖。

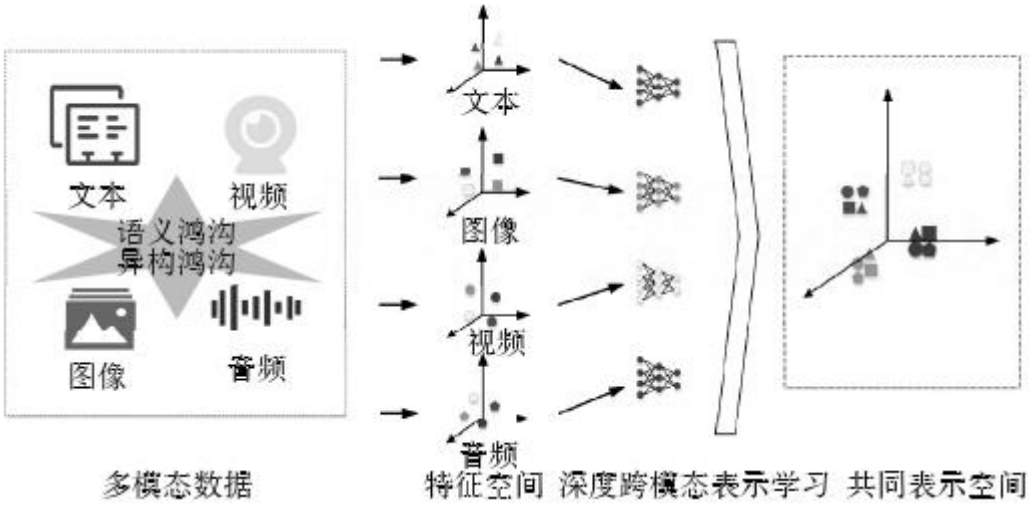

图 1 多模态深度表示学习一般框架

3 多模态知识图谱

知识图谱技术已经被广泛用于处理结构化数据（采用本体+D2R 技术）和文本数据（采用文本信息抽取技术），但是还有一类非结构化数据，即视觉数据，则相对关注度较低，而且相缺乏有效的技术手段来从这些数据中提取结构化知识。最近几年，虽然有一些多模态视觉技术提出，这些技术主要还是为了提升图像分类、图像生成、图像问答的效果，不能很好地支撑多模态知识图谱的构建。视觉数据库通常是图像或视频数据的丰富来源，并提供关于知识图谱中实体的充分视觉信息。显然，如果可以在在更大范围内进行链接预测和实体对齐，进而进行实体关系抽取，可以使现有的模型在综合考虑文本和视觉特征时获得更好的性能，这也是我们研究多模态知识图谱（multi-modal knowledge graph）的意义所在。

目前，已经有很多开放知识图谱，而且不少企业也有自己的企业知识图谱。然而，这些知识图谱很少有可视化的数据资源。

多模态知识图谱与传统知识图谱的主要区别是，传统知识图谱主要集中研究文本和数据库的实体和关系，而多模态知识图谱则在传统知识图谱的基础上，构建了多种模态（例如视觉模态）下的实体，以及多种模态实体间的多模态语义关系。例如在最新的一个多模态百科图谱 Richpedia[6]中，首先构建了图像模态伦敦眼图像与文本模态知识图谱实体（DBpedia 实体：London eye）之间的多模态语义关系（rpo:imageof），之后还构建了图像模态实体伦敦眼与图像模态实体大本钟之间的多模态语义（rpo:nextTo）。

随着人工智能技术的不断发展，知识图谱作为人工智能领域的知识支柱，以其强大的知识表示和推理能力受到学术界和产业界的广泛关注。近年来，知识图谱在语义搜索、问答、知识管理等领域得到了广泛的应用。其中在描述多媒体的可用数据集中，现有的工作重点是捕获多媒体文件的高级元数据(如作者、创建日期、文件大小、清晰度、持续时间)，而不是多媒体内容本身的音频或视觉特性。以下会介绍几个重要的开源多模态知识图谱。

3.1 DBpedia

DBpedia 作为近十年来语义网研究的中心领域，其丰富的语义信息也将会成为今后多模态知识图谱的链接端点，其完整的本体结构对于构建多模态知识图谱提供了很大的便利。DBpedia 项目是一个社区项目，旨在从维基百科中提取结构化信息，并使其可在网络上访问。DBpedia 知识库目前描述了超过 260 万个实体。对于每个实体，DBpedia 定义了一个唯一的全局标识符，可以将其解引用为网络上一个 RDF 描述的实体。DBpedia 提供了 30 种人类可读的语言版本，与其他资源形成关系。在过去的几年里，越来越多的数据发布者开始建立数据集链接到 DBpedia 资源，使 DBpedia 成为一个新的数据 web 互联中心。目前，围绕 DBpedia 的互联网数据源网络提供了约 47 亿条信息，涵盖地理信息、人、公司、电影、音乐、基因、药物、图书、科技出版社等领域。

