软件学报 ISSN 1000-9825, CODEN RUXUEW
*Journal of Software*, 2022,33(7):2646−2666 [doi: 10.13328/j.cnki.jos.006313]

E-mail: jos@iscas.ac.cn
http://www.jos.org.cn
Tel: +86-10-62562563

# 基于互联网群体智能的知识图谱构造方法*

蒋 逸[1,2], 张 伟[1,2], 王 佩[1,2], 张馨月[1,2], 梅 宏[1,2]

[1](高可信软件技术教育部重点实验室(北京大学), 北京 100871)

[2](北京大学 计算机学院, 北京 100871)

通信作者: 张伟, E-mail: zhangw.sei@pku.edu.cn; 梅宏, E-mail: meih@pku.edu.cn

**摘 要**: 知识图谱是一种基于图的结构化知识表示方式. 如何构造大规模高质量的知识图谱, 是研究和实践面临的一个重要问题. 提出了一种基于互联网群体智能的协同式知识图谱构造方法. 该方法的核心是一个持续运行的回路, 其中包含自由探索、自动融合、主动反馈 3 个活动. 在自由探索活动中, 每一参与者独立进行知识图谱的构造活动. 在自动融合活动中, 所有参与者的个体知识图谱被实时融合在一起, 形成群体知识图谱. 在主动反馈活动中, 支撑环境根据每一参与者的个体知识图谱和当前时刻的群体知识图谱, 向该参与者推荐特定的知识图谱片段信息, 以提高其构造知识图谱的效率. 针对这3个活动, 建立了一种层次式的个体知识图谱表示机制, 提出了一种以最小化广义熵为目标的个体知识图谱融合算法, 设计了情境无关和情境相关两种类型的信息反馈方式. 为了验证所提方法及关键技术的可行性, 设计并实施了 3 种类型的实验: 仅包含结构信息的仿真图融合实验、大规模真实知识图谱的融合实验, 以及真实知识图谱的协同式构造实验. 实验结果表明, 该知识图谱融合算法能够有效利用知识图谱的结构信息以及节点的语义信息, 形成高质量的知识图谱融合方案; 基于“探索-融合-反馈”回路的协同方法能够提升群体构造知识图谱的规模和个体构造知识图谱的效率, 并展现出较好的群体规模可扩展性.

**关键词**: 人类群体智能; 互联网; 知识图谱; 知识图谱融合

**中图法分类号**: TP18

中文引用格式: 蒋逸, 张伟, 王佩, 张馨月, 梅宏. 基于互联网群体智能的知识图谱构造方法. 软件学报, 2022, 33(7): 2646–2666. http://www.jos.org.cn/1000-9825/6313.htm

英文引用格式: Jiang Y, Zhang W, Wang P, Zhang XY, Mei H. Knowledge Graph Construction Method via Internet-based Collective Intelligence. Ruan Jian Xue Bao/Journal of Software, 2022, 33(7): 2646−2666 (in Chinese). http://www.jos.org.cn/1000-9825/6313.htm

## Knowledge Graph Construction Method via Internet-based Collective Intelligence

JIANG Yi[1,2], ZHANG Wei[1,2], WANG Pei[1,2], ZHANG Xin-Yue[1,2], MEI Hong[1,2]

[1](Key Laboratory of High Confidence Software Technology (Peking University), Ministry of Education, Beijing 100871, China)

[2](School of Computer Science, Peking University, Beijing 100871, China)

**Abstract**: Knowledge graph is a graph-based structural representation of knowledge. One of the key problems about knowledge graph in both research and practice is how to construct large-scale high-quality knowledge graphs. This paper presents an approach to construct knowledge graphs based on Internet-based human collective intelligence. The core of this approach is a continuously executing loop, called the EIF loop or EIFL, consisting of three activities: free exploration, automatic integration, and proactive feedback. In free exploration activity, each participant tries to construct an individual knowledge graph alone. In automatic integration activity, all participants' current individual knowledge graphs are integrated in real-time into a collective knowledge graph. In proactive feedback activity, each participant is provided with personalized feedback information from the current collective knowledge graph, in order to improve the participant's efficiency of constructing an individual knowledge graph. In particular, a hierarchical knowledge graph representation mechanism is proposed, a knowledge graph merging algorithm is designed driven by the goal of minimizing the collective

* 基金项目: 科技创新 2030——“新一代人工智能”重大项目(2020AAA0109402); 国家自然科学基金(61690200)
收稿时间: 2020-08-14; 采用时间: 2021-01-27; jos 在线出版时间: 2021-08-03

knowledge graph's general entropy, and two ways for context-dependent and context-independent information feedback are introduced, repectively. In order to investigate the feasibility of the proposed approach, three kinds of experiment are designed and carried out: (1) the merging experiment on simulated graphs with structural information only; (2) the merging experiment on real large-scaled knowledge graphs; (3) the construction experiment of knowledge graphs with different number of participants. The experimental results show that: (1) the proposed knowledge graph merging algorithm can find high-quality merging solutions of knowledge graphs by utilizing both structural information of knowledge graphs and semantic information of elements in knowledge graphs; (2) EIFL-based collective collaboration improves both the efficiency of participants in constructing individual knowledge graphs and the scale of the collective knowledge graph merged from individual knowledge graphs, and shows sound scalability with respect to the number of participants in knowledge graph construction.

**Key words**: human collective intelligence; Internet; knowledge graph; knowledge graph merging

知识图谱(knowledge graph)是一种基于图(graph)的结构化知识表示方式. 一个图通常由一组节点以及节点间的关系构成. 采用图的方式对知识进行表示, 反映了一种以关系为核心的知识观, 即知识蕴含在关系中.

人类文明发展到目前的阶段, 已经累积形成了海量的知识资源. 其中, 相当部分的知识以自然语言这种非结构化的方式存在. 随着人类社会的持续发展, 人类知识的规模和复杂度也在不断增长. 持续增长的非结构化知识资源对知识的管理、传播与再生产的负面影响日益显著. 通过将知识表示为一组节点及其之间的关系, 知识图谱能够帮助人类和计算机更好地管理、理解与使用海量的知识资源, 对于促进人类文明的持续发展具有重要意义.

设想一项知识图谱构造任务: 建立《红楼梦》一书中所有人物之间的关系图. 粗略一想, 大概有如下几种方式去完成这项任务.

(1) 一个人手工完成. 找到一本《红楼梦》图书, 逐页阅读, 提取其中的人物及人物之间的关系信息. 可以想象, 即使是一个对红楼梦非常了解的人, 也需要耗费数月甚至更长的时间去完成这一任务. 即便如此, 也不能保证结果的正确性和完整性.

(2) 基于软件算法的自动构造. 采用某种自然语言处理算法, 自动从《红楼梦》的文字信息中抽取出人物关系信息. 这是一个看起来非常完美的解决方案. 但其有效性依赖于一个基本假设, 即自然语言处理算法在该问题上具备了相当于(或超过)人类个体的自然语言理解及分析能力. 目前的技术进展还不能满足这一假设.

(3) 基于软件算法的自动构造+人工修正. 这种方式将上述两种方式结合起来, 能够进一步提高所构造的知识图谱的质量.

(4) 几个好友一起手工完成. 几个好友分别阅读《红楼梦》的不同章节, 提取其中的人物关系信息. 与单人方式相比, 采用多人方式去完成这项任务, 在满足如下条件的情况下会有更高的效率: 一, 这几个好友对红楼梦有一定程度的了解; 二, 这几个好友愿意花费一段时间全身心地投入到这项任务中; 三, 按照章节的方式去分工, 不会导致人物关系信息的大量丢失; 四, 这几个好友具有良好的协同能力.

在互联网环境下, 还有另外一种方式去完成这项任务, 即采用协同式众包的方式. 这种方式大概可以理解为是"几个好友一起手工完成"在互联网技术支持下的规模扩展版本. 在互联网技术的支持下, 任何个体都可以自由加入到这项任务中, 在其中贡献自己所知道的红楼梦人物关系信息片段, 或对其他人创建的信息片段的正确性/准确性进行判断; 然后, 通过某种方式将所有参与者提交的信息片段拼接在一起, 形成完整的红楼梦人物关系图. 本文关注的也正是这样一种知识图谱构造方式.

需要指出的是, 这种协同式众包与目前主流的两种众包实践(即竞争性众包、微任务众包)具有一定的差异性. 首先, 协同式众包不是竞争性众包. 所谓竞争性众包, 是指由若干团队各自独立地完成一项任务, 然后通过某种方式确定完成质量最好的一个团队, 向其支付酬金; 其他团队的工作结果不会被采纳, 也不会获得任何酬金. 而在协同式众包中, 协同的特点更显著一些, 竞争的特点则相对微弱. 其次, 协同式众包也不完全是微任务众包. 所谓微任务众包, 是指众包任务本身就是由一组离散的微任务组成. 例如, 对于"为一个图片库中的所有图片添加文字标注"这一众包任务, 其实际上是由一组"为一个特定图片添加文字标注"的微任务所组成; 完成了所有的微任务, 就相当于完成了这一众包任务. 在协同式众包中, 可能并不存在一组事先定义

的子问题, 而是由参与者自发地识别出当前众包任务的子问题并提交相应的解决方案信息. 另外, 与微任务众包相比, 协同式众包还增加了信息拼接的内容, 即需要采用某种方式把不同个体提交的片段信息拼接在一起. 相比较而言, 协同式众包是一种更关注协同、更为智能的众包. 在本文中, 我们将这种类型的众包定位为一种互联网群体智能, 进而将这种构造知识图谱的方式称为“基于互联网群体智能的知识图谱构造”.

抽象而言, 本文探索采用基于互联网人类群体智能的方式来构造知识图谱并促使其持续演化, 即通过人类个体基于互联网的大规模群体协同, 来构造和演化知识图谱. 在该方式中, 每一人类个体都可以自由加入到知识图谱的构造活动中, 在其中贡献自己的力量, 形成某种形式的大规模群体协同. 该方式的可行性体现在 3 个方面. (1) 在互联网环境下, 涌现出了面向众多复杂问题求解的群体智能现象, 为基于群体智能的知识图谱构造和演化提供了参考性示例. (2) 人类个体, 在某种意义上, 是一个天然的高质量自然语言分析程序. (3) 知识图谱具有的图结构, 使得知识图谱的构造问题具有良好的可分解性, 使得每一参与者都可以低成本地参与到知识图谱的构造活动中: 每一参与者可以把自己知道的信息转化为相应的知识图谱片段; 然后自动化算法对个体片段信息进行拼接, 形成更为完整的知识图谱. 这种基于群体智能的知识图谱构造方式, 其核心技术难点在于如何对大规模参与者群体提交的海量信息片段进行有效的融合与反馈, 使得在群体层面上形成一致、准确的高质量知识图谱.

