智能系统学报
CAAI TRANSACTIONS ON INTELLIGENT SYSTEMS

领域知识图谱快速构建和应用框架

于皓，张杰，吴明辉，吴信东

引用本文:
于皓, 张杰, 吴明辉, 等. 领域知识图谱快速构建和应用框架[J]. 智能系统学报, 2021, 16(5): 871–884.
YU Hao, ZHANG Jie, WU Minghui, et al. A framework for rapid construction and application of domain knowledge graphs[J]. *CAAI Transactions on Intelligent Systems*, 2021, 16(5): 871–884.

在线阅读 View online: https://dx.doi.org/10.11992/tis.202103024

吴文俊人工智能技术发明奖一等奖

成果名称：知识图谱自动构建及行业应用
获 奖 人：吴信东、张杰、付骁弈、于皓、吴明辉
完成单位：北京明略软件系统有限公司

吴信东

博士，明略科学院院长、明略科技集团首席科学家，IEEE Fellow、AAAS Fellow，HFUT教授、博士生导师。1993年英国爱丁堡大学人工智能博士，研究领域为大数据分析、知识工程，数据挖掘。2001年至2010年为美国佛蒙特大学计算机科学系正教授和系主任。2016年至2018年为路易斯安那大学计算机科学讲席教授、终身教授，计算与信息学院院长。为国际期刊IEEE TKDE前任主编，ACM TKDD现任联合主编。国际期刊KAIS创办人和现任主编，国际会议IEEEICDM创办人和现任指导委员会主席。2012年获IEEE计算机学会技术进步奖。2014年获IEEEICDM10年最有影响力论文奖，2014年CCF(中国计算机学会)优秀博士学位论文导师奖。国家重点研发计划重点专项项目-大数据知识工程(BigKE:Knowledge Engineering with Big Data)首席科学家、项目负责人(2016/7/1–2020/12/31)。“大数据知识工程”教育部重点实验室主任。

团队简介

明略科学院（Mininglamp Academy of Sciences，以下简称MAS）依托于行业人工智能独角兽公司明略科技集团，以科技、商业两翼齐飞的前沿科技阵地为定位，以大数据和人工智能研究为抓手，以“大数据、大知识、大智慧”基础理论为发力点，促进产、学、研、用一体化，助力政府机构和企业实现智能计算和知识服务。目前明略科学院已有来自中国科学院、中国工程院、英国皇家学会、加拿大皇家学会、澳大利亚科学院等机构的十几名院士入选MAS Fellows，共同推动中国行业人工智能的前沿技术发展。他们包括中国科学院院士、中国科学院数学与系统科学研究院研究员陆汝钤，中国工程院院士、清华大学教授吴建平，中国科学院院士、西安交通大学教授徐宗本，中国工程院院士方滨兴等。此外，130余位来自清华、北大、卡内基梅隆大学等国内外著名学府毕业的博士硕士加入成为MAS 技术骨干，许多成员拥有在IBM、NEC、Oracle、Schlumberger 等500强企业的实战经验。

DOI: 10.11992/tis.202103024
网络出版地址: https://kns.cnki.net/kcms/detail/23.1538.TP.20210713.1604.007.html

领域知识图谱快速构建和应用框架

于皓，张杰，吴明辉，吴信东
（明略科技集团，北京 110000）

摘　要：快速构建和应用领域知识图谱 (domain knowledge graphs, DKG) 已成为企业的迫切需求。然而，领域知识图谱的快速构建和应用仍然存在问题：1）构建前，复杂领域场景涉及知识维度广，业务专家短时间内难以构建完备的领域 schema；2）构建中，业务和技术专家深度耦合，图谱构建缓慢，难以适应业务快速发展需求；3）构建后，图谱应用严重依赖技术人员开发，业务专家无法自主基于领域知识图谱探索问题解决方案。为解决上述问题，本文提出了领域知识图谱的快速构建和应用框架，其中包括：多人协作构建领域 schema 解决领域知识的复杂性问题，将业务和技术专家解耦合，提高领域知识图谱的构建效率，最后通过建立基于行业 schema 的 auto-KBQA((knowledge base question answering) 解决领域知识图谱在知识问答应用的快速落地。通过实际项目的应用落地，验证了该框架可有效加快领域知识图谱的落地应用，期望该框架给领域知识图谱的快速构建和应用带来一定的启发和帮助。

关键词：领域知识图谱；schema；实体链指；知识补全；KBQA；多人协作；业务专家；技术专家
中图分类号：TP391　　**文献标志码**：A　　**文章编号**：1673−4785(2021)05−0871−14

中文引用格式：于皓，张杰，吴明辉，等. 领域知识图谱快速构建和应用框架 [J]. 智能系统学报，2021，16(5)：871–884.
英文引用格式：YU Hao，ZHANG Jie，WU Minghui，et al. A framework for rapid construction and application of domain knowledge graphs[J]. CAAI transactions on intelligent systems，2021，16(5)：871–884.

A framework for rapid construction and application of domain knowledge graphs

YU Hao, ZHANG Jie, WU Minghui, WU Xindong
(Mininglamp Technology, Beijing 110000, China)

Abstract: There is an urgent need for enterprises to rapidly construct and apply domain knowledge graphs (DKG). However, there are still problems in meeting this need: firstly, extensive knowledge is required to deal with complex domain scenarios and it is difficult for business experts to rapidly construct a complete domain schema; secondly, the construction process is so slow that it's difficult to meet the needs of rapid business development, because information exchanges are ambiguous between business experts and technical experts; thirdly, after their construction, the heavy dependence of the applications of graphs on technical personnel refrains business experts from exploring approaches to solving business problems. In order to solve the above problems, this paper proposes a framework for the rapid construction and applications of DKG. It builds a domain schema through multiple-person cooperation to solve complex problems and improves the construction efficiency by decoupling business experts and technical experts in the construction and applications of DKG. Finally, it builds the schema-based auto-KBQA((knowledge base question answering) for fast applications of DKG in question and answering mode. Through real-world applications, it is verified that the framework can effectively accelerate the construction and applications of DKG. The framework is expected to bring insights and extend assistance to the rapid construction and application of DKG.

Keywords: domain knowledge graph; schema; entity linking; knowledge completion; KBQA; multiple-person cooperation; business experts; technical experts

近年来，无论是政府或是企业对于数据处理和数据价值挖掘都十分重视，但由于数据总量大且呈现类型多样化等特点，使许多关键数据背后的隐性知识并不能很好地被发现以及利用。知识

收稿日期：2021−03−15.　**网络出版日期**：2021−07−13.
基金项目：国家重点研发计划项目（2016YFB1000901）；国家自然科学基金项目（91746209）；教育部创新团队项目（IRT17R3）；国家发改委项目（2019040416510045 4）.
通信作者：吴信东. E-mail: wuxindong@mininglamp.com.

图谱是连接大数据和人工智能的技术纽带，是从感知智能到认知智能的基石，在复杂度高的行业场景中，领域知识图谱将借助于其天然的知识可解释性和推理等技术方向的优势，在解决实际业务问题、辅助智能决策方向上发挥巨大作用。知识图谱技术在产业界正经历着应用的高速增长。然而，研发领域知识图谱，并在实际应用场景中部署和使用仍面临着：1) 构建前，复杂领域场景涉及到的知识维度广，业务专家短时间内难以构建出完备的领域 schema；2) 构建中，业务专家和技术专家深度耦合，图谱构建缓慢，难以适应业务快速发展需求；3) 构建后，图谱应用严重依赖技术人员开发，业务专家无法自主进行领域知识图谱在业务问题解决方案中的探索。

本文立足于将知识图谱相关前沿技术成果应用落地，设计开发了面向领域的知识图谱快速构建和应用框架，主旨是利用知识图谱相关技术，从异构多源数据中提取知识，快速构建出领域知识图谱，并持续将碎片化知识融合到领域知识图谱，形成体系化的领域知识。通过知识图谱向量化方法对领域知识进行丰富和深层次的领域语义表示，突破传统的基于字符串匹配的浅层语义，更加高效地辅助用户发现潜在领域知识价值，在应用于搜索、推荐、推理等传统任务之外，其将在领域流程优化、辅助决策、预测分析等应用服务发挥更大空间。该框架已成功应用在公共安全、金融、工业、广告营销等领域，发挥出巨大的商业价值和社会价值，在该框架中集成了 KBQA 的快速构建落地方法，以对话的形式降低人工智能产品使用门槛，高效提供业务决策支持，有效降低知识劳动力成本，提高知识转化为企业竞争力的效率。

1　相关工作

1.1　HAO 智能理论

HAO(human intelligence, artificial intelligence, organizational intelligence) 智能理论[1] 是大数据到大智慧的理论框架，如图 1 所示，通过对人类智能 (human intelligence, HI)、机器智能 (artificial intelligence, AI) 和组织智能 (organizational intelligence, OI) 三位一体的集成，构建了新的人工智能理论。在未来万物互联的时代，只有打通感知智能与认知智能，将深度学习与行业知识图谱相结合，才能扩大行业人工智能的应用领域，加速人工智能技术商业化落地。行业人工智能的应用领域，加速人工智能技术商业化落地。