3.2 Wikidata

Wikidata 中也存在大量的多模态资源，Wikidata 是维基媒体基金会(WMF)联合策划的一个知识图谱，是维基媒体数据管理策略的核心项目。充分利用 Wikidata 的资源，主要挑战之一是提供可靠并且强大的数据共享查询服务，维基媒体基金会选择使用语义技术。活动的 SPARQL 端点、常规的 RDF 转储和链接的数据 api 是目前 Wikidata 的核心技术，Wikidata 的目标是通过创造维基百科全球管理数据的新方法来克服数据不一致性。Wikidata 的主要成就包括:Wikidata 提供了一个可由所有人共享的免费协作知识库;Wikidata 已经成为维基媒体最活跃的项目之一;越来越多的网站在浏览页面时都从 Wikidata 获取内容，以增加大数据的可见性和实用性。

3.3 IMGpedia

IMGpedia 是一个大型的链接数据集，它从 Wikimedia Commons 数据集中的图像中收集大量的可视化信息。它构建并生成了 1500 万个视觉内容描述符，图像之间有 4.5 亿个视觉相似关系，此外，在 IMGpedia 中单个图像与 DBpedia 之间还有链接。IMGpedia 旨在从维基百科发布的图片中提取相关的视觉信息，从 Wikimedia 中收集所有术语和所有多模态数据(包括作者、日期、大小等)的图像，并为每张图像生成相应的图像描述符。链接数据很少考虑多模态数据，但多模态数据也是语义网络的重要组成部分。为了探索链接数据和多模态数据的结合，构建了 IMGpedia，计算 Wikipedia 条目中使用的图像描述符，然后将这些图像及其描述与百科知识图谱链接起来。

IMGpedia 是一个多模态知识图谱的先例。将语义知识图谱与多模态数据相结合，面对多种任务下的挑战和机遇。IMGpedia 使用四种图像描述符进行基准测试，这些描述符的引用和实现是公开的。IMGpedia 提供了 Wikidata 的链接。由于 DBpedia 中的分类对一些可视化语义查询不方便，所以 IMGpedia 旨在提供一个更好的语义查询平台。IMGpedia 在多模态方向上是一个很好的先例，但也存在一些问题，比如关系类型稀疏，关系数量少，图像分类不清晰等，也是之后需要集中解决的问题。

3.4 MMKG

MMKG 主要用于联合不同知识图谱中的不同实体和图像执行关系推理，MMKG 是一个包含所有实体的数字特征和(链接到)图像的三个知识图谱的集合，以及对知识图谱之间的实体对齐。因此，多关系链接预测和实体匹配社区可以从该资源中受益。MMKG 有潜力促进知识图谱的新型多模态学习方法的发展，作者通过大量的实验验证了 MMKG 在同一链路预测任务中的有效性。

MMKG 选择在知识图谱补全文献中广泛使用的数据集 FREEBASE-15K(FB15K)作为创建多模态知识图谱的起点。知识图谱三元组是基于 N-Triples 格式的，这是一种用于编码 RDF 图的基于行的纯文本格式。MMKG 同时也创建了基于 DBpedia 和 YAGO 的版本，称为 DBpedia-15K(DB15K)和 YAGO15K，通过将 FB15K 中的实体与其他知识图谱中的实体对齐。其中对于基于 DBpedia 的版本，主要构建了 sameAs 关系，为了创建 DB15K，提取了 FB15K 和 DBpedia 实体之间的对齐，通过 sameAs 关系链接 FB15K 和 DBpedia 中的对齐实体；构建关系图谱，来自 FB15K 的很大比例的实体可以与 DBpedia 中的实体对齐。

但是，为了使这两个知识图谱拥有大致相同数量的实体，并且拥有不能跨知识图谱对齐的实体，在 DB15K 中包括了额外的实体；构建图像关系，MMKG 从三大搜索引擎中获取相应文本实体的图像实体，生成对应的文本-图像关系。但是，它是专门为文本知识图谱的完成而构建的，主要针对小数据集(FB15K, DBPEDIA15K, YAGO15K)。MMKG 在将图像分发给相关文本实体时也没有考虑图像的多样性。

4 多模态知识图谱的典型应用

多模态知识图谱技术可以服务于各种场景，例如多模态实体链接技术可以融合多种模态下的相同实体，能够广泛应用于新闻阅读、商品推荐等场景；通过远程监督可以补全多模态知识图谱，完善现有的多模态知识图谱，利用动态更新技术使其更加的完备，在端到端实体分类、多模态摘要中也有实际应用。