具体而言, 本文提出了一种基于群体智能的知识图谱构造方法. 该方法的核心是一个持续运行的回路(如图 1 所示), 称为“探索-融合-反馈”[1]回路. 该回路包含了 3 个并行的活动: 自由探索、自动融合、主动反馈. 其中, 第 1 个活动由人类参与者实施, 后两个活动由支撑环境自动实施. 在自由探索活动中, 每一参与知识图谱构造的人类个体独立进行知识图谱的构造活动, 不与其他参与者发生直接的交互. 在任一时刻, 对于每一参与者而言, 其探索活动的输出是一个个体知识图谱. 在自动融合活动中, 支撑环境实时地将所有参与者当前各自的探索结果融合在一起, 形成当前时刻的群体知识图谱. 在主动反馈活动中, 支撑环境根据每一参与者当前的个体知识图谱以及当前的群体知识图谱, 向该参与者推荐特定的知识图谱片段信息, 以提高其构造知识图谱的效率. 每一参与者自主决定是否接受、拒绝或忽略支撑环境提供的反馈信息. 参与者对反馈信息的响应会被记录下来, 用于评估个体的知识偏好以及群体对特定信息的接受程度.

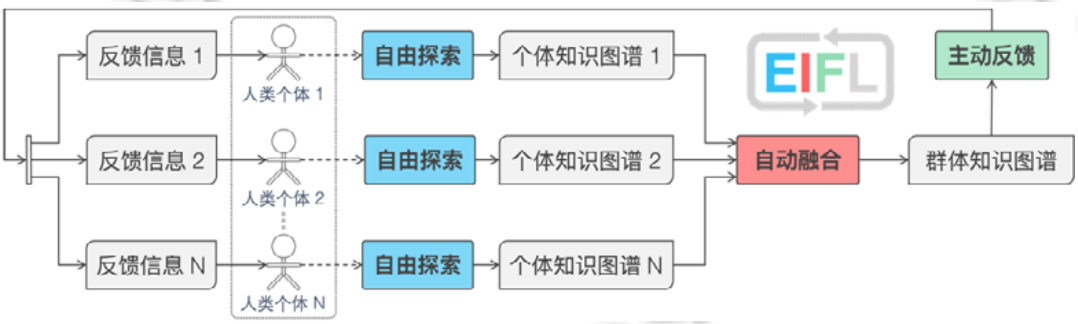



图 1 基于群体智能的知识图谱构造框架

为了验证所提方法及关键技术的可行性, 我们设计并实施了 3 种类型的实验: 仅包含结构信息的仿真图融合实验、大规模真实知识图谱的融合实验以及真实知识图谱的协同式构造实验. 第 1 类实验的目的是为了观察本文提出的知识图谱融合算法对图结构信息的利用能力; 第 2 类实验的目的是为了验证算法对图结构信息和节点语义信息的融合能力; 第 3 类实验的目的是为了考察本文提出的协同式知识图谱构造方法的可行性.

为了实施第 3 类实验, 我们开发了一个支持“探索-融合-反馈”回路的多人在线知识图谱构造环境, 并分别在 1、2、4、8 人规模的参与者群体中进行了真实的知识图谱构造实验. 实验结果表明: (1) 本文提出的知识图谱融合算法能够有效利用知识图谱的结构信息以及节点的语义信息, 形成高质量的知识图谱融合方案(在两个真实知识图谱融合数据集上, 相比较目前最好的知识图谱融合算法, 本文算法在 Hit@1 指标上分别实现了

2.24%和 11.4%的提升); (2) 基于“探索-融合-反馈”回路的协同方法能够提升群体构造知识图谱的规模和个体构造知识图谱的效率, 并展现出较好的群体规模可扩展性(在相同时间内, 相比较单人独立构造知识图谱, 8 人协同构造形成的群体知识图谱的规模提升了约 11 倍, 且参与者的单人构造效率提升了约 1.5 倍).

本文的主要贡献包含如下 4 点: 一种基于“探索-融合-反馈”回路的协同式知识图谱构造方法; 一种层次式的个体知识图谱表示机制; 一种以最小化广义熵为目标的迭代式个体知识图谱融合算法; 一个支持“探索-融合-反馈”回路的多人在线知识图谱构造环境.

本文第 1 节对知识图谱和群体智能两方面的相关研究工作进行简要总结. 第 2 节提出基于“探索-融合-反馈”回路的协同式知识图谱构造方法, 并对其中的关键技术进行详细阐述. 第 3 节通过 3 类实验对本文所提方法和关键技术的可行性进行验证. 第 4 节总结全文, 并对下一步研究工作进行简要说明.

## 1 相关工作

### 1.1 知识图谱的构建

知识图谱最早可以追溯到 20 世纪 60 年代的语义网络(semantic network)以及 20 世纪 70 年代的专家系统(expert system). 在这一时期, 领域专家是知识的主要来源, 知识图谱主要通过单一个体或小规模群体手工构造的方式完成. 2000 年左右, Tim Berners-Lee 提出了语义网(semantic Web)和关联数据(linked data)的概念[2], 其目是为互联网中存在的海量数据信息提供一种标准的描述框架, 从而促成大规模知识的结构化表示、互联与共享. 2012 年, 谷歌正式提出了知识图谱(knowledge graph)的概念, 将其用于语义化搜索, 展现出泛在的应用前景. 在此之后, 知识图谱得到了工业界和学术界的广泛关注.

知识图谱在实践和研究中的一个重要问题是: 如何构造大规模高质量的知识图谱. 目前, 知识图谱的构造方式大致可分为两类: 人工构造和自动化构造.

#### 1.1.1 人工构建

早期的知识图谱主要依靠单一个体或小规模群体进行人工构造. 这一时期的典型工作包括 Cyc 和 WordNet 这两个知识图谱构造项目. Cyc 通过手工构造的方式将专家知识表示为一阶逻辑形式[3]. WordNet 则主要依靠语言学专家手工输入词语之间的语义关系[4]. 随着互联网的普及与发展, 众包成为一种新的知识图谱构造方式. 例如, Freebase 项目采用类似维基百科的方式将知识图谱的创建、修改、查看权限对外开放, 使得互联网上的任一用户都可以自由创建和编辑知识图谱[5]. DBpedia 项目将知识图谱构造任务进行微任务化, 由大规模志愿者群体手工完成对维基百科中自然语言知识的结构化表示[6].

通过人工方式构造形成的知识图谱具有较高的准确性、可用性和可信性. 但是, 受到构造者个体能力的限制, 这种方式存在知识覆盖面窄, 更新缓慢等问题. 虽然互联网众包大大提高了知识图谱的构造规模, 但这种方式仍然存在对一个小规模核心专家群体的强依赖. 例如, 不同用户提交的数据之间存在的不一致性, 仍然需要由社区核心成员进行裁决[7,8].

#### 1.1.2 自动化构造

知识图谱的自动化构造算法大致可以分为基于规则和基于统计两种类别. 在基于规则的构造算法中, 需要由领域专家事先给定适用于特定数据集的知识抽取、融合以及补全规则[9−12], 然后算法将这些规则应用到特定的数据集上, 形成知识图谱. 基于统计的构造算法则自动识别特定领域数据源的统计特征, 并自动完成知识图谱的构造[13−16]. 目前, 主流的基于统计的自动化构造算法普遍采用监督学习的方式, 依赖于事先人工标注的大规模训练数据集, 且针对不同的问题领域需要建立不同的训练数据集. 针对开放领域存在的样本数据稀疏问题, 也有学者探索采用弱监督学习的方式进行知识图谱的自动化构造[17,18].

自动化算法在一定程度上提高了知识图谱的构造效率, 降低了构造成本, 但仍然存在两个基本问题. (1) 自动化算法, 特别是采用监督学习的知识图谱构造算法, 严重依赖于训练数据集的规模和质量. (2) 在可以预见的将来, 自动化算法所具有的对一般性非结构化知识的理解能力还远远达不到人类个体的能力, 这在很大程度上限制了自动化算法的应用范围. 在谷歌搜索引擎使用的知识图谱中, 就大量包含了 Freebase 项目

中由人工方式构造的知识谱图信息[19,20]. 一些研究工作也表明, 在自动化构造知识图谱的过程中, 加入人类的反馈信息, 能够明显提升知识图谱的构造质量[21−23].

### 1.2 知识图谱的表示

早期对知识表示的研究, 主要关注于建立形式化的逻辑语义表示机制, 从而支持对知识的有效推理. 20 世纪 60 年代 Collins 等人[24]提出了语义网络(semantic network)的概念, 试图通过网络结构表示实体之间的语义关系. 20 世纪 70 年代兴起的专家系统[25]提出了更为形式化的知识表示机制, 主要包括: 产生式表示法(production rule pepresentation), 通过 IF-THEN 的结构支持知识的推理; 框架表示语言(frame representation language), 通过“槽”描述对象可能存在的属性和关联; 一阶逻辑(first-order logic), 支持量化和断言的命题逻辑, 通过演算支持知识的推理. 1985 年, Brachman 等人[26]在 KL-ONE 系统中使用描述逻辑(description logic)对知识进行表示, 其主要包含公理集合和断言集合两部分. 描述逻辑是一阶逻辑的一个可判定子集, 能够支持对一致性(consistency)、可满足性(satisfiability)、包含检测(subsumption)、实例检测(instance checking)等性质的判断.

随着互联网的发展, 知识表示的一个重要任务是为互联网中存在的海量数据信息提供一种统一的描述框架, 从而促进大规模知识的结构化表示、互联与共享. 与早期的知识表示相比, 现代知识图谱(如 Freebase、Yago、Wikidata 等)均弱化了对逻辑语义表达的要求, 而强调大规模的事实型知识. 其中, 资源描述框架(resource description framework, RDF)是对事实型知识的一种主流表示方式, 即通过⟨主语, 谓语, 宾语⟩三元组的形式, 表示知识图谱中实体及其之间的关系. 同时, 通过 RDF 范式(RDF schema)、元数据(metadata)等方式对 RDF 的语义信息进行轻量级的描述[2].