HAO 智能通过打通感知、认知、行动系统，帮助组织进行分析决策，实现 AI 闭环落地，其中机器智能需要人类智能进行大量的语料标注，以使机器智能不断地学习，提升机器智能的水平，机器智能为人类提供的知识而非数据，标注的语料作为人类智能的载体传递给机器，机器通过对标注数据的学习产生机器智能，然后分析挖掘新的知识输出给人类进行学习，以优化人类智能，从而产生人类智能和机器智能的协同优化升级，人和机器通过该框架打造成一个统一的有机组织，通过将专家知识和使用者的知识数字化，使最终的人类智能和机器智能在组织中不断迭代优化，将目前人工智能模型无法解决的复杂问题，通过人类的智慧在组织智能中协同解决，形成最终 AI 闭环，从而产生最大化的应用价值。

1.2　知识抽取

1) 实体识别

命名实体识别 (named entity recognition, NER) 旨在从预定义好语义类型 (例如人、位置、组织等) 的文本中识别出相应实体类型的提及词[2]。1996 年在第 6 次信息理解会议[3] 上 NER 作为从文本中提取人员、地名、货币、时间和百分比等信息的任务而被首次使用。自此，人们对 NER 的兴趣不断增加，投入了大量的精力进行研究。早期的 NER 采用基于规则和字典的方法，随着机器学的发展，人们开始尝试将一些机器学习的方法用于 NER 中，例如：隐马尔可夫模型 (hidden markov model, HMM)[4]、决策树[5]、最大熵模型 (maximum entropy models, ME)[6]、支持向量机[7]、条件随机场 (conditional random field, CRF)[8]，其中，CRF 是最有效的 NER 算法之一。后来，随着深度学习的快速发展，很多工作都提出利用神经网络完成 NER 任务，Lample[9] 提出了提出了长短时记忆神经网络（long short-term memory networks, LSTM）和 CRF 组合的神经网络模型 LSTM-CRF，利用双向 LSTM 对输入文本进行表征学习，然后将其输入到 CRF 中，对句子中的每个词进行分类，最终输出分类结果，完成实体识别。Zhang[10] 对 LSTM 做了进一步的改进，提出了 Lattice-LSTM 以融合词汇信息。Devlin 等[11] 提出了基于 Transformer 的预训练模型 BERT(bidirectional encoder representations from transformers)，刷新了多项 NLP（natural language processing）任务的记录，并在 CoNLL-2003 的 NER 数据集仅仅通过 fine-tuning 就得到了接近 state-of-the-art 的成绩。在此之后，很多

研究都是基于 Transformer 来展开的。Zhang[12] 将知识库信息与 Bert 语言模型进行融合, 使得模型掌握更多的先验知识, 提高模型表达效果。Li 等[13] 提出了 FLAT(flat-lattice transformer for Chinese NER)模型, 基于 Transformer 设计了一种巧妙位置编码结构来融合 Lattice 结构, 从而引入词汇信息, 在 MSRA(microsoft research)-NER 任务中, FLAT+BERT 实现了最新的 SOTA(state of the art)。

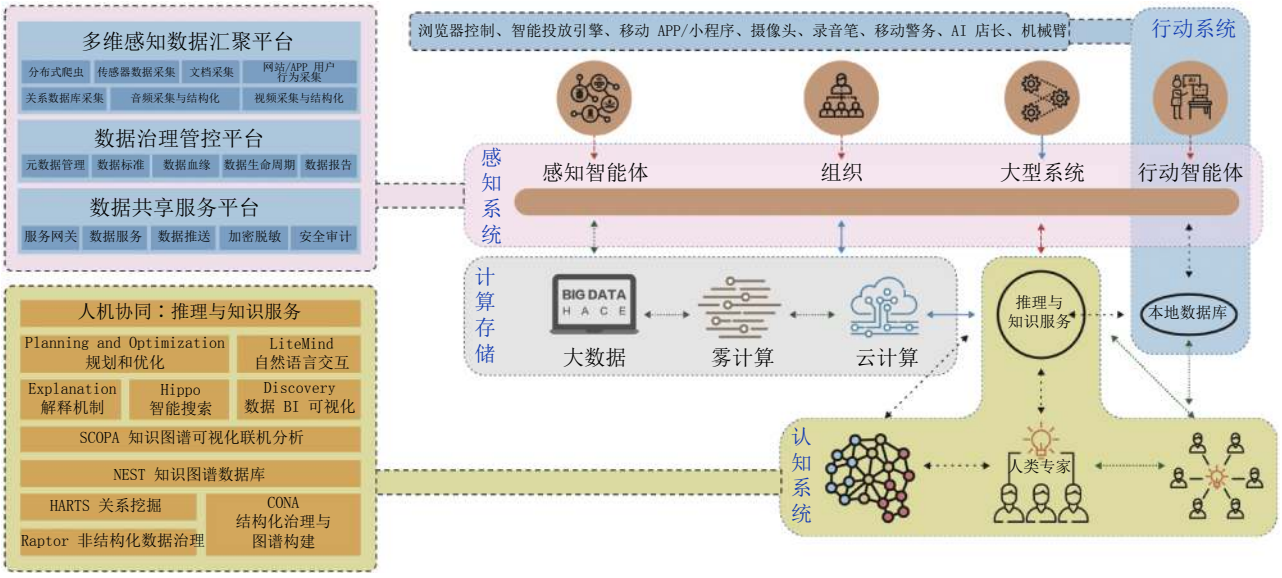

图 1　HAO 智能

Fig. 1　HAO intelligence

2) 关系抽取

在缺少标注数据的场景下, 半监督的方法能够取得一定的效果。Ye 等[14] 提出一种 Intra-Bag&Inter-Bag Attention 的远程监督方法, 在 bag 内的句子以及每个 bag 都添加 Attention 机制, 来减少错误标注数据对关系分类模型的影响。Qin 等[15] 提出度强化学习方法识别远程监督方法中错误的样本, 他们认为 Attention 机制并不是最优的选择, 标注样本的错误数据始终是模型的瓶颈。Alt 等[16] 提出将预训练模型应用到远程监督中, 预训练模型能够更好地捕捉句子的语义和语法信息, 以解决在关系分类中的长尾现象, 但是预训练模型对硬件要求高, 在工业场景下落地较困难。Ye 等[17] 提出了 MLMAN(multi-level matching and aggregation network) 结构的小样本学习方法, 该方法采用原型网络[18] 思想, 分别计算查询实例的嵌入向量和各支持集的原型向量。但是在实际应用场景中, 每个关系类别的标注实例很可能是极度不均匀的, 少样本典型的 N-way K-shot 场景可能并不完全适用。

1.3　实体链指

实体链接任务旨在研究如何将文本中对实体有歧义的“提及”链接到目标知识库所对应的实体上。在研究方法上, 实体链接任务经历了从传统的基于特征工程的方法到目前基于神经网络的端到端方法的过渡。Shen 等[19] 梳理了深度学习时代当中基于传统机器学习算法的实体链指技术, 包括候选实体生成、候选实体排序等。伴随着神经网络的发展, 实体链指技术引入了基于大规模预训练语言模型的表征算法以及注意力网络来捕捉提及、实体以及二者的相似度, 继而大幅提高链指精度[20]。而近年来, 链指当中的不可链接预测 NIL、标准数据不足的问题成为当下的研究热点, Gu 等[21] 提出利用多轮阅读理解 MRC(machine reading comprehension) 框架, 并设计不可链接预测验证和门控机制, 通过轮询已识别提及对应的实体描述信息优化对后续提及的判别, 设计新颖; Wu 等[22] 提出两步的 zero-shot 模型, 提出 cross-encoder 将提及上下文和候选描述融合, 并实现知识蒸馏, 验证了低资源链指的可行性。

1.4　知识补全

受限于业务语料规模, 领域知识图谱存在不完备性问题。知识补全方法通过预测三元组缺失部分, 可对领域知识图谱进行有效补充。基于平移距离的方法 TransE[23], TransH[24] 和 TransR[25] 等, 仅依赖于网络结构, 对于只有少量关系的实体表现不佳。Shi[26] 提出了 ConMask 算法, 将实体描述信息嵌入到模型, 丰富了语义表示。知识表示

无法解决复杂路径问题，为此 Gardner[27] 基于路径排序算法 PRA(path ranking algorithm)[28]，提出了将向量相似性计算与随机游走结合。然而随机游走的搜索空间过大，DeepPath[29] 首次将强化学习应用于链接预测中。知识补全的另一大痛点是长尾问题，许多关系出现的频次很低，却更需要补全。GMatching[30] 结合实体嵌入式表示和局部网络结构信息，提出了基于度量的小样本学习方法。