4.1 推荐系统

多模态知识图谱通过将其他模态的信息引入传统的知识图谱，能够提供的丰富的特征和信息，将其应用于推荐系统可以有效缓解推荐系统的数据稀疏和冷启动问题，从而使推荐结果更准确，并提供可解释性支撑[7]。目标项目及其属性可以映射到知识图谱中以理解项目之间的相互关系。此外，还可以将用户和用户侧信息集成到知识图谱中，以更准确地捕获用户和项目之间的偏好关系。实体的图像和描述可以为知识表示学习提供重要的视觉或文本信息，Sun 等人[8]提出多模态知识图谱注意力网络 MKGAT，作为第一个将多模态知识图引入推荐系统的工作，其包含一个多模态知识图谱嵌入模块和推荐模块，以协作知识图谱作为输入，通过多模态知识图实体编码器和多模态知识图谱注意层为每个实体学习一个新的实体表示，通过聚合实体的邻居的信息，同时保留自身的信息以表示知识推理关系。根据传统的推荐模型生成用户与项目之间的匹配分数，多模态知识图谱在推荐效果上较单模态知识图谱有更好的表现。

4.2 跨模态检索

跨模态检索研究的基本内容是寻找不同模态样本之间的关系，以一种类型数据作为查询来检索其他类型的数据，例如使用文本去检索相关图片或视频。跨媒体检索能够打破检索结果的媒体限制，从而增强搜索体验和结果的全面性。跨模态检索在方法上主要分为两大类：a）实值表示学习;b）二值表示学习，也称为跨模态哈希方法。实值表示学习直接对从不同模态提取到的特征进行学习；而二值表示学习是对从不同模态提取到的特征先映射到汉明二值空间，然后在此空间中进行学习。如文献中提出将深度玻尔兹曼机（deep Boltz-mann machine，DBM）结构扩充到多模态领域，通过多模态DBM,可以学习到多模态的联合概率分布。针对文本或图片输入，利用条件概率生成相应特征，通过检索出最靠近该特征向量的两个实例，可以得到符合条件的结果。Wei 等人以 Im-age Net 为数据源，使用深度语义匹配的方法，将已完成训练的全连接层中的图像特征与文本的语义信息进行配对比较，完成跨模态检索。多模态知识图谱在跨模态检索工作中有较大的应用前景，对于文本检索，输出结果中能够呈现与关键词相关的视觉等其他模态信息，能够有效帮助用户进行实体识别与消歧；对于图像检索，通过一个特征提取模块对检索内容进行特征提取和编码，目标识别和视觉问答工作也可以从中受益。

4.3 人机交互

利用多模态知识图谱融合不同模态数据的特性，可以推动知识驱动的人机交互(human-computer interaction，HCI) 。人类通过人机交互界面与计算机系统进行交流和操作，在实际场景中，通过多传感器的使用，机器能够感知到多模态、数字化的世界，借助多模态知识图谱作为背景知识，有助于机器加强对真实场景的理解，作出更令人舒适、更自然的反馈。例如通过分析人类的语言和面部表情数据对使用者进行情感分析，从而调整环境灯光舒适度等。原来人机交互接入信息更多是从文本、页面中获得，多模态技术会带来新的内容形态，通过听觉和视觉等综合作用，在未来强调沉浸感的人机交互中发挥重要作用。

5 结论

在人工智能从单一模态逐步向多模态演进、从感知智能向认知智能发展的大背景下，多模态数据学习与知识图谱的交互作用为大数据的价值在应用上的落地提供了极富想象力的可能性。本文对目前多模态知识图谱构建与应用相关的研究现状进行了全面的调研和分析，讨论了当前对多模态知识图谱概念的基本认识，其可看做是具有多模态化实体和属性的知识图谱；介绍了图数据库和知识图谱两个视角出发的构建工作，并归纳了基于属性和基于实体的两种构建方法的主要思路；分析了多模态信息抽取、表示学习、实体链接等多模态知识图谱构建和应用中的关键技术和相关工作，并对技术发展面临的标注数据缺失、噪声影响等挑战进行了讨论；列举了多模态知识图谱在推荐系统、跨模态检索、人机交互三个场景中的应用。希望能为多模态知识驱动的人工智能相关领域研究者提供研究思路。

参考文献：

[1]侯海燕，刘则渊，陈悦,等. 当代国际科学学研究热点演进趋势知识图谱[J]. 科研管理，2006，27(3):7.
[2]秦长江，侯汉清. 知识图谱——信息管理与知识管理的新领域[J]. 大学图书馆学报，2009，27(1):9.
[3]胡泽文，孙建军，武夷山. 国内知识图谱应用研究综述[J]. 图书情报工作，2013.
[4]韩燕娟，卜彩丽，张宝辉. 我国微课的研究热点、主题和发展趋势——基于共词分析的知识图谱研究[J]. 现代远距离教育，2015(6):9.
[5]朱永生. 多模态话语分析的理论基础与研究方法[J]. 外语学刊，2007(5):5.
[6]张德禄. 多模态话语理论与媒体技术在外语教学中的应用[J]. 外语教学，2009(4):6.
[7]王东峰. 多模态和大型图像配准技术研究[D]. 中国科学院研究生院(电子学研究所).
[8]王东峰. 多模态和大型图像配准技术研究[J]. 中国科学院研究生院（电子学研究所），2002.