随着基于深度神经网络的表示学习技术的发展, 知识的向量化表示成为一个重要研究方向. 通过知识嵌入(embedding), 将实体和关系的语义信息表示为对应的向量, 实体之间的关系可以通过向量计算得到, 减少了对图的拓扑结构的依赖. 知识的向量化表示能够有效地支持大规模知识图谱中的知识查询和知识补全. Trans 系列工作是知识向量化表示的典型代表. 该系列工作基于翻译模型, 将知识图谱中的实体转换为词向量, 并将实体间的关系视作两个实体间的翻译关系. 在 TransE 方法[13]中, 源实体通过关系被直接翻译为目标实体, 所以当源/目标实体和关系确定时目标/源实体也是确定的. 这导致 TransE 方法无法支持一个实体拥有多个同类关系的情况, 与知识图谱的实际表达能力不符. Wang 等人提出了 TransH 方法[27], 以应对实体间可能存在多种同类关系这一客观情况. TransH 的核心思想是在翻译过程中仅关心实体中与当前关系相关的维度信息, 且在翻译前需要先将实体投影到关系所在的超平面. Lin 等人提出了 TransR 方法[28], 其核心思想是将实体和关系建模在两个不同的空间中, 从而减小了空间维度, 能够在一定程度上避免过拟合问题, 在实际数据中取得了更好的补全效果.

### 1.3 群体智能

#### 1.3.1 自然界中的群体智能

长久以来, 科学家在很多社会性昆虫群体中观察到了一种看似矛盾的现象: 每一昆虫个体不具有或仅具有有限的智能, 但一个昆虫群体却能在群体层次上展现出远超个体的智能行为. 这种在昆虫群体层次上展现出的智能行为, 被称为群体智能(swarm/collective intelligence)[29,30]. 从群体智能现象中可以观察到群体智能具有的一个基本性质, 即对个体智能的放大效果.

研究者提出了环境激发效应[31]这一概念, 用于解释社会性昆虫的群体智能现象. 环境激发效应指代了一种发生在昆虫个体之间以物理环境为媒介的间接交互机制. 基于这一概念, 昆虫群体中的群体智能现象通过如下过程涌现形成:昆虫个体在物理环境中留下自己的踪迹, 或对物理环境作出某种改变; 这些踪迹或改变被群体中的个体感知到, 并刺激这些个体在环境中留下新的踪迹或对环境作出进一步的改变; 因此, 个体行为之间实现了有效的协同, 并形成了一个正反馈回路, 进而在群体层次上表现出智能的自组织行为. 环境激发效应解释了群体智能具有的另外一个基本性质: 群体协同规模的可扩展性.

物理空间中存在的群体智能现象指出了信息空间(cyberspace)中一种潜在的大规模人类群体协同方式[1]. 主要基于如下两点原因: (1) 基于当前的研究, 群体智能蕴含了一种能够有效放大个体智能的大规模群体协同机制. (2) 与物理空间中大规模群体聚集的高成本相比, 在信息空间中更容易实现大规模人群的低成本聚集. 如果能够将群体智能的基本原理成功应用到信息空间中的大规模人类群体上, 实现对人类个体智能的有效放大, 那么, 我们认为, 这将极大地释放人类社会具有的潜在创造力, 促进人类文明的进一步发展[32].

1.3.2 基于互联网的人类群体智能

互联网上已经出现了很多人类群体智能现象或系统, 为很多领域带来了创新性的问题求解方法. 其中, 一些群体智能现象/系统是长期的社会-技术协同演化的产物, 另一些则是针对特定的问题精心设计的群智化求解系统. 例如, 在软件工程领域, 经过数十年的演化, 开源软件开发[33]已经成为一种重要的社会-技术现象; 在其中, 地理分布的大规模开发者群体通过互联网进行有效的协同, 成功开发出数量众多的高质量复杂软件应用. 在单项选择题求解领域, UNU 系统[34]提供了一个有趣的多人在线环境, 可以支持一个大规模群体通过持续协同的方式确定一个单项选择题的答案, 在很多实际场景中的预测和决策问题上表现出很高的准确率. 在生物学研究领域中, EteRNA 系统[35]提供了一个多人在线游戏, 通过大规模非专业个体的持续协同求解复杂的蛋白质结构问题.

群体智能的研究还远远落后于实践; 现有的研究成果几乎没有对人工群体智能系统的构造产生实质性的影响. 目前存在的较为成功的人工群体智能系统都不是在任何成熟的群体智能理论的指导下构造形成的. 主要原因在于, 目前的研究工作主要关注群体智能的解释型理论(即如何解释某一群体智能现象的形成机理), 而较少触及群体智能的构造型理论(即如何可控地构造求解特定问题的群体智能系统). 一个典型案例是环境激发效应. 这一概念在提出时是用于解释社会性昆虫群体中群体智能现象[31], 而且近年来也被广泛用于分析和解释人类群体智能现象[36,37]. 我们认为, 环境激发效应提供了一种针对群体智能的解释性模型, 能够对已经存在的群体智能现象进行有效的事后分析. 但是, 这一概念能够在何种程度上有效指导一个人工群体智能系统的构造, 仍然需要进一步的观察和确认.

## 2 方 法

本节介绍一种基于互联网群体智能的知识图谱构造方法. 该方法的核心是一个持续运行的回路, 包含 3 个并行的活动: 自由探索、自动融合、主动反馈. 本节分别对这 3 个活动及其中的基本概念和关键技术进行说明.

### 2.1 自由探索

在自由探索活动中, 每一参与知识图谱构造的人类个体独立进行知识图谱的构造活动, 不与其他参与者发生直接的交互. 在任一时刻, 对于每一参与者而言, 其探索活动的输出是一个个体知识图谱.

2.1.1 个体知识图谱

个体知识图谱的表示需要考虑两个方面的因素. 一方面, 所采用的表示机制应该具备有效的抽象性和良好的可扩展性, 从而支持对不同领域中存在的多样性知识片段进行有效的建模. 另一方面, 这种表示机制应该能够支持算法有效识别不同知识图谱之间的共性和差异性, 从而实现对群体知识的有效融合与反馈. 基于上述考虑, 我们设计了一种层次式的个体知识图谱, 支持对二元关系、多元关系以及高阶关系的统一标识, 且可以被方便地转换为一种边上带标签的有向图, 从而基于图结构进行多源信息的分析、融合与反馈.

**定义 1(个体知识图谱)**. 个体知识图谱是一个五元组 $K\doteq(K0,K1,K2,K3,K4)$. 其每个元素的定义如下.

1. $K0\doteq(L,V,\ell,\succ,\sqsupseteq,\mathbb{P},\sqcup,\sqcap,\eta,\alpha)$: 个体知识图谱框架, 满足如下条件.

(a) $L\doteq\{0,1,2,3,4\}$: 个体知识图谱中节点具有的 5 个层次. 其中, 0、1、2、3、4 分别表示道层(tao level)、元元模型层(meta-meta-model level)、元模型层(meta-model level)、模型层(model level)、实例层(instance level).

(b) $V$: 个体知识图谱的节点集合.

(c) $\ell$:$V\to L$: 层次映射函数,将个体知识图谱节点映射到其所在的层次. 为方便下文叙述, 令 $V(i)\doteq\{v\in V|\ell(v)=i\}$, 且 $V(-i)\doteq\{v\in V|\ell(v)\neq i\}$, $i\in L$. 前者表示由 $V$ 中处于 $i$ 层的元素构成的集合; 后者表示由 $V$ 中所有不处于 $i$ 层的元素构成的集合.

(d) $\succ\subseteq\bigcup_{i=0}^{3}V(i)\times V(i+1)$: 个体知识图谱节点之间的实例化关系. 对于任何$(u,v)\in\succ$(也记为 $u\succ v$), 表示 $v$ 是 $u$ 的一个实例, 或 $u$ 是 $v$ 的一个类型. 为方便下文描述, 令 $V(\succ v)\doteq\{u\in V|u\succ v\}$, 且 $V(u\succ)\doteq\{v\in V|u\succ v\}$. 前者表示由 $V$ 中所有 $v$ 的类型构成的集合; 后者表示由 $V$ 中所有 $u$ 的实例构成的集合(下文会根据需要将这种表示符号应用到其他集合与二元关系上)). 实例化关系不具有自反性、对称性、传递性. 对任何 $u\succ v$, 有 $\ell(v)=\ell(u)+1$ 成立.

(e) $\sqsupseteq\subseteq\bigcup_{i=1}^{3}V(i)\times V(i)$: 个体知识图谱节点之间的一般特殊关系. 对任何$(g,s)\in\sqsupseteq$(也记为 $g\sqsupseteq s$), 称 $g$ 是 $s$ 的一般概念, 或 $s$ 是 $g$ 的特殊概念, 满足:对任何 $s\succ w$, 有 $g\succ w$ 成立. 也即一个概念的任何一个实例一定是这个概念的一般概念的实例. 对任何 $u,v\in\in V$, 如果 $u\sqsupseteq v$ 且 $v\sqsupseteq u$, 则称 $u,v$ 等价, 记为 $u=v$. 一般特殊关系具有自反性、传递性, 但不具有对称性.

(f) $\mathbb{P}$:$\bigcup_{i=1}^{3}V(i)\nrightarrow V(i-1)$: 个体知识图谱节点之间的幂集关系, 一个部分函数(partial function). 对任何 $(u,v)\in\mathbb{P}$(也记为$\mathbb{P}(u)=v$), 称 $v$ 是 $u$ 的幂概念, 满足:对任何 $v\succ w$, 有 $u\sqsupseteq w$ 成立. 也即一个概念的幂概念的任何一个实例一定是这个概念的一个特殊概念.

(g) $\sqcup$:$\bigcup_{i=1}^{3}V(i-1)\nrightarrow V(i)$: 个体知识图谱节点之间的并集关系, 一个部分函数. 对任何 $u\mapsto v\in\sqcup$(也记为 $\sqcup(u)=v$), 称 $v$ 是 $u$ 的所有实例的并集, 满足: (1) 对任何 $x,y\in V$, 如果 $u\succ x$ 且 $x\succ y$, 则 $v\succ y$ 成立; (2) 对任何 $y\in V$, 如果 $v\succ y$, 则存在 $x\in V$, 有 $u\succ x$ 且 $x\succ y$ 成立. 也即一个概念的所有实例的并集是由这些实例的所有实例构成的集合.