1.5　KBQA

传统的 KBQA(knowledge base question answering) 模型可以大致分为两种类别，即基于语义分析[31-33] 和基于信息检索[34-36] 的方法。之前的模型着重于将问题映射到其形式逻辑表达形式以便于在知识图谱上进行查询。相反，近些年的模型主要研究问题文本与知识图谱中三元组的语义相似性。例如：Dong 等[35] 使用多列卷积神经网络嵌入文本而无需使用任何人工特征和词典，利用 CNN(convolutional neural networks) 模型捕捉问句与答案属性间的字面关联性，验证了字面关联性能带来效果上的提升。Zhang 等[36] 提出了利用注意力 (Attention) 机制来解决字符级别的语义匹配问题，通过对候选答案的不同维度表示增强了对问题的动态表示能力。随着深度学习能力的提高，基于 SimpleQuestions 数据集的回答性能已接近上限，此后的研究向基于多条件、多跳推理的复杂问题转移[37-38]，多元关系推理和知识库的结构表示被逐渐重视起来。

2　领域知识图谱构建框架

为解决复杂业务问题构建的领域知识图谱，需要建立在业务知识的框架下，否则很难将领域知识图谱应用到实际的问题解决中，自上而下的 schema 设计和自下而上的知识抽取相融合的模式是构建领域知识图谱重要方法。复杂领域场景涉及到的知识维度广，需要业务各方向的专家参与制定领域 schema，一方面，行业专家对构建知识图谱的技术了解较少，需要时间进行学习，另一方面，需要行业专家协同构建体系化的领域 schema。目前缺少有效工具帮助其协同开发，因此导致领域专家短时间内难以构建出完备的领域 schema。为了有效提高领域专家系统构建领域 schema 的效率，本文在领域知识图谱构建框架中，设计了多人协作构建领域 schema 方法。

2.1　多人协作构建领域 schema

多人协作构建领域 schema 存在诸多问题：schema 术语多样性难以统一、领域专家构建的 schema 存在业务边界难以融合体系化、构建的领域 schema 存在缺失、领域 schema 和底层数据无关联性难以维护。基于以上问题，在该框架中，设计了权限管理、协作模式、schema 融合、schema 推理、融合日志和抽取器功能模块。

1) 权限管理

根据不同的领域知识图谱角色，在协同开发领域 schema 的过程中，分配角色不同的协同权限，其目的是保证全域的 schema 具有权威性，从而保证底层的数据可以较准确地映射到业务层面的知识体系中。

2) 协作模式

在协作构建过程中，设计的概念实体需要满足规范性、完全性、一致性、可扩展性和语义区分性，领域专家可以独立构建其领域 schema 子图，再通过 schema 融合，形成全域的行业知识图谱，也可在一个领域 schema 图中构建全域知识 schema。

3) schema 融合

为了提高协同构建概念图谱的效率，框架中预设通用域和特定域的知识图谱，用户可以检索相应的本体概念，将相应的子分支下的概念体系融合到正设计的概念维度，也可在已有的概念图谱中申请协作开发，形成最新的图谱 schema。

4) schema 推理

在复杂的领域知识图谱的 schema 设计中，容易遗漏概念间的隐性关系，在该框架中，设计了基于规则的推理技术，通过设定领域概念规则，可以通过推理的方式发现新的关系，对领域 schema 进行有效的知识补全。

5) 日志

协同开发过程中面临着对概念认知不一致的情况，从而导致在构建领域 schema 的过程中，存在概念实体的分歧，通过保存所有参与用户的操作日志，可以帮助协同人员对分歧的概念实体进行讨论确定，以达到共识，避免在下层的数据层面出现分歧点。

6) 抽取器

在概念图谱构建之后，就建立了从上层的业务到业务知识的映射，为了打通从底层的大数据到业务知识的映射，该框架设计了基于图谱 schema 构建抽取器进行映射，通过在图谱 schema 中构建相应的抽取器，实现从底层数据到业务知识的无缝隙映射。

2.2　基于抽取器的图谱自动构建

在复杂业务场景中，构建领域知识图谱周期较长，通常以半年周期倍数计算，具体建设周期

和领域业务复杂度、底层数据质量和投入的资源等方面相关, 而在一些特定领域, 业务发展速度快, 对领域图谱的构建需要以周为周期, 否则难以适应业务发展速度, 构建的领域知识图谱具有较大的延迟性, 无法满足对业务的支撑, 为了解决企业对快速构建领域知识图谱的需求, 本文设计了基于抽取器的快速构建知识图谱的框架, 业务专家依据业务需求选择相应的抽取器灵活构建知识图谱, 该框架有效降低了知识图谱构建的技术门槛, 大幅提升领域知识图谱的构建效率。

在本文设计的框架中, 将领域知识图谱的构建进行模块化, 主要分为数据层、语料层、算法层、组件层、抽取器层, 如图 2 所示, 对每层的功能点进行封装, 从而形成了从数据层到业务层的组装式映射, 为业务专家提供灵活的领域知识选择, 快速生成满足业务需求的精简的领域知识图谱, 避免大而全的领域知识图谱所产生的噪音、效率低等问题。

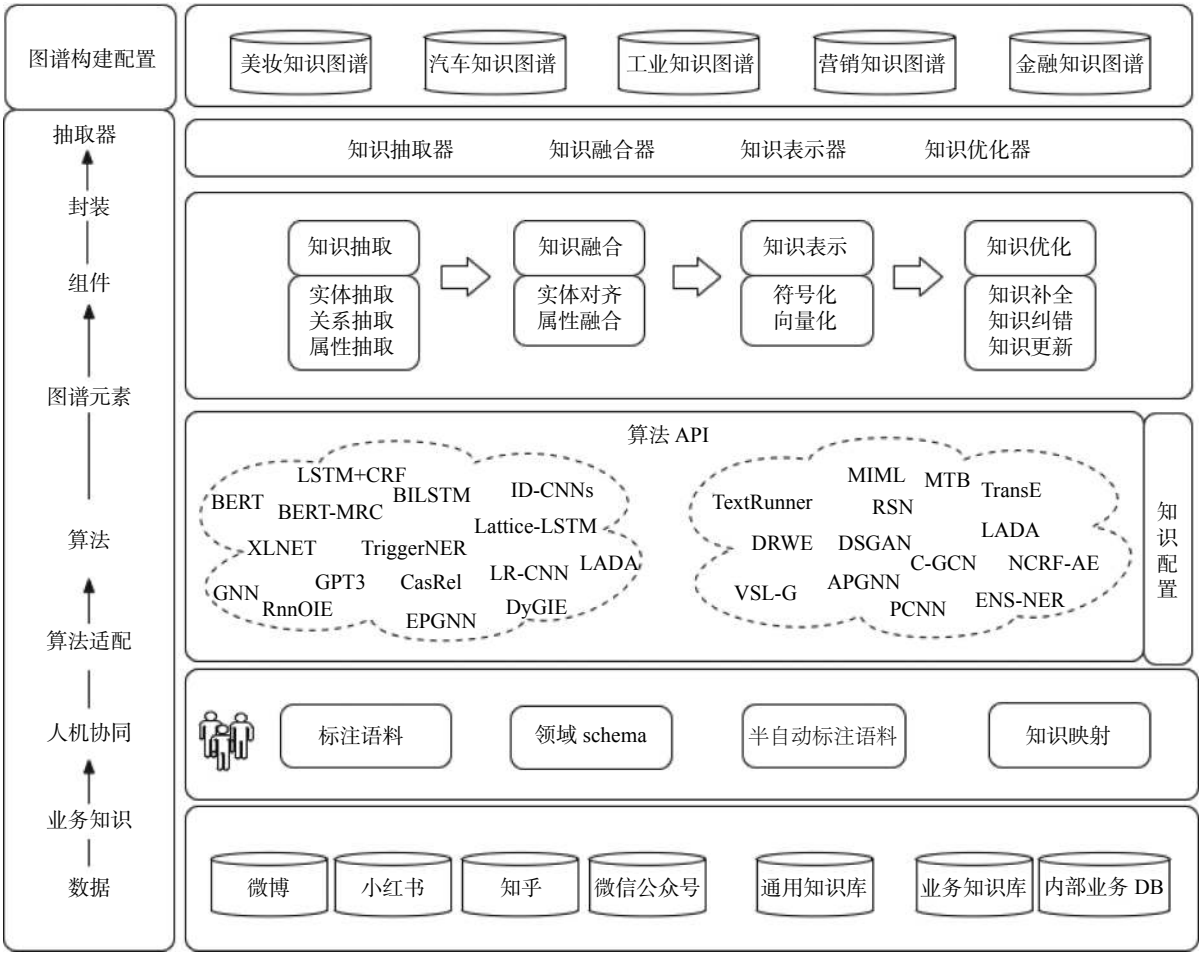

图 2　领域知识图谱快速构建框架

Fig. 2　Construction framework for domain knowledge graphs

在企业内主要存在两种数据类型: 1) 业务相关的结构化数据; 2) 从互联网获取的公开数据。业务数据在知识层面更深, 开放数据在知识层面更广, 两者可以互相补充, 形成较完备的领域知识。通过将领域业务数据和公开数据融合构建领域知识图谱, 是目前行业通用的模式, 企业积累的结构化数据是基于业务逻辑关联的, 可以和领域 schema 有效融合, 通过简单的映射, 就可以将业务结构化数据映射到领域 schema 知识体系中, 这部分的知识抽取相对简单, 而对于公开的非结构数据, 如何映射到领域知识图谱中是难点: 1) 可以在领域 schema 的规范下进行人工标注, 生成领域的非结构化语料, 优点是可以形成高质量的训练数据, 为后期算法训练提供良好的数据条件, 缺点是需要耗费大量的人力; 2) 借助通用知识库或者远程监督等技术手段, 半自动生成标注语料, 优点是可以快速生成深度学习需要的语料, 节省人力成本, 缺点是标注的语料存在一定程度的错误, 会影响后期算法训练的模型效果。