(h) $\sqcap$:$\bigcup_{i=1}^{3}V(i-1)\nrightarrow V(i)$: 个体知识图谱节点之间的交集关系, 一个部分函数. 对任何 $u\mapsto v\in\sqcap$(也记为 $\sqcap(u)=v$), 称 $v$ 是 $u$ 的所有实例的交, 满足: (1) 对任何 $x\in V$, 如果对所有 $y\in V(u\succ)$, $y\succ x$ 成立, 则有 $v\succ x$ 成立; (2) 对任何 $x\in V$, 如果 $v\succ$x, 则对任何 $y\in V(u\succ)$, 有 $y\succ x$ 成立. 也即一个概念的所有实例的交集是由这些实例的共有实例构成的集合.

(i) $\eta$:$V\to V(Str\succ)$: 标识符函数. 将个体知识图谱节点映射到字符串上. $Str$ 是模型层知识图谱的一个节点, 表示由所有字符串构成的集合. 该函数的主要目的是为个体知识图谱中的每一个节点关联一个人类可理解的描述信息.

(j) $\alpha$:$\bigcup_{v\in V(\circledcirc\succ)}V(v\succ)\to V(Str\succ)$: 符号字面量函数. 将 $V$ 中符号概念$\circledcirc$实例的实例映射到字符串上. 符号概念$\circledcirc$是元模型层知识图谱的一个节点. 该函数的主要目的是为每一个符号概念实例的实例关联一个对应的字面量. 不失一般性, 令$\alpha\subseteq\eta$. 也即一个符号的字面量即提供对该符号的一种描述信息.

2. $K1\doteq(\circ_1,\varnothing_1)$: 元元模型层知识图谱, 满足: $\{\circ_1,\varnothing_1\}\subseteq V$. $\circ_1$ 表示元元模型层的满节点, 满足: (1) $\ell(\circ_1)=\ell$; (2) 对于任何 $v\in V(1)$, 有$\circ_1\sqsupseteq v$ 成立. 可知, 对任何$v\in V(2)$, 有$\circ_1\succ v$ 成立. 元素$\varnothing_1$表示元元模型层的空节点, 满足: (1) $\ell(\varnothing_1)=1$; (2) 对于任何 $v\in V(1)$, 有 $v\sqsupseteq\varnothing_1$ 成立. 可知, 不存在 $v\in V(2)$, 使得$\varnothing_1\succ v$ 成立.

3. $K2\doteq(\circ_2,\varnothing_2,\odot,\ominus,\circledcirc,\circledast)$: 元模型层知识图谱, 满足: $\{\circ_2,\varnothing_2,\odot,\ominus,\circledcirc,\circledast\}\subseteq V$. $\circ_2$ 表示元模型层的满节点, 满足: (1) $\ell(\circ_2)=2$; (2) 对任何 $v\in V(2)$, 有$\circ_2\sqsupseteq v$ 成立. 可知, 对任何 $v\in V(3)$, 有$\circ_2\succ v$ 成立. $\varnothing_2$ 表示元模型层的空节点, 满足: (1) $\ell(\varnothing_2)=2$; (2) 对任何 $v\in V(2)$, 有 $v\sqsupseteq\varnothing_2$ 成立. 可知, 不存在 $v\in V(3)$, 使得$\varnothing_2\succ v$ 成立. $\odot$、$\ominus$、$\circledcirc$、$\circledast$分别表示实体概念、关系概念、角色概念、符号概念, 满足$\circ_1\succ\odot$, $\circ_1\succ\ominus$, $\circ_1\succ\circledcirc$, $\circ_1\succ\circledast$.

4. $K3\doteq(\circ_3,\varnothing_3,Str,Int,\twoheadrightarrow,\pi,\kappa,\mathbb{T})$: 模型层知识图谱, 满足如下条件.

(a) $(\circ_3,\varnothing_3,Str,Int)\subseteq V$. $\circ_3$ 表示模型层的满节点, 满足: (1) $\ell(\circ_3)=3$; (2) 对任何 $v\in V(3)$, 有$\circ_3\sqsupseteq v$ 成立. 可知, 对任何 $v\in V(4)$, 有$\circ_3\succ v$ 成立. $\varnothing_3$ 表示模型层的空节点, 满足: (1) $\ell(\varnothing_3)=3$; (2) 对任何 $v\in V(3)$, 有$v\sqsupseteq\varnothing_3$ 成立. 可知, 不存在$v\in V(4)$, 使得$\varnothing_3\succ v$成立. 元素 $Str$、$Int$ 分别表示字符串、整数, 满足$\circledast\succ Str$, $\circledast\succ Int$. 令 $Ints=\mathbb{P}(int)$, 也即 $Ints$ 是 $Int$ 的幂概念.

(b) $\twoheadrightarrow$:$V(\ominus\succ)\leftarrow V(\circledcirc\succ)$: 关系概念实例与角色概念实例之间的关联关系. 其逆关系$\twoheadrightarrow^{-1}$是一个函数, 即任何一个角色概念实例只与一$\twoheadrightarrow$:$V(\ominus\succ)\leftarrow V(\circledcirc\succ)$: 关系概念实例与角色概念实例之间的关联关系. 其逆关系$\twoheadrightarrow^{-1}$是一个函数, 即任何一个角色概念实例只与一个关系概念实例相关.

(c) $\pi$:$V(\circledcirc\succ)\rightarrow V(3)$: 角色概念实例的承担者函数, 将一个角色概念实例映射到模型层知识图谱的节点上. 其具体含义见实例层知识图谱.

(d) $\kappa$:$V(\circledcirc\succ)\rightarrow V(Ints\succ)$: 角色概念实例的承担者数量限制函数, 将一个角色概念实例映射到一个整数集合上. 其具体含义见实例层知识图谱.

(e) $\mathbb{T}\doteq(\tau,\leqslant_\tau,\rightsquigarrow,\nearrow,\nwarrow)$: 关于时间点、时间点先后关系、以及时间区间的模型层知识图谱. 其中, $\tau$表示时间点, 满足$\circledast\succ\tau$. $\leqslant_\tau\subseteq V(\tau\succ)\times V(\tau\succ)$表示时间点之间的先后关系; $\leqslant_\tau$是一个偏序关系(具有自反性、传递性, 但不具有对称性). 对任何$(t_0,t_1)\in\leqslant_\tau$(也记为 $t_0\leqslant_\tau t_1$), 若满足 $t_1\leqslant_\tau t_0$, 则称 $t_0$ 和 $t_1$ 相等(记为 $t_0=t_1$). $\rightsquigarrow$表示时间区间, 满足$\circledast\succ\rightsquigarrow$. $\nearrow$:$V(\rightsquigarrow\succ)\rightarrow V(\tau\succ)$表示一个函数, 将时间区间实例映射到对应的开始时间点实例上. $\nwarrow$:$V(\rightsquigarrow\succ)\rightarrow V(\tau\succ)$表示一个函数, 将时间区间实例映射到对应的结束时间点实例上. 对任何 $p\in V(\rightsquigarrow\succ)$, 有$\nearrow(p)\leqslant_\tau\nwarrow(p)$成立.

5. $K4\doteq(\rho,\circlearrowleft)$: 实例层知识图谱, 满足如下条件.

(a) $\rho\doteq\{(v,r)\mapsto w|\ominus\succ u,u\succ v,u\twoheadrightarrow r,\pi(r)\sqsupseteq w,\kappa(r)\succ|V(w\succ)|\}$: 关系概念实例的实例到角色承担者的映射函数. 对于其中的一个元素$(v,r)\mapsto w$, $v$ 表示一个关系概念的实例 $u$ 的实例, $r$ 表示 $u$ 的一个角色, $w$ 表示角色 $r$ 在 $v$ 上的承担者集合, 且满足: (1) $w$ 是$\pi(r)$的一个特殊概念; (2) $w$ 的实例的数量是$\kappa(r)$中的一个元素. 可以看到, 模型层知识图谱中定义的角色概念实例的承担者函数$\pi$和承担者数量限制函数$\kappa$对$\rho$包含的元素进行了限制.

(b) $\circlearrowleft$:$V(4)\rightarrow\rightsquigarrow$: 实例层节点到其生命周期的映射函数.

该定义给出了一种层次式的知识图谱, 其中包含 5 个层次: 道层、元元模型层、元模型层、模型层、实例层.

个体知识图谱包含的每一个节点都处于且仅处于一个层次中. 相邻层次的节点之间通过实例化关系相互关联. 实例化关系的定义建立在概念外延的基础上, 即将一个概念理解为由其所有实例形成的集合; 若一个元素属于概念的外延集合, 则表明该元素是该概念的一个实例. 除实例层外(不包括实例层), 处于其他层的节点均是概念, 且指代了概念的外延. 个体知识图谱还定义了概念之间的一般特殊关系、幂集关系、并集关系、交集关系. 对于个体知识图谱中的每一个节点, 通过标识符函数, 将该节点与对应的字符串描述信息进行关联. 对于个体知识图谱中的每一个节点, 如果是符号概念$\circledast$实例的实例, 则通过标识符函数将其与对应的字面量进行关联. 对于元元模型层、元模型层、以及模型层, 分别定义了若干基本节点以及节点之间的关系; 需要指出的是, 这些元素不是一个全集, 可以根据实际需要向其中添加新的元素. 实例层包含两个函数: $\rho$函数将关系概念$\ominus$实例的实例映射到涉及角色的承担者; $\circlearrowleft$函数将实例层节点映射到其生命周期. 另外, 对于道层, 由于其中包含的元素(处于元元元模型层或之上)过于抽象, 且不会对知识图谱的构造产生直接的影响, 所以我们没有对其中的元素进行定义.

2.1.2 个体知识图谱的图表示

给定个体知识图谱 $K\doteq(K0,K1,K2,K3,K4)$, 其图表示(graph representation)是一个边上带标签的有向图$\mathcal{G}(K)=(\mathcal{V},\mathcal{E},\mathcal{T})$. 其中, $\mathcal{V}$表示节点集合、$\mathcal{E}\subseteq\mathcal{V}\times\mathcal{T}\times\mathcal{V}$表示带标签的有向边集合、$\mathcal{T}$表示边上标签的集合. $\mathcal{G}(K)$的构造方法见算法 1.

基于个体知识图谱生成对应的图表示的基本思想如下: 把个体知识图谱内置的每一种二元关系包含的每一个元素转化为图表示中两个节点之间一条带标签的有向边; 有向边上的标签即是对应的关系名. 除此之外, 算法 1 还包含对两种例外情况的处理. (1) 对于函数 $\ell$, 把其值域中的 5 个整数分别转化为符号概念实例 1 的 5 个实例 $l_i$, $i\in L$;然后, 把 $\ell$ 中的每个元素$(v,i)$转化节点 $v$ 和 $l_i$ 之间一条标签为“$l$”的有向边. (2) 对于函数$\rho$中的每一个元素$(v,r,w)$, 创建 $r$ 的一个实例$\gamma$;然后, 在节点 $v$ 和$\gamma$之间建立一条标签为“$\twoheadrightarrow$”的有向边, 在节点$\gamma$和 $w$ 之间建立一条标签为“$\rho$”的有向边. 图 2 给出了个体知识图谱图表示的一个示例.

**算法 1**. 个体知识图谱图表示的构造算法.

输入: $K\doteq(K0,K1,K2,K3,K4)$.

输出: $\mathcal{G}(K)\doteq(\mathcal{V},\mathcal{E},\mathcal{T})$.

1. $\mathcal{V}\leftarrow V$, $\mathcal{E}=\varnothing$, $\mathcal{T}\leftarrow\{\ell,\succ,\sqsupseteq,\mathbb{P},\sqcup,\sqcap,\eta,\alpha,\twoheadrightarrow,\circlearrowleft,\pi,\kappa,\leqslant_r,\lceil,\neg,\rho\}$;
2. $\mathcal{R}\leftarrow\mathcal{T}\backslash\{\ell,\rho\}$;
3. **for** $R\in\mathcal{R}$ **do**
4. $\quad\mathcal{E}\leftarrow\mathcal{E}\cup\{(u,R,v)|(u,v)\in R\}$
5. **end for**
6. Create 6 new vertices l and $l_i$, $i\in L$;
7. $\mathcal{V}\leftarrow\mathcal{V}\cup\{\text{l}\}\cup\{l_i|i\in L\}$;
8. $\mathcal{E}\leftarrow\mathcal{E}\cup\{(\circledast,\succ,l)\}\cup\{(l,\succ,l_i)|i\in L\}$;
9. $\mathcal{E}\leftarrow\mathcal{E}\cup\{(v,\ell,l_i)|\ell(v)=i\}\cup\{(l,\ell,l_3)\}\cup\{(l_i,\ell,l_4)|i\in L\}$;
10. $\mathcal{E}\leftarrow\mathcal{E}\cup\{(l,\eta,\text{"知识图谱节点层次"})\}$;
11. $\mathcal{E}\leftarrow\mathcal{E}\cup\{(l_0,\eta,\text{"道层"}),(l_1,\eta,\text{"元元模型层"}),(l_2,\eta,\text{"元模型层"}),(l_3,\eta,\text{"模型层"}),(l_4,\eta,\text{"实例层"})\}$;
12. $\mathcal{E}\leftarrow\mathcal{E}\cup\{(l_i,\alpha,\text{des})|i\in L,(l_i,\eta,\text{des})\in E\}$;
13. **for** $(v,r,m)\in\rho$ **do**
14. $\quad$Create a new vertex $\gamma$,
15. $\quad\mathcal{V}\leftarrow\mathcal{V}\cup\{\gamma\}$;
16. $\quad\mathcal{E}\leftarrow\mathcal{E}\cup\{(r,\succ,\gamma),(\gamma,l,l_4),(\gamma,\eta,\eta(r)),(v,\twoheadrightarrow,\gamma),(\gamma,\rho,w)\}$;
17. **end for**

可以看到, 个体知识图谱及其图表示之间存在一一对应关系, 即从一个个体知识图谱出发, 可以通过算法 1 生成对应的图表示; 从一个个体知识图谱的图表示出发, 也可以还原出对应的个体知识图谱. 在下文中, 为表述简洁, 在没有特殊说明的情况下, 我们用“个体知识图谱”指代“个体知识图谱的图表示”.

### 2.2 自动融合

在自动融合活动中, 支撑环境将所有参与者通过自由探索活动形成的个体知识图谱实时融合在一起, 形成当前时刻的群体知识图谱. 个体知识图谱的融合包含节点融合与边融合两个方面. 节点融合指的是把不同个体知识图谱中具有相同或相似语义的节点融合在一起, 形成群体知识图谱的节点. 对于不同个体知识图谱中具有相同标签的一组边, 如果其对应的源节点和目标节点已经分别被融合在相同的群体知识图谱节点中, 则这组边被融合在一起, 形成群体知识图谱的一条边(即这些边的源节点和目标节点所在的群体知识图谱节点之间一条同方向、同标签的边); 如果个体知识图谱的一条边没有与其他边发生融合, 这条边也会独立形成群体知识图谱的一条边(即这条边的源节点和目标节点所在的群体知识图谱节点之间一条同方向、同标签的边).

#### 2.2.1 群体知识图谱

**定义 2(个体知识图谱的融合方案)**. 给定个体知识图谱集合$\boldsymbol{G}=\{\mathcal{G}_i=(\mathcal{V}_i,\mathcal{E}_i,\mathcal{T}_i)|i\in[1,N]\}$, 称集合$\boldsymbol{M}\subseteq(\mathcal{V}_1\cup\{\perp\})\times(\mathcal{V}_2\cup\{\perp\})\times...\times(\mathcal{V}_N\cup\{\perp\})$(其中, $\perp$表示一个虚节点)为集合$\boldsymbol{G}$中个体知识图谱的一个融合方案, 当且仅当其满足如下条件.

1. $\forall\mathcal{V}_i\in\mathcal{G}_i$, $\forall v\in\mathcal{V}_i$, $\exists!(m_1,m_2,...,m_i,...,m_N)\in\boldsymbol{M}$, $v=m_i$. 即$\boldsymbol{G}$中任一个体知识图谱的任一节点, 存在且仅存在于$\boldsymbol{M}$的一个元素中.

2. $\forall(m_1,m_2,...,m_N)\in\boldsymbol{M}$, $\exists i\in[1,N]$, $m_i\neq\perp$. 即 $\boldsymbol{M}$ 中不存在所有分量均为虚节点的元素.

3. $\forall(m_1,m_2,...,m_N)\in\boldsymbol{M}$, $\forall i,j\in[1,N]$, $(m_i\neq\perp)\wedge(m_j\neq\perp)\Rightarrow\ell(m_i)=\ell(m_j)$. 即 $\boldsymbol{M}$ 中任一元素的所有非虚节点分量具有相同的层次.

给定 $N$ 个个体知识图谱, 其融合方案中的每一个元素是一个长度为 $N$ 的向量: 其中, 位置 $i$ 上的分量, 是一个来自个体知识图谱 $i$ 的节点, 或者是一个虚节点. 融合方案中的一个元素的所有非虚节点分量, 表示一组应该融合在一起的个体知识图谱节点. 给定融合方案中的一个元素 $m$, $m_i$ 表示该元素位置 $i$ 上的分量, $|m|_{\perp}$ 表示该元素中非虚节点分量的个数.

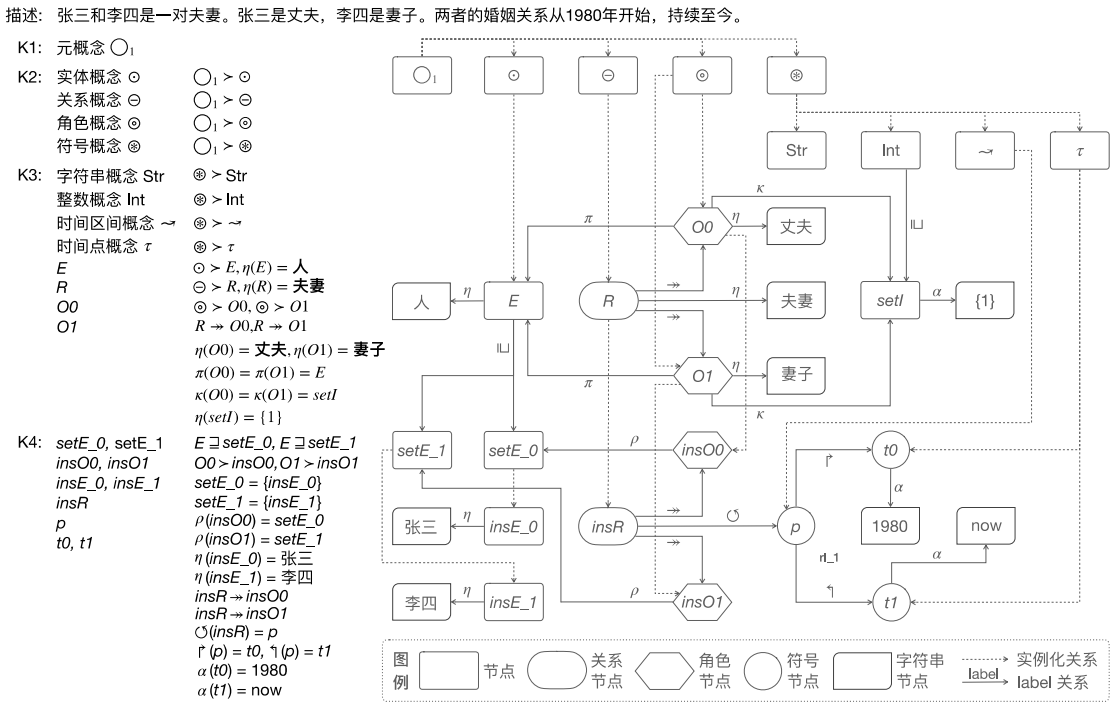



图 2 个体知识图谱的图表示示例

**定义 3(群体知识图谱)**. 给定个体知识图谱集合 $\boldsymbol{G}=\{\mathcal{G}_i=(\mathcal{V}_i,\boldsymbol{\mathcal{E}}_i,\mathcal{T}_i)|i\in[1,N]\}$, 以及 $\boldsymbol{G}$ 的一个融合方案 $\boldsymbol{M}$, 则 $\boldsymbol{G}$ 在 $\boldsymbol{M}$ 下的群体知识图谱是一个图 $\mathcal{G}(\boldsymbol{M})=(\mathbb{V},\mathbb{E},\mathbb{T},edges)$, 其中,

1. $\mathbb{V}=\boldsymbol{M}$: 群体知识图谱的节点集合.