从数据层到人机协同层, 将数据的信息借助

于人力智能，形成了机器可学习的语料。通过选择相应的算法进行训练，可以将数据空间映射到业务知识空间。在算法层中，框架集成了近几年主流的监督、半监督和无监督的知识图谱构建相关的技术体系，封装算法接口，为构建知识图谱的组件提供算法调用。

知识图谱的构建流程主要分为知识抽取、知识融合、知识表示和知识优化，知识抽取主要是借助于算法层的实体抽取算法、关系抽取算法、属性抽取算法或者联合抽取算法，对标注好的语料数据进行模型训练，生成相应知识抽取组件。知识融合主要解决在知识抽取过程后的知识对齐和属性融合问题，形成一致性较好的领域知识图谱，在知识构建之后建立符号化和向量化的表述组件，满足不同的业务知识表示需求。知识优化则为了在构建的领域知识图谱进行知识质量的优化提升，挖掘领域知识图谱中隐漏的领域知识，发现知识冲突并对领域知识进行更新，从而形成了一整套的知识图谱构建组件。

在组件层通过算法构建了从底层数据中学到业务语义知识识别模型，将模型进行服务化的封装生成抽取器。将模型的输入、输出以及对输入数据的预处理等功能模块封装为可独立运行的抽取器。抽取器配置相应的业务功能说明，业务专家在这个层面可以根据业务问题，选择相应的抽取器自主构建领域知识图谱。这样做的优点是借助于业务专家对业务的了解，生成实际可解决业务问题的知识图谱，防止技术人员缺乏业务知识，生成的领域知识图谱难以和业务有效结合的缺点，技术人员和业务人员在构建领域知识图谱的过程中分工明确，各司其职，减少两者协调工作的复杂度。

本文框架设计主旨是为业务专家提供高效的领域知识图谱构建方法，将企业中业务专家和技术专家既联合又分割，在对数据分析时技术专家需要业务专家协助，在构建领域知识图谱时，业务专家需要技术专家指导。在领域知识图谱构建的整个流程，两种角色相对独立，技术专家负责从数据到抽取器的构建，而业务专家负责从业务问题选择相应的抽取器构建领域知识图谱。详细的领域知识图谱构建流程如图 3 所示。

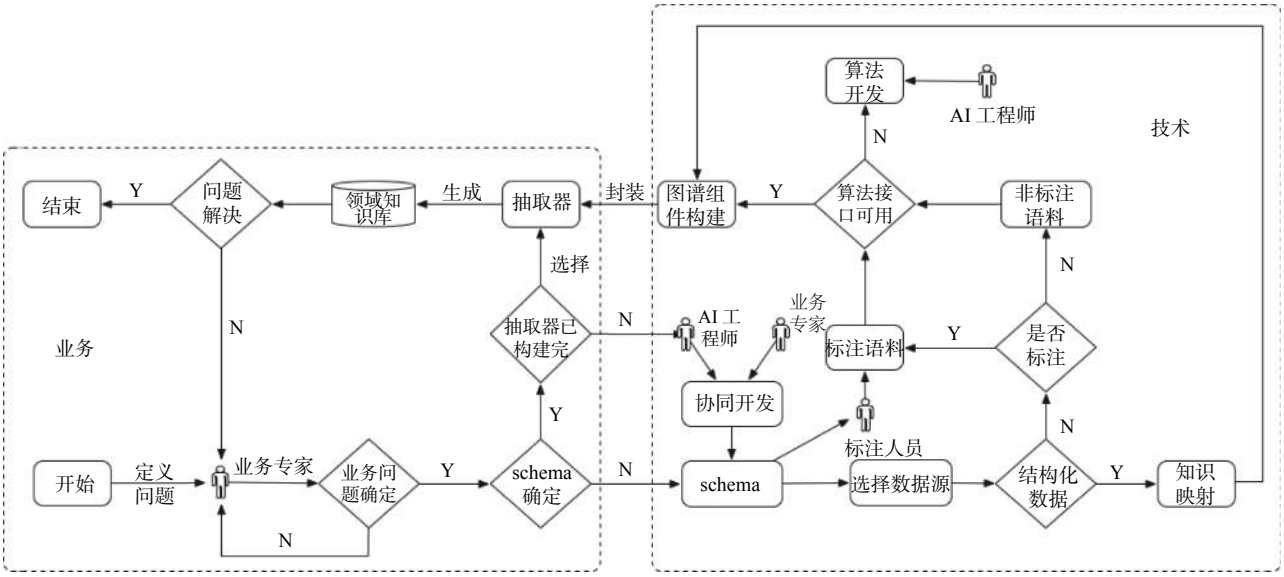

图 3　领域知识图谱构建流程

Fig. 3　Construction processes for domain knowledge graphs

2.3　SA-KBQA

目前，领域知识图谱构建之后，主要应用于搜素、推荐、问答和以可视化方式进行人机交互，为解决业务问题提供可解释和辅助决策的支撑，这个过程需要 AI 工程师的深度参与，难以以统一的形式赋能下游的实际业务问题。业务专家无法独立完成领域知识图谱到业务应用的转化，本文提出了 SA-KBQA 帮助业务专家自主构建领域知识图谱之后，可配置将领域知识图谱应用到对话，用来解决业务问题，详细的构建流程如图 4 所示。首先业务专家根据要解决的业务问题，依据设计的领域 schema 进行业务问题配置，框架根据配置文件，自动生成问题模式集合，然后将问题集合自动生成标准化的查询语句，从而构建了从业务问题到领域知识图谱查询的完整流程，在这个过程中完全由业务专家进行操作和制定。在

问答系统应用阶段, 通过业务专家配置的命名实体识别和实体链指等抽取器将问句中的关键实体映射到领域知识图谱中的标准实体, 然后通过图谱 schema 结构对当前关键实体的所有路径生成候选的查询语句, 对部分不符合查询规范或者无查询结果的语句进行剪枝, 最后将问句、查询语句和结果进行排序, 获得最佳结果输出。通过 schema 结构对实体查询语句的召回方式可以实现多跳问题的查询, 提升了解决领域问题的难度。

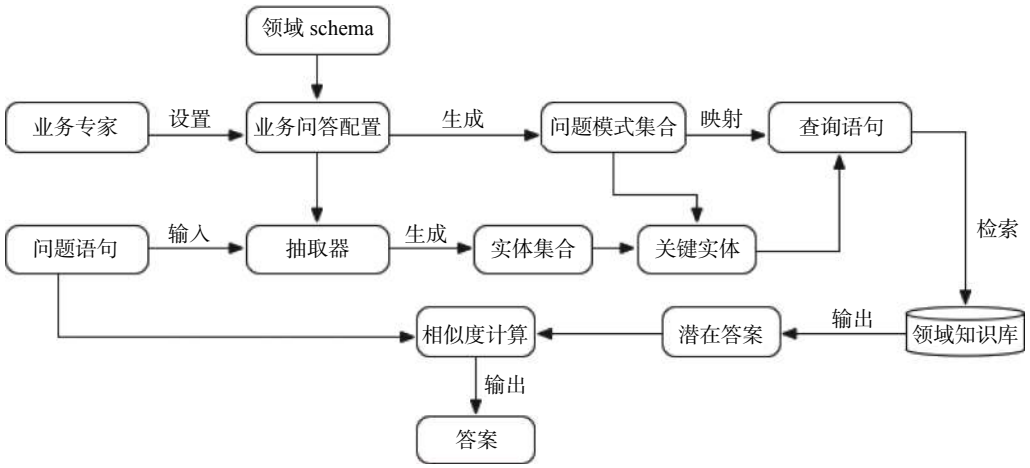

图 4　SA-KBQA 构建流程

Fig. 4　Construction processes for SA-KBQA

3　应用案例

3.1　线上美妆营销

美妆行业品牌众多、品类丰富、产品繁杂, 对于消费者面临琳琅满目的产品, 无法选择合适产品, 而对于化妆品销售员, 无法对所有产品清晰熟知, 面对消费者具有多条件和深层跨域知识维度的提问时, 销售顾问无法给出令消费者满意的答案, 而知识图谱可以将美妆领域的所有品牌、品类、产品、成分、功效等实体有机关联, 形成系统化、全域性的美妆知识体系, 再借助于知识图谱在推理方面的优势, 可以有效解决上述的业务痛点问题。除此之外, 基于网络社交场景美妆营销数据, 分析用户对美妆行业的品牌、产品的评价, 从而帮助企业对产品进行改进, 提高用户的满意度, 另外通过大数据分析用户的需求, 给产品研发提供客观的数据支撑, 帮助企业更全面了解消费者的需求。