2. $\mathbb{E}=\{(m,t,n)|(m,t,n)\in\mathbb{V}\times\mathbb{T}\times\mathbb{V}, (\exists i\in[1,N], (m_i,t,n_i)\in\boldsymbol{\mathcal{E}}_i)\}$: 群体知识图谱的边集合.

3. $\mathbb{T}=\bigcup_{i=1}^{N}\mathcal{T}_i$: 群体知识图谱中边上标签的集合.

4. $edges=\{(m,t,n)\mapsto\{(m_i,t,n_i)|(m_i,t,n_i)\in\boldsymbol{\mathcal{E}}_i, i\in[1,N]\}|(m,t,n)\in\mathbb{E}\}$: 群智知识图谱中的每一条边到对应的个体知识图谱边集合的映射.

给定一组个体知识图谱的一个融合方案, 则可以唯一确定一个群体知识图谱. 群体知识图谱的节点集合即是融合方案, 即融合方案中的元素与群体知识图谱的节点之间存在一一映射关系. 群体知识图谱的两个节点之间存在一条具有特定标签的边, 当且仅当这两个节点的某个分量 $i$ 上均是非虚节点, 且这两个非虚节点之间存在一条同方向、同标签的边.

### 2.2.2 群体知识图谱的熵

给定个体知识图谱集合 $\boldsymbol{G}=\{\mathcal{G}_i=(\mathcal{V}_i,\boldsymbol{\mathcal{E}}_i,\mathcal{T}_i)|i\in[1,N]\}$, 由其所有可能的融合方案构成的集合, 称为 $\boldsymbol{G}$ 的融合方案空间, 记为 $\mathcal{M}(\boldsymbol{G})$. 对 $\boldsymbol{G}$ 中的个体知识图谱进行融合的目的是从 $\mathcal{M}(\boldsymbol{G})$ 中寻找到一个具有足够高质量的融合方案. 我们通过计算一个融合方案对应的群体知识图谱的熵, 对融合方案的质量进行度量. 在一般意义上, 熵衡量了一个系统的混乱度/有序度: 熵越大, 表明系统的混乱度越高/有序度越低;反之, 则表明混乱度越低/有序度越高. 对于一个群体知识图谱, 我们首先计算其中每一节点的熵, 然后综合形成群体知识图谱的熵. 对于群体知识图谱的每一节点, 我们通过考察在不同个体知识图谱中该节点上下文之间的相似度/差异度, 计算形

成该节点的熵.

**定义 4(群体知识图谱的熵)**. 给定个体知识图谱集合$\boldsymbol{G}=\{\mathcal{G}_i=(\mathcal{V}_i,\mathcal{E}_i,\mathcal{T}_i)|i\in[1,N]\}$、融合方案$\boldsymbol{M}\in\mathcal{M}(\boldsymbol{G})$、以及$\boldsymbol{G}$在$\boldsymbol{M}$下的群体知识图$\mathbb{G}(\boldsymbol{M})=(\mathbb{V},\mathbb{E},\mathbb{T},edges)$, $\mathbb{G}$的熵$\mathcal{H}(\mathbb{G})$的计算方式如下.

$$\mathcal{H}(\mathbb{G})\doteq\sum_{m\in V}\mathcal{H}(m)\cdot(1+\theta(m)) \tag{1}$$

$$\theta(m)\doteq\Delta-\Delta^{|m|_{\perp}/N} \tag{2}$$

$$\mathcal{H}(m)\doteq\begin{cases}\dfrac{\sum\limits_{Rol\in\mathbb{V}(\circledcirc\succ)}\mathcal{H}_4(m\mid Rol)}{\sum\limits_{Rol\in\mathbb{V}(\circledcirc\succ)}\mathbf{1}(ref(Rol\succ*\xrightarrow{\rho}*\succ m)>0)}, & \ell(m)=4\text{ and }ref(*\xrightarrow{\rho}*\succ m)>0\\ \mathcal{H}_3(m\mid\circledcirc), & \ell(m)=3\text{ and }ref(*\xrightarrow{\pi}m)>0\\ 0, & \text{otherwise}\end{cases} \tag{3}$$

$$\mathcal{H}_4(m\mid Rol)\doteq\frac{-\sum\limits_{X\subseteq\boldsymbol{\Omega}(m|Rol)}\left[p_4(X\mid m,Rol)\cdot\log\left(\sum\limits_{Y\subseteq\boldsymbol{\Omega}(m|Rol)}p_4(Y\mid m,Rol)\cdot S_4(X,Y\mid m,Rol)\right)\right]}{log(ref(Rol\succ*\xrightarrow{\rho}*\succ m))} \tag{4}$$

$$\mathcal{H}_3(m\mid\circledcirc)\doteq\frac{-\sum\limits_{X\subseteq\boldsymbol{\Omega}(m|\circledcirc)}\left[p_3(X\mid m,\circledcirc)\cdot\log\left(\sum\limits_{Y\subseteq\boldsymbol{\Omega}(m|\circledcirc)}p_3(Y\mid m,\circledcirc)\cdot S_3(X,Y\mid m)\right)\right]}{\log(ref(*\xrightarrow{\pi}m))} \tag{5}$$

$$\boldsymbol{\Omega}(m\mid Rol)\doteq\{rel\mid(rel\twoheadrightarrow rol\xrightarrow{\rho}*\succ m)\in\mathbb{G},rol\in\mathbb{V}(Rol\succ)\} \tag{6}$$

$$\boldsymbol{\Omega}(m\mid\circledcirc)\doteq\{Rel\mid(Rel\twoheadrightarrow Rol\xrightarrow{\pi}m)\in\mathbb{G},Rol\in\mathbb{V}(\circledcirc\succ)\} \tag{7}$$

$$p_4(X\mid m,Rol)\doteq\frac{\left|\left\{\mathcal{G}_i\middle|\begin{array}{c}i\in[1,N],(Rol_i\succ*\xrightarrow{\rho}*\succ m_i)\in\mathcal{G}_i,\\(\forall rel\in X:(rel_i\twoheadrightarrow*\xrightarrow{\rho}*\succ m_i)\in\mathcal{G}_i),\\(\forall rel\in\boldsymbol{\Omega}(m\mid Rol)\setminus X:(rel_i\twoheadrightarrow*\xrightarrow{\rho}*\succ m_i)\notin\mathcal{G}_i)\end{array}\right\}\right|}{ref(Rol\succ*\xrightarrow{\rho}*\succ m)} \tag{8}$$

$$S_4(X,Y\mid m,Rol)\doteq\frac{\sum\limits_{(x,y)\in\max M(X,Y|m,Rol)}s_4(x,y\mid m,Rol)}{|X|+|Y|-|\max M(X,Y\mid m,Rol)|} \tag{9}$$

$$\max M(X,Y\mid m,Rol)\doteq\mathop{\arg\max}_{M\in\mathbb{MATCH}(X,Y)}\sum_{(x,y)\in M}s_4(x,y\mid m,Rol) \tag{10}$$

$$\mathbb{MATCH}(X,Y\mid m,Rol)\doteq\left\{M\middle|\begin{array}{l}M\subseteq X\times Y,\\\forall(x,y),(u,v)\in M:(x,y)\neq(u,v)\Rightarrow x\neq u\wedge y\neq v\end{array}\right\} \tag{11}$$

$$s_4(x,y\mid m,Rol)\doteq\frac{\sum\limits_{Rol'\in Rols(x,y|m,Rol)}sim(v_x^{Rol'},v_y^{Rol'})}{\left|Rols(x,y\mid m,Rol)\right|} \tag{12}$$

$$Rols(x,y\mid m,Rol)\doteq\left\{Rol'\middle|\begin{array}{c}Rol'\in\mathbb{V}(\circledcirc\succ),Rol'\neq Rol,\\(x\twoheadrightarrow rol'_x)\in\mathbb{G},rol'_x\in\mathbb{V}(Rol'\succ),\\(y\twoheadrightarrow rol'_y)\in\mathbb{G},rol'_y\in\mathbb{V}(Rol'\succ)\end{array}\right\} \tag{13}$$

$$v_x^{Rol'}\doteq\{v\mid(x\twoheadrightarrow rol'\xrightarrow{\rho}*\succ v)\in\mathbb{G},rol'\in\mathbb{V}(Rol'\succ)\} \tag{14}$$

$$sim(v_1,v_2)\doteq\begin{cases}\max\begin{pmatrix}sim(\alpha(v_1),\alpha(v_2)),\\sim(\alpha(v_2),\alpha(v_1))\end{pmatrix}, & \exists u\in\mathbb{V}(\circledast\succ):u\sqsupseteq(v_1\cup v_2)\\ \text{Jaccard}(v_1,v_2), & \text{otherwise}\end{cases} \tag{15}$$

$$p_3(X\mid m,\circledcirc)\doteq\frac{\left|\left\{\mathcal{G}_i\left|\begin{array}{c}i\in[1,N],(\circledcirc\succ *\xrightarrow{\pi}m_i))\in\mathcal{G}_i,\\(\forall Rel\in X,(Rel_i\twoheadrightarrow *\xrightarrow{\pi}m_i)\in\mathcal{G}_i),\\(\forall Rel\in\boldsymbol{\Omega}(m\mid\circledcirc)\setminus X,(Rel_i\twoheadrightarrow *\xrightarrow{\pi}m_i)\notin\mathcal{G}_i)\end{array}\right.\right\}\right|}{ref(Rol\succ *\xrightarrow{\rho}*\succ m)} \tag{16}$$

$$S_3(X,Y\mid m)\doteq\text{Jaccard}(X,Y) \tag{17}$$

群体知识图谱的熵$\boldsymbol{\mathcal{H}}(\mathcal{G})$等于群体知识图谱中所有节点 $m\in\mathbb{V}$的正则化熵$\boldsymbol{\mathcal{H}}(m)\cdot(1+\theta(m))$之和(公式(1)). 其中, $\boldsymbol{\mathcal{H}}(m)$为节点$m$的熵; $(1+\theta(m))$为对应的正则化系数. $\theta(m)$的定义见公式(2). 其中, $\Delta$为一个常量; $|m|_{\perp}$表示$m$的$N$个分量中, $N$不等于$\perp$的分量的数量. $\boldsymbol{\mathcal{H}}(m)$的定义见公式(3), 其度量了节点$m$以不同角色参与到关系中产生的熵的均值. 对于实例层节点$m(\ell(m)=4)$而言, $\boldsymbol{\mathcal{H}}_4(m|Rol)$表示节点$m$以$Rol$的实例的承担者参与到关系中产生的熵;$\boldsymbol{\mathcal{H}}(m)$表示$m$在不同$Rol$下产生的熵的平均值. 对于模型层节点$m(\ell(m)=3)$而言, $\boldsymbol{\mathcal{H}}(m)$表示$m$以⊙的实例的承担者参与到不同关系中产生的熵, 也即$\boldsymbol{\mathcal{H}}_3(m|\circledcirc)$. 对于元模型层和元元模型层, 目前不支持对其内容进行定制, 因此其中节点的熵均为0.