以往构建领域知识图谱需要经过业务专家和技术专家数周甚至数月时间构建领域 schema, 借助于本文提出的多人协作构建领域 schema 的方法, 在实际项目中仅用 7 天时间完成全部美妆领域 schema 的构建。

1) 多人协作快速构建美妆 schema

构建美妆知识图谱的业务目标, 是将大数据底层的信息抽取出美妆产品知识, 帮助企业进行产品运营和创新, 业务专家从领域问题出发多人协作快速构建美妆概念: 品牌、品类、产品、成分、包装、场景等 29 类, 如图 5 所示。AI 工程师结合专家上层设计对网络社交数据例如微博、小红书等数据分析, 确定底层数据对美妆实体和关系抽取的可行性, 最终确定美妆领域知识图谱的 schema, 美妆业务专家和技术专家分别从业务顶层和数据底层两个角度协同开发, 即确保从底层数据到顶层的业务知识可以无缝隙打通, 又加快了美妆 schema 的构建效率。

2) 美妆知识图谱快速构建流程

首先美妆行业知识官和 AI 工程师两种角色多人协同快速完成美妆领域知识图谱的 schema 设计, 然后 AI 工程师从两个方面构建底层的数据映射到美妆的业务知识层面的抽取器。对于结构化数据进行知识映射, 而对于非结构化数据, 通过调用框架集成知识抽取算法 pipeline 进行知识抽取, 构建知识图谱组件, 封装成美妆知识抽取器。基于社交数据提取的美妆知识缺乏大量知识, 例如产品的功效, 社交对产品的功效的交互主要集中在产品功效的优缺点, 对于没有鲜明对比优势的功效很少提及, 因此需要通过知识补全技术。对产品的功效等知识进行补全, 构建了美妆知识补全抽取器, 从而形成了从美妆数据到美妆领域 schema 的语义空间映射, AI 工程师负责

将底层数据构建成美妆抽取器，美妆业务专家根据业务需要，可灵活、快捷选择相应的美妆抽取器，快速建立美妆领域知识图谱，其详细流程如图 6 所示。

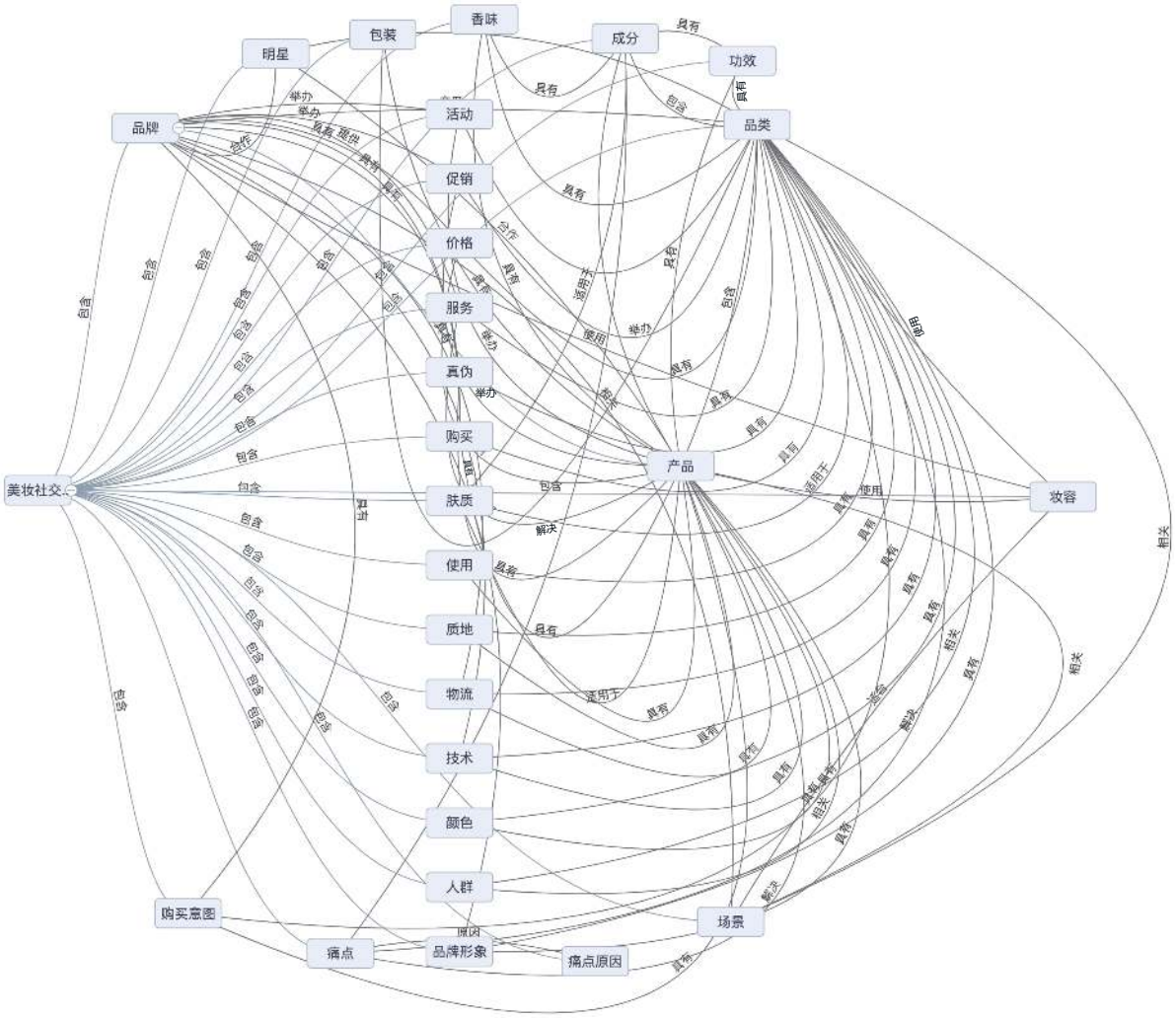

图 5　美妆 schema 设计

Fig. 5　Makeup schema design

3) 知识管理

在自动化构建的美妆知识图谱中，无法保证抽取的美妆知识的完全准确性，在框架中集成人机交互的模式对美妆知识进行修正和补充，美妆知识官通过知识管理模块对知识图谱进行管理，如图 7 所示，即可以帮助其从美妆整体维度审查领域知识的完备性，又可以辅助其进行局部的知识管理以及在解决业务问题方面的辅助洞察功能，例如帮助行业知识官了解其产品的受众群体的用户画像等。

4) 美妆知识图谱快速应用

为了提高美妆行业知识图谱的快速构建和落地，降低行业知识图谱构建的技术门槛，框架集成了基于对话问答模式的行业知识图谱模式，通过输入“打开知识流程”对话，如图 8 所示，问答机器人会提供完整的行业知识图谱构建流程，行业专家可以按照标准化的流程创建领域知识图谱，具体的后台构建流程如图 9 所示。

美妆知识图谱构建完成后，基于 SA-KBQA 框架，可以迅速构建领域的知识问答，针对领域知识图谱已有知识，对用户提出的问题进行回答，同时该框架中集成了推理功能，对用户的问题，给出相应的答案推理路径如图 10 所示。

若回答的问题错误，同时提供人机交互的方式，以完善行业知识图谱的知识。通过该框架，实际美妆知识图谱项目落地应用周期缩短 40%，客户对销售顾问的满意度提升 23%，同时在营销洞察方面有效提升企业对消费者的痛点感知，在优化美妆产品和提升企业产品质量方面发挥了较大作用。

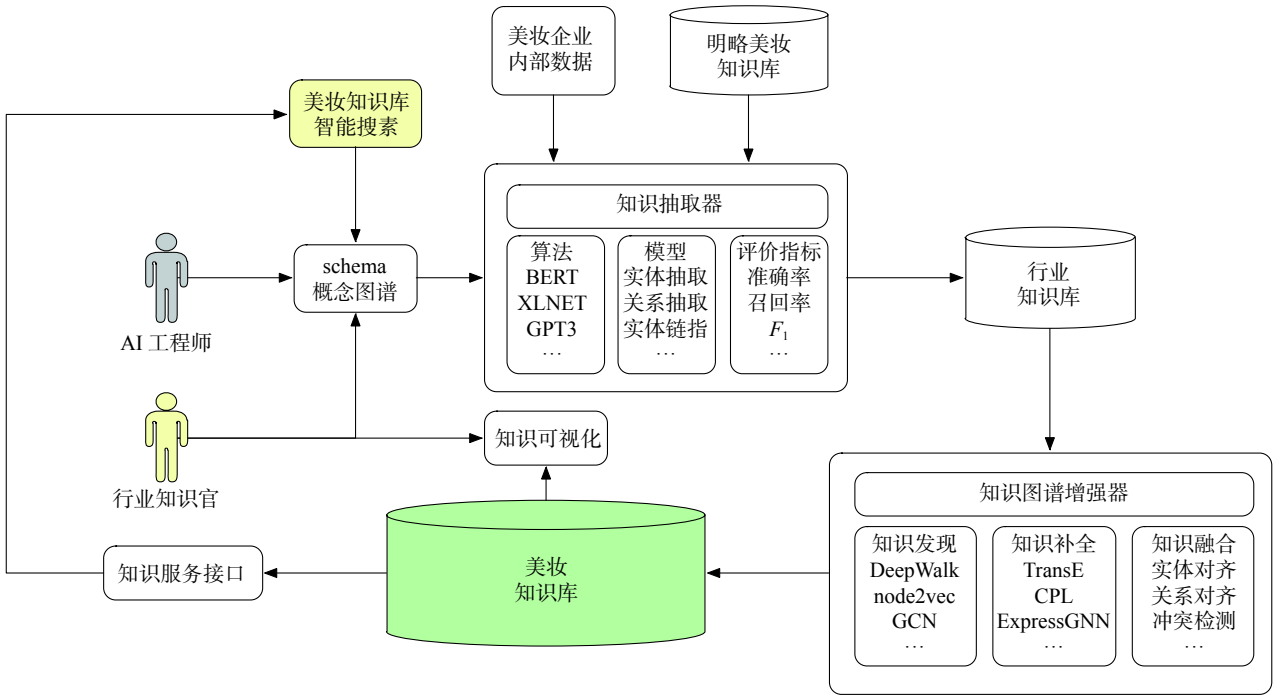

图 6　美妆知识图谱构建流程

Fig. 6　Construction processes for makeup knowledge graphs

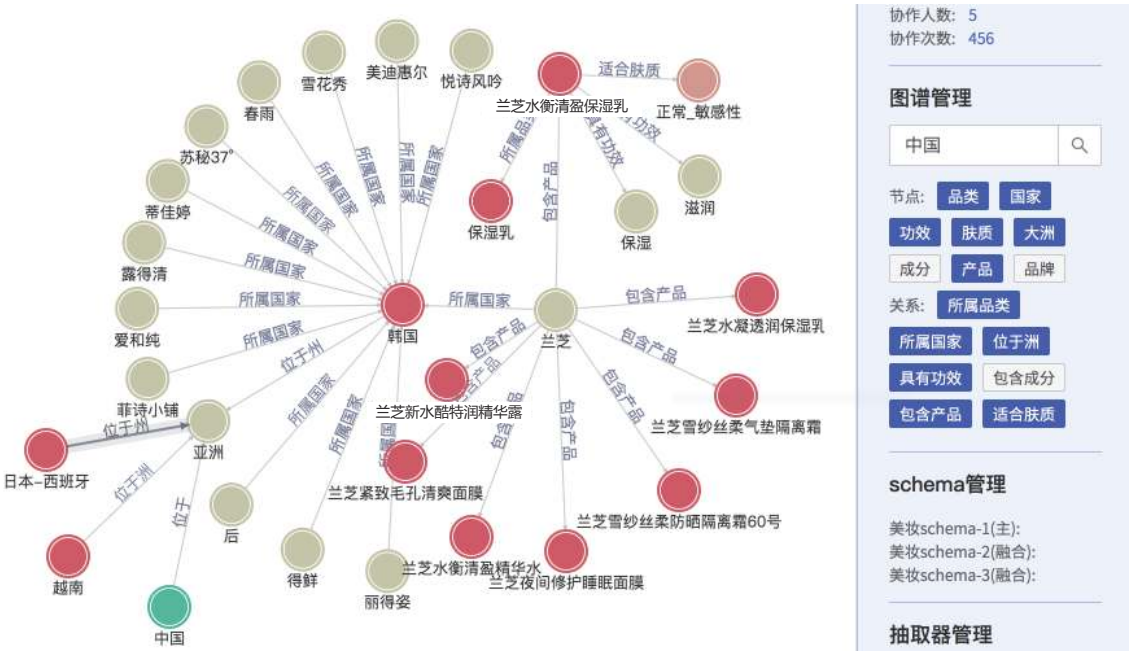

图 7　美妆知识管理

Fig. 7　Makeup knowledge management

3.2　汽车销售

汽车领域属于知识密集型产业, 汽车产品众多、功能繁杂, 消费者难以抉择。而销售顾问一方面无法掌握全域的汽车知识, 另一方面对消费者的推荐产品往往存在一定的主观意愿, 无法从消费者的实际需求提供最佳的产品推介, 从而降低了对消费者的服务质量, 阻碍了汽车企业的快速发展。因此通过科技手段赋能汽车销售顾问, 帮助其对自己的销售过程进行回盘, 自动化分析其在销售过程中的细节, 借鉴优秀汽车销售人员的整套服务过程细节, 可以有效帮助其提升业务销售能力, 同时对消费者的深度洞察, 帮助汽车企业进产品优化, 满足消费者对汽车的更高要求。

图 8　美妆知识图谱构建

Fig. 8　Construction makeup knowledge graph

借助于本文提出的多人协作构建领域 schema 的方法，汽车销售专家从领域问题出发，自上而下设计了汽车领域 schema，AI 工程师结合车企内部数据和互联网公开数据，将底层数据抽取的实体和关系映射到业务专家设计的汽车 schema 中，并将知识抽取和知识补全两个维度封装为多个抽取器，企业销售业务专家根据实际业务需求，快速构建了汽车领域知识图谱，其目的是拓展汽车消费顾问的知识边界，同时将销售过程中的话题基于时间序列构建事理图谱，帮助销售顾问了解销售过程的细节，以发现其销售过程的缺点，进行销售过程优化，销售过程框架如图 11 所示。

汽车销售业务专家可自主、便捷选择基于车企内部数据和网络社交数据封装的各类抽取器快速构建汽车领域知识图谱，如图 12 所示。该图谱的目的是帮助汽车销售顾问全面、细致了解汽车产品间在基本参数的差异以及消费者在社交平台上对汽车的评价信息，从而帮助汽车销售顾问熟悉产品在消费者中的真实感受，以帮助其在销售过程对消费者进行个性化服务，不断提升业务销售能力，提高消费者对服务的满意度。