对于实例层和模型层节点在特定角色下产生的熵$\boldsymbol{\mathcal{H}}_4(m|Rol)$和$\boldsymbol{\mathcal{H}}_4(m|\circledcirc)$, 我们采用一种针对离散型随机变量且考虑随机变量不同取值之间相似度的广义熵进行计算, 并将其值归一化到区间[0,1]上. 对于任意离散型随机变量$X$, 其传统信息熵$\boldsymbol{\mathcal{H}}(X)$定义为$-\Sigma_{x\in X}p_X(x)\cdot\log(p_X(x))$. 给定关于$X$的所有取值之间的一个相似度计算函数 $S:X\times X\to[0,1]$, 则可以定义另外一种形式的熵$\boldsymbol{\mathcal{H}}(X,S)=-\Sigma_{x\in X}p_X(x)\cdot\log(\Sigma_{y\in X}p_X(y)\cdot S(x,y)))$称为广义熵[38]. 广义熵$\boldsymbol{\mathcal{H}}(X,S)$具有如下3个基本性质. (1) 可退化性. 如果随机变量$X$所有两个不同的取值$x$, $y$之间的相似度$S(x,y)=0$, 则广义熵退化为传统熵, 即$\boldsymbol{\mathcal{H}}(X,S)=\boldsymbol{\mathcal{H}}(X)$. (2) 可合并性. 如果随机变量$X$的两个不同的取值$x$, $y$之间的相似度 $S(x,y)=1$, 则可以将这两个值合并为一个值(新值的概率为$x$, $y$的概率之和), 且广义熵的值不变. (3) 相似度越大, 广义熵越小. 给定离散型随机变量$X$以及两个相似度函数 $S_0$, $S_1$, 对于$X$的任意两个取值$x$, $y$, 如果满足 $S_0(x,y)\leqslant S_1(x,y)$, 则一定有$\boldsymbol{\mathcal{H}}(X,S_0)\geqslant\boldsymbol{\mathcal{H}}(X,S_1)$. 为了方便理解, 可以将随机变量的不同取值理解为不同的观点, 不同取值之间的相似度即表示不同观点之间的相似度, 则随机变量的广义熵刻画了在特定群体中由于不同个体具有不同观点而产生的不一致性.

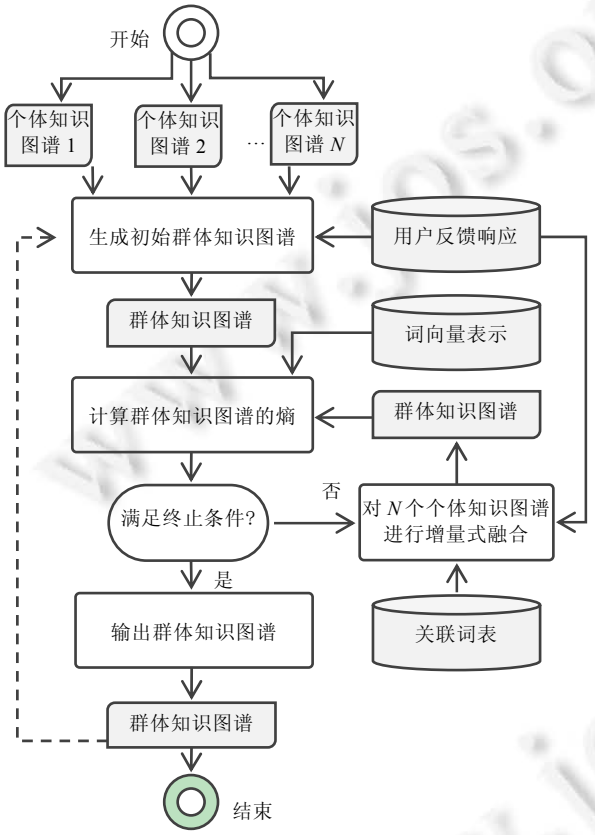
对于实例层节点 $m$ 而言, 其在 $Rol$ 下产生的熵$\boldsymbol{\mathcal{H}}_4(m|Rol)$对应的离散型随机变量的概率分布函数 $p_4(X|m,Rol)$的定义见公式(8), 其物理含义为观点$X$在所有包含"$m$节点通过$Rol$实例连接至特定关系"这一结构的个体知识图谱中的出现比例. 其中, $X$表示群体对$m$以$Rol$的实例的承担者参与到不同关系中的一种观点; 分母表示包含"$m$节点连接至$Rol$实例"结构的个体知识图谱数量; 分子表示在这些个体知识图谱中观点$X$的出现次数. 两种不同观点$X$, $Y$之间相似度的计算方式见公式(9). 其中, 分子表示在两个观点中元素之间的最大相似度匹中, 所有匹配的关系对$(x,y)$相似度之和; 分母表示两个观点中包含的元素之和减去最大相似度匹中的元素的数量. 该公式是两个集合之间 Jaccard 系数的一种泛化形式. 一对关系$x$, $y$的相似度的计算方式见公式(12), 其物理含义为这两个关系中除去角色$Rol$之外其他各个角色的承担者$v_x^{Rol'},v_y^{Rol'}$相似度的平均值. 当这两个承担者集合中包含的元素均为符号概念实例的实例时, 每个元素存在对应的字面量, 且每个字面量包含的单词具有相应的嵌入式向量表示; 此时, 我们计算这两个集合元素对应的字面量之间的相似度以及对应的嵌入式向量之间的相似度, 并取这两个相似度的较大值作为两个承担者集合之间的相似度. 否则, 如果两个承担者集合中包含的元素不是符号概念实例的实例, 则采用 Jaccard 系数作为两个承担者集合之间的相似度.

对于模例层节点 $m$ 而言, 其在⊙下产生的熵$\boldsymbol{\mathcal{H}}_3(m|\circledcirc)$对应的离散型随机变量的概率分布函数 $p_3(X|m,\circledcirc)$的定义见公式(16), 其物理含义为观点$X$在所有包含"$m$节点通过⊙实例连接至特定关系"这一结构的个体知识图谱中的出现比例. 其中, $X$表示群体对$m$以⊙实例的承担者参与到不同关系中的一种观点; 分母表示包含"$m$节点连接至⊙实例"结构的个体知识图谱数量; 分子表示在这些个体知识图谱中观点$X$的出现次数. 公式(17)给出了两种观点之间相似度的计算方式. 由于模型层中节点并不具有字面量, 我们直接采用两个集合的Jaccard系数作为两种观点之间的相似度.

2.2.3 个体知识图谱的融合算法

给定个体知识图谱集合$\boldsymbol{G}=\{\mathcal{G}_i=(\mathcal{V}_i,\boldsymbol{\mathcal{E}}_i,\mathcal{T}_i)|i\in[1,N]\}$，其融合问题是一个组合优化问题. 其优化目标是在融合方案空间$\mathbb{M}(\boldsymbol{G})$中寻找到一个熵值尽可能小的融合方案，进而形成相应的群体知识图谱. 另外，需要注意一点：在群体协同式知识图谱建模过程中，每一参与者的个体知识图谱都有可能发生持续的演化；相应地，个体知识图谱融合问题自身也会不断地演化.

图 3(a)给出了对个体知识图谱进行融合的基本流程，包含 4 个主要活动：初始融合活动，在上一时刻个体知识图谱融合结果的基础上，形成初始群体知识图谱；熵值计算活动，根据第 2.2.2 节的公式计算形成当前知识图谱的熵；终止条件判定活动，根据迭代轮数以及熵值变化情况，判断是否需要继续进行下一轮融合；增量融合活动，根据当前群体知识图谱中每一融合节点的熵值，筛选出可能存在融合错误的融合节点，并对其中包含的所有个体在知识图谱节点进行再融合. 其中，初始融合和增量融合两个活动会根据参与者在个体知识图谱构造过程中对反馈信息的响应(见第 2.3 节)进行相应的融合决策；熵值计算活动会根据词语的向量嵌入表示和词的标记距离计算不同词语之间的相似度；增量融合活动还会利用通过词向量建立的关联词表，缩小节点对齐关系的搜索范围，提高再融合的效率.




(a) 融合流程

输入: $\boldsymbol{G}=\{\mathcal{G}_i=(\mathcal{V}_i,\boldsymbol{\mathcal{E}}_i,\mathcal{T}_i)|i\in[1,N]\}$，个体知识图谱集合；$\mathcal{G}'$，前一时刻的群体知识图谱；$R$，反馈响应信息；$maxRound$=5，$\varepsilon$=0，$per$=0.05，基础参数配置.