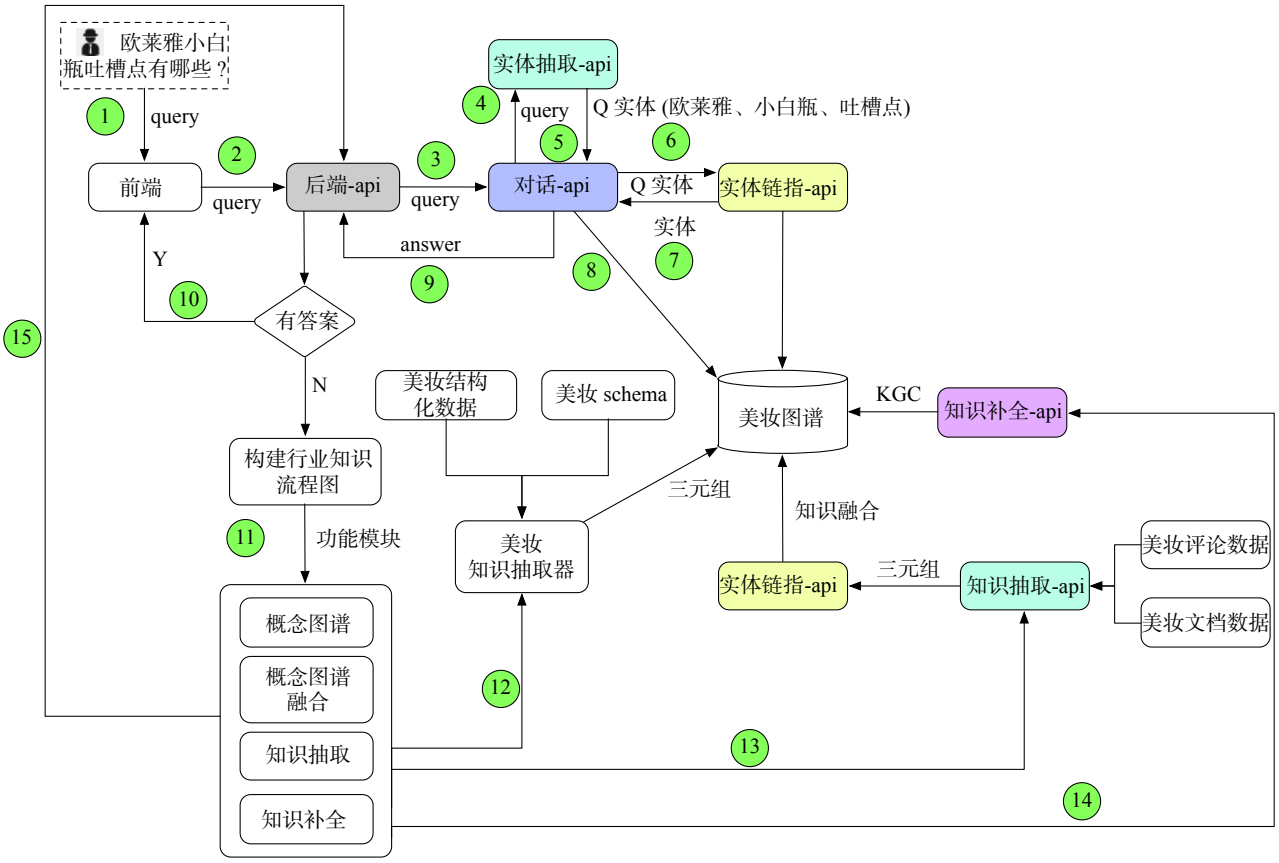

图 9　美妆知识图谱构建流程

Fig. 9　Construction processes for makeup knowledge graph

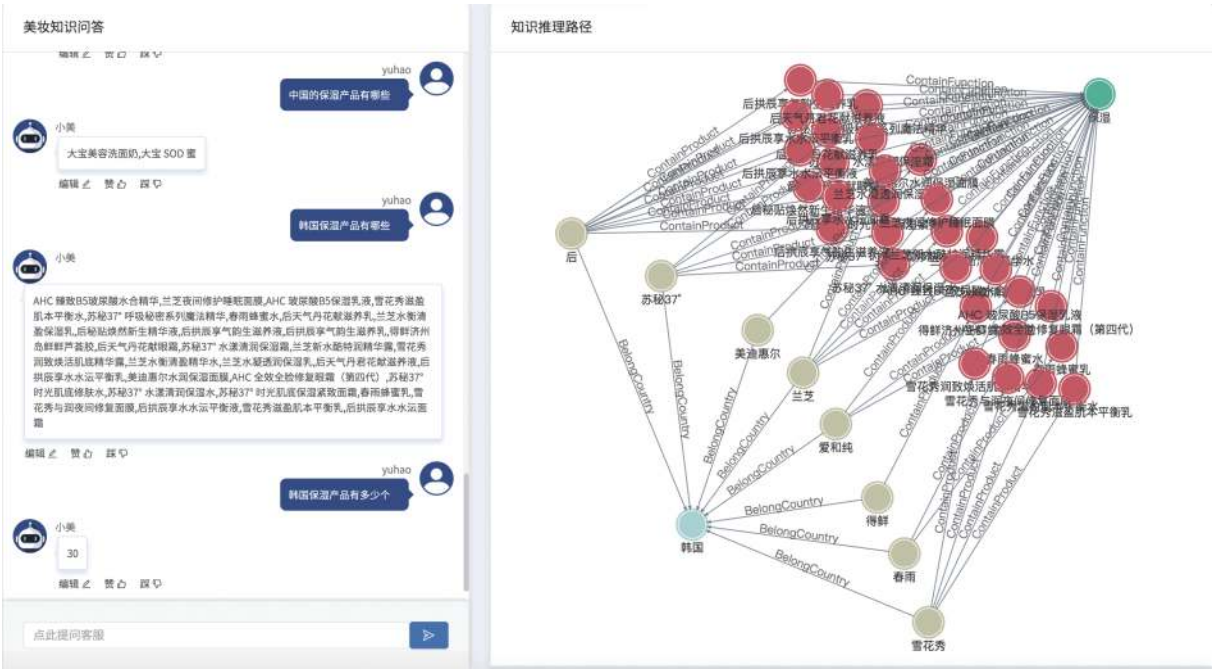

图 10　美妆知识图谱问答案例

Fig. 10　Case of makeup KBQA

图 11　汽车销售过程数字化

Fig. 11　Digital management platform for automobile sales

汽车销售顾问对顾客的销售, 在整个汽车销售过程中起到了非常重要的作用, 通过事例图谱将汽车销售过程进行数字化, 可以帮助销售顾问对自己的销售问题追因, 帮助其改进销售过程, 在该事例图谱中, 将销售过程进行话题标签识别, 在此基础上基于时间序列, 生成销售话题标签的转移矩阵, 形成完整的销售过程话题事理图谱如图 13 所示, 为后期企业对销售过程的优化和新人培训等业务场景问题提供科学的知识辅助。

该项目中的领域知识图谱构建和销售过程的话题实例图谱的构建相比计划提前 30%, 整个项目部署应用之后, 销售顾问的销售技能得到有效提升, 其表现在消费者的满意度和销售业绩以良好的态势在逐步提升。

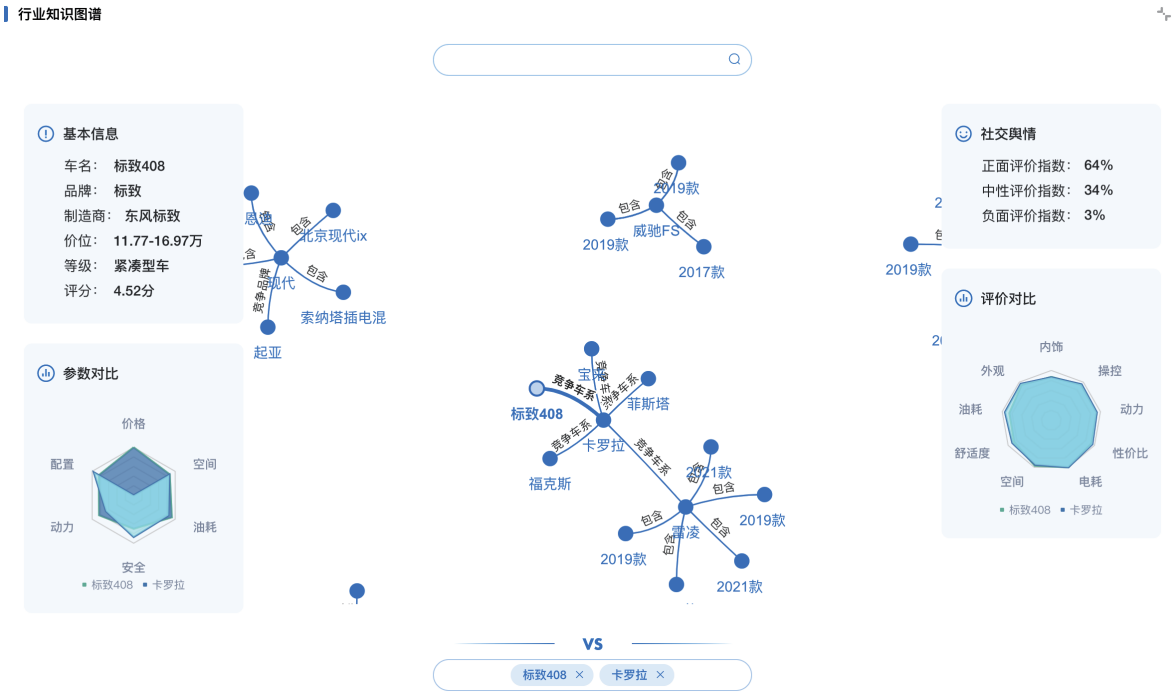

图 12　汽车领域知识图谱

Fig. 12　Automobile knowledge graphs

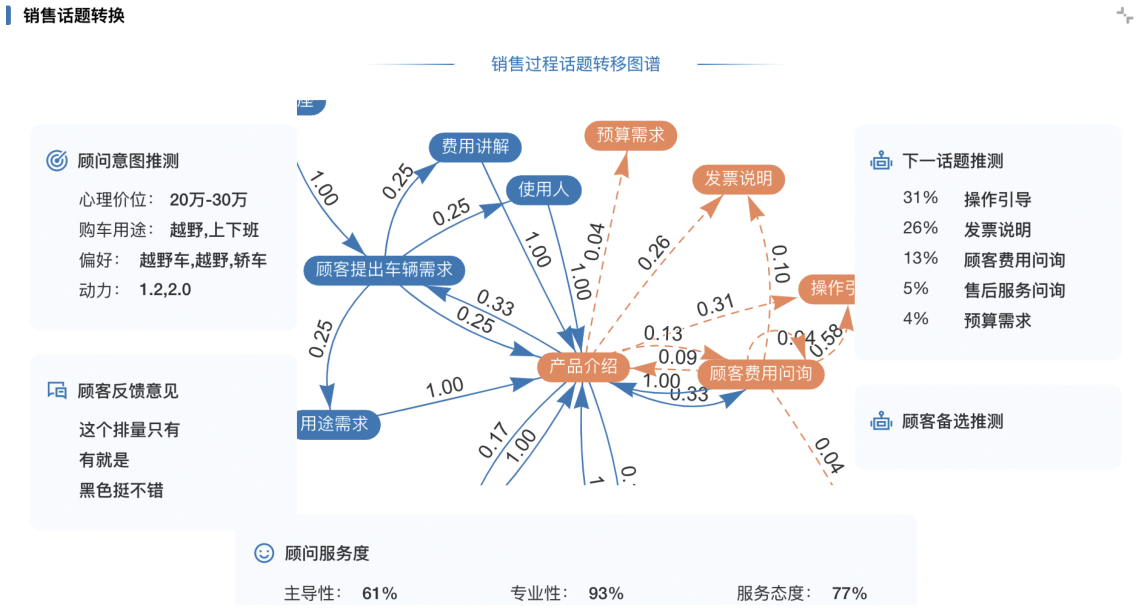

图 13　汽车销售事例图谱

Fig. 13　Event evolutionary graph for automobile sales

4　结束语

知识图谱已成为探索从感知智能到认知智能的重要途径，为解决企业对领域知识图谱的快速构建和应用需求，本文研发了领域知识图谱快速构建和应用框架，设计了多人协作模式构建领域 schema 方法，解决行业知识的复杂性导致的图谱构建过程缓慢问题，通过解耦合业务专家和技术专家，赋能业务专家依据业务问题灵活配置构建领域知识图谱，通过建立基于行业 schema 的 SA-KBQA 解决行业知识图谱在知识问答方向的快速落地问题，最后通过美妆和汽车领域实际项目验证了该框架可以有效加快行业知识图谱的落地和应用。同时，在领域知识图谱构建的过程中，如何将通用领域的知识图谱和领域知识图谱进行有效结合、基于小样本的知识抽取以及领域间的知识图谱的迁移是未来该框架继续研究的方向。

参考文献:

[1] WU Minghui, WU Xindong. On big wisdom[C]//Proceedings of the IEEE International Conference on Data Mining. Singapore, Singapore, 2018:1−2.

[2] NADEAU D, SEKINE S. A survey of named entity recognition and classification[J]. Lingvisticæ investigationes, 2007, 30(1): 3–26.

[3] GRISHMAN R, SUNDHEIM B. Message understanding conference-6: a brief history[C]//Proceedings of the 16th Conference on Computational Linguistics. Copenhagen, Denmark, 1996:466−471.

[4] RABINER L R. A tutorial on hidden Markov models and selected applications in speed recognition[J]. Proceedings of IEEE, 1989, 77(2): 257–286.

[5] SEKINE S, GRISHMAN R, SHINNOU H. A decision tree method for finding and classifying names in Japanese texts[C]//Proceedings of the 6th Workshop on Very Large Corpora.Montreal. Quebec, Canada, 1998:171−178.

[6] BORTHWICK A, STERLING J, AGICHTEIN E, et al. NYU: description of the MENE named entity system as used in MUC-7[C]//Proceedings of the 7th Message Understanding Conference. Washington, DC, USA, 1998.

[7] ASAHARA M, MATSUMOTO Y. Japanese named entity extraction with redundant morphological analysis[C]//Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology. Edmonton, Canada, 2003:8−15.

[8] LAFFERTY J D, MCCALLUM A, PEREIRA F C N. Conditional random fields: probabilistic models for segmenting and labeling sequence data[C]//Proceedings of the Eighteenth International Conference on Machine Learning. San Francisco, CA, United States, 2001: 282−289.

[9] LAMPLE G, BALLESTEROS M, SUBRAMANIAN S, et al. Neural architectures for named entity recognition[C]//Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. San Diego, California, USA, 2016: 260−270.

[10] ZHANG Yue, YANG Jie. Chinese NER using lattice LSTM[C]//Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics. Melbourne, Australia, USA, 2018: 1554−1564.

[11] DEVLIN J, CHANG Mingwei, LEE K, et al. BERT: pre-
training of deep bidirectional transformers for language understanding[C]//Proceedings of the 2019 Conference of
the North American Chapter of the Association for Computational Linguistics: Human Language Technologies Minneapolis, Minnesota, USA, 2019: 4171−4186.

[12] ZHANG Zhengyan, HAN Xu, LIU Zhiyuan, et al. ERNIE: enhanced language representation with informative entities[C]//Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. Florence, Italy, 2019: 1441−1451.

[13] LI X , YAN H , QIU X , et al. FLAT: Chinese NER using flat-lattice transformer[C]// Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. Seattle, Washington, USA, 2020:6836−6842.

[14] YE Z X , LING Z H . Distant supervision relation extraction with intra-bag and inter-bag attentions[C]// Proceedings of the 2019 Conference of the North. Minneapolis, Minnesota, USA,2019:2810−2819.

[15] QIN Pengda, XU Weiran, WANG W Y. Robust distant supervision relation extraction via deep reinforcement learning[C]// Meeting of the Association for Computational Linguistics. Melbourne, Australia , 2018: 2137−2147

[16] ALT C, HÜBNER M, HENNIG L. Fine-tuning pre-trained transformer language models to distantly supervised relation extraction[J]. arXiv preprint arXiv: 1906.08646, 2019.

[17] YE Zhixiu, LING Zhenhua. Multi-level matching and aggregation network for few-shot relation classification[J]. arXiv preprint arXiv:1906.06678, 2019.

[18] SNELL J, SWERSKY K, ZEMEL R. Prototypical networks for few-shot learning[C]//Proceedings of the 31st Conference on Neural Information Processing Systems. Long Beach, CA, USA, 2007: 4077−4087.

[19] SHEN Wei, WANG Jianyong, HAN Jiawei. Entity linking with a knowledge base: issues, techniques, and solutions[J]. IEEE transactions on knowledge and data engineering, 2015, 27(2): 443−460.

[20] SEVGILI O, SHELMANOV A, ARKHIPOV M, et al. Neural entity linking: a survey of models based on deep learning[J]. arXiv:2006.00575, 2021.

[21] GU Yingjie, QU Xiaoye, WANG Zhefeng, et al. Read, retrospect, select: an MRC framework to short text entity linking[J]. arXiv:2101.02394, 2021.

[22] WU L, PETRONI F, JOSIFOSKI M, et al. Scalable zero-shot entity linking with dense entity retrieval[C]//Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing. PuntaCana, The Dominican Republic,2020: 6397−6407.

[23] LIN Yankai, LIU Zhiyuan, SUN Maosong, et al. Learning entity and relation embeddings for knowledge graph completion[C]//Proceedings of the 29th AAAI Conference on Artificial Intelligence. Austin, Texas, 2015: 2181−2187.

[24] SOCHER R, CHEN Danqi, MANNING C D, et al. Reasoning with neural tensor networks for knowledge base completion[C]//Proceedings of the 26th International Conference on Neural Information Processing Systems. Lake Tahoe, Nevada, 2013: 926−934.

[25] WANG Zhen, ZHANG Jianwen, FENG Jianlin, et al. Knowledge graph embedding by translating on hyperplanes[C]//Proceedings of the 28th AAAI Conference on Artificial Intelligence. Quebec, Canada, 2014: 1112−1119.

[26] SHI B X, WENINGER T. Open-world knowledge graph completion[C]//Proceedings of the 32nd International Conference on Artificial Intelligence. New Orleans, Louisiana, USA, 2018: 1957−1964.

[27] LAO Ni, COHEN W W. Relational retrieval using a combination of path-constrained random walks[J]. Machine learning, 2010, 81(1): 53−67.

[28] GARDNER M, TALUKDAR P, KRISHNAMURTHY J, et al. Incorporating vector space similarity in random walk inference over knowledge bases[C]//Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing. Doha, Qatar, 2014: 397−406.

[29] XIONG Wenhan, HOANG T, WANG W Y. DeepPath: a reinforcement learning method for knowledge graph reasoning[C]//Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing. Copenhagen, Denmark, 2017: 564−573.

[30] XIONG Wenhan, YU Mo, CHANG Shiyu, et al. One-shot relational learning for knowledge graphs[C]//Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing. Brussels, Belgium, 2018: 1980−1990.

[31] YIH W T, HE Xiaodong, MEEK C. Semantic parsing for single-relation question answering[C]//Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics. Baltimore, Maryland, 2014: 643−648.

[32] REDDY S, LAPATA M, STEEDMAN M. Large-scale semantic parsing without question-answer pairs[J]. Transactions of the association for computational linguistics, 2014, 2: 377−392.

[33] XU Kun, WU Lingfei, WANG Zhiguo, et al. Exploiting rich syntactic information for semantic parsing with graph-to-sequence model[C]//Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing. Brussels, Belgium, 2018: 918−924.

[34] BORDES A, USUNIER N, CHOPRA S, et al. Large-scale simple question answering with memory networks[J]. arXiv preprint arXiv:1506.02075, 2015.

[35] DONG Li, WEI Furu, ZHOU Ming, et al. Question answering over freebase with multi-column convolutional neural networks[C]//Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing. Beijing, China, 2015: 260−269.

[36] ZHANG Yuanzhe, LIU Kang, HE Shizhu, et al. Question answering over knowledge base with neural attention combining global knowledge information[J]. arXiv preprint arXiv:1606.00979, 2016.

[37] ZHANG Yuyu, DAI Hanjun, KOZAREVA Z, et al. Variational reasoning for question answering with knowledge graph[C]//Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence.San Francisco, California, USA, 2018:1−13.

[38] LIN X V, SOCHER R, XIONG Caiming. Multi-hop knowledge graph reasoning with reward shaping[C]//Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing. Brussels, Belgium, 2018: 3243−3253.

作者简介：

于皓，博士研究生，主要研究方向为机器学习、知识图谱和自然语言处理。

张杰，博士研究生，主要研究方向为知识工程、自然语言处理。

吴信东，教授，博士生导师，主要研究方向为数据挖掘、大数据分析、知识工程。明略科技集团首席科学家和明略科学院院长，教育部大数据知识工程重点实验室（合肥工业大学）主任，营销智能国家新一代人工智能开放创新平台负责人，国家重点研发计划重点专项项目“大数据知识工程基础理论及其应用研究”首席科学家。