输出: $\mathcal{G}$，融合后的群体知识图谱.

round←0; $\mathcal{G}_{pre}\leftarrow\mathcal{G}_{cur}\leftarrow initial(\boldsymbol{G},\mathcal{G}',R)$;
**repeat**:
  $round\leftarrow round+1$;
  $\mathbb{V}_{rm}\leftarrow$top($\mathcal{G}_{cur}$,$per$);
  **for** $i\in[1,N]$ **do** $\boldsymbol{V}_i\leftarrow\{\}$;
  **For** $\mathcal{V}\in\mathcal{G}_{cur}$ **do**
  **If** $\mathcal{V}\in\mathbb{V}_{rm}$ **or** $|\{v\in\mathcal{V}\,|v\neq\perp\}|=1$ **then**
  **For** $v_i\in\mathcal{V}$ **do if** $v_i\neq\perp$ then $\boldsymbol{V}_t\leftarrow\boldsymbol{V}_t\cup\{v_i\}\}$
  **end if**
  **end for**
  $\boldsymbol{MV}\leftarrow\{\}$;
  **For** $v\in\boldsymbol{V}_1$ **do** $\boldsymbol{MV}\leftarrow\boldsymbol{MV}\cup\{\{v\}\}$;
  **for** $i\in[2,N]$ **do**
    $W_i\leftarrow\{w_{xy}=+\infty|mv_x\in\boldsymbol{MV},\ v_y\in\boldsymbol{V}_i\}$;
    **for** $mv_x\in\boldsymbol{MV}$ **and** $v_y\in V_i$ **do**
      **if** related($mv_x$,$v_y$) **then**
        $w_{xy}\leftarrow\mathcal{H}(mv_x\cup\{v_y\},\mathcal{G}_{cur})+R.weight(mv_x,v_y)$;
      **end if**
  **end for**
  $\mathbf{MV}\leftarrow MinMatching(\boldsymbol{MV},\boldsymbol{V}_i,W_i)$
  **end for**
  $\mathcal{G}_{pre}\leftarrow\mathcal{G}_{cur}$, update($\mathcal{G}_{cur}$,$\boldsymbol{MV}$);
**until** $round\geqslant maxRound$ **or** $\mathcal{H}(\mathcal{G}_{cur})/\mathcal{H}(\mathcal{G}_{pre})\geqslant 1-\varepsilon$
$\mathcal{G}=(\mathcal{G}_{cur})<\mathcal{H}(\mathcal{G}_{pre})?\ \mathcal{H}(\mathcal{G}_{cur}):\mathcal{H}(\mathcal{G}_{pre})$

(b) 融合算法

图 3 个体知识图谱的自动融合

图 3(b)给出了个体知识图谱融合算法的更多细节信息. 该算法的输入包括当前个体知识图谱集合$\boldsymbol{G}$、前一时刻的群体知识图谱$\mathcal{G}'$、反馈响应信息 $R$、以及 3 个固定参数(最大迭代轮数 $maxRound$、熵值降低敏感度$\varepsilon$、以及再融合比例 $per$). 该算法的输出是对$\boldsymbol{G}$进行以熵最小化为目标的融合后得到的一个群体知识图谱$\mathcal{G}$. 首

先, 基于当前个体知识图谱集合$\boldsymbol{G}$、上一时刻的群体知识图谱$\mathcal{G}'$, 以及反馈响应情况 $R$, 得到初始群体知识图谱$\mathcal{G}_{\text{cur}}$ (第 1 行); 然后, 对$\mathcal{G}_{\text{cur}}$进行迭代式更新. 在每一轮迭代中, 首先筛选出熵值排名前 $per$ 比例的融合节点(第 4 行)以及未形成对齐关系的融合节点(第 7 行), 作为需要进行再融合的节点. 对于每个需要再融合的节点, 拆分得到其包含的个体知识图谱节点$v_i$, 并根据$v_i$ 的来源不同, 分别将其放入所在个体知识图谱对应的节点集合$\boldsymbol{V}_i$中(第 8 行). 然后, 依次对节点集合$\boldsymbol{V}_1,\boldsymbol{V}_2,...,\boldsymbol{V}_N$进行融合形成新的融合节点集合$\boldsymbol{MV}$. 其中, 第 1 次融合后, $\boldsymbol{MV}$中的元素与$\boldsymbol{V}_1$中的节点之间存在一一对应关系: $\boldsymbol{MV}$中的每一个元素包含且仅包含$\boldsymbol{V}_1$中的对应节点(第 12 行). 在第 $i$ 次融合时, 首先建立$\boldsymbol{MV}$中融合节点 $mv_x$ 和$\boldsymbol{V}_i$中待融合的节点$v_y$之间的权重 $w_{xy}$: 权重 $w_{xy}$ 依赖于 $mv_x$ 和$v_y$ 融合后的熵值$\mathcal{H}$和反馈响应信息 $R$(第 17 行). 同时, 为了提升在大规模知识图谱融合中增量融合的效率, 基于关联词表对可能存在的融合关系进行筛选, 从而避免进行不必要的融合(第 16 行). 最后, 通过权值最小匹配得到$\boldsymbol{MV}$与$\boldsymbol{V}_i$的最优融合方案, 并形成新的融合$\boldsymbol{MV}$(第 20 行). 在每一轮迭代的最后, 更新当前群体知识图谱$\mathcal{G}_{\text{cur}}$, 并记录更新前的群体知识图谱$\mathcal{G}_{\text{pre}}$. 然后, 根据终止条件满足情况决定是否进行下一轮更新. 我们将终止条件设定为: 超过预先设定的最大迭代轮数, 或迭代后形成的群体知识图谱的熵值不再发生显著下降(第 23 行). 若终止条件满足, 则输出本次融合后形成的群体知识图谱$\mathcal{G}$(第 24 行).

### 2.3 主动反馈

在主动反馈活动中, 支撑环境在当前群体知识图谱的基础上, 根据每一参与者当前的个体知识图谱, 向该参与者反馈特定的知识图谱片段信息, 以提高其构造知识图谱的效率. 每一参与者自主决定是否接受、拒绝或忽略支撑环境提供的反馈信息. 被接受的反馈信息会进入参与者的个体知识图谱. 参与者对反馈信息的响应也会被记录下来, 用于评估个体的知识偏好以及群体对特定信息的接受程度.

我们设计了两种类型的信息反馈方式:情境无关的反馈、情境相关的反馈. 在情境无关的反馈中, 支撑环境不考虑参与者的当前个体知识图谱和当前操作, 直接向其反馈一定数量的高质量知识图谱节点. 为了实现情境无关的反馈, 支撑环境通过下面的公式确定群体知识图谱中每一个融合节点的反馈强度(反馈强度越高的节点, 其被反馈的概率也越高).

$$ri(m)\doteq\frac{1}{1+\mathcal{H}(m)}\cdot ref(m) \tag{18}$$

其中, $\mathcal{H}(m)$表示节点 $m$ 的熵值, $1/(1+\mathcal{H}(m))\in(0,1]$表示节点的一致程度. $ref(m)=|\{m_i|m_i\in m,m_i\neq\perp\}|$表示节点在个体知识图谱中出现的次数. 可知, 节点的反馈强度正比于节点本身的一致性和出现次数.

在情境相关的反馈中, 支撑环境根据参与者当前关注的实例层知识图谱实体节点(所谓实例层知识图谱实体节点, 即实体概念$\odot$实例的实例. 类似地, 所谓实例层知识图谱关系节点, 即关系概念$\ominus$实例的实例), 为其反馈可能相关的实例层关系节点和实体节点. 为了实现情境相关的反馈, 支撑环境通过下面的两个公式分别确定与当前实体节点相关的关系/实体节点对其的反馈强度.

$$ri(rel\twoheadrightarrow rol\,|\,ent)\doteq\frac{ref(rol\xrightarrow{\rho}*\succ ent)^2}{\sum\limits_{rol_i\in\{rol'|v\succ rol,v\succ rol'\}}ref(rol_i\xrightarrow{\rho}*\succ ent)} \tag{19}$$

$$ri(ent'\,|\,ent)\doteq ri(ent')\cdot\prod_{\substack{rel\in\mathbb{V}(\ominus\succ\succ),\,rol,rol'\in(rel\rightarrow)\\ ref(rol'\xrightarrow{\rho}*\succ ent')>0}}ri(rel\twoheadrightarrow rol\,|\,ent) \tag{20}$$

给定当前关注的实体节点 $ent$, 若其以角色节点 $rol$ 连接至关系节点 $rel$, 则 $rel$ 对 $ent$ 的反馈强度通过公式(19)确定. 其中, $ref(rol\xrightarrow{\rho}*\succ ent)$表示“$ent$ 承担 $rol$ 角色”这一结构在个体知识图谱中出现的次数. 可知, 给定实体节点和角色节点, 相关关系节点对其的反馈强度正比于关系结构的出现次数和相对比例. 给定当前关注的实体节点 $ent$, 则任意其他实体节点 $ent'$对 $ent$ 的反馈强度等于 $ent'$的情境无关反馈强度乘以两者共同连接的所有关系节点对 $ent$ 的反馈强度之和(见公式(20)). 可知, 当相关实体节点的情境无关反馈强度较高时, 实体节点之间关系越密切, 则相关实体节点的反馈强度越高.

## 3 实验与评估

本节通过 3 种类型的实验对上述基于群体智能的知识图谱构建方法的有效性和可行性进行验证. 其中, (1) 仿真图融合实验通过对仿真生成的仅包含结构信息的带类型有向图进行融合, 验证本文提出的知识图谱融合算法对图结构信息的有效利用能力; (2) 知识图谱融合实验通过对真实存在的大规模知识图谱进行自动融合, 验证本文提出的知识图谱融合算法的有效性; (3) 知识图谱构造实验通过多人在受控条件下的在线知识图谱构造活动, 验证基于“探索-融合-反馈”回路的协同式知识图谱构建的可行性.

### 3.1 仿真图融合实验

#### 3.1.1 实验设计与数据

本文提出的知识图谱融合算法综合利用了图中节点自身的语义信息以及图的结构信息. 为考察该算法对图的结构信息的有效利用能力, 我们随机生成了一组具有不同差异度的带类型有向图对(其中, 图中的节点不存在任何语义信息), 分析该算法在具有不同差异度的图对上进行融合时, 所形成的融合方案的准确率的变化情况.

在实验数据生成上, 我们将图中节点的类型数设置为 6, 边的类型设置为 3, 在节点数量上选取了 1 000、5 000、10 000 这 3 种取值, 在节点的平均度上选取了 5、10、15、20 这 4 种取值;基于上述配置, 随机生成了 12 个母图. 然后, 设计了 12 种图变异方案, 即采用 5%、10%、15%、20%、25%、30%这 6 种比例值, 从一个图中随机删除相应比例的节点或边. 母图和图变异方案进行组合, 形成 144 个实验组. 每一个实验组中包含 6 个图: 1 个母图以及 5 个基于当前母图和图变异方案形成的变异图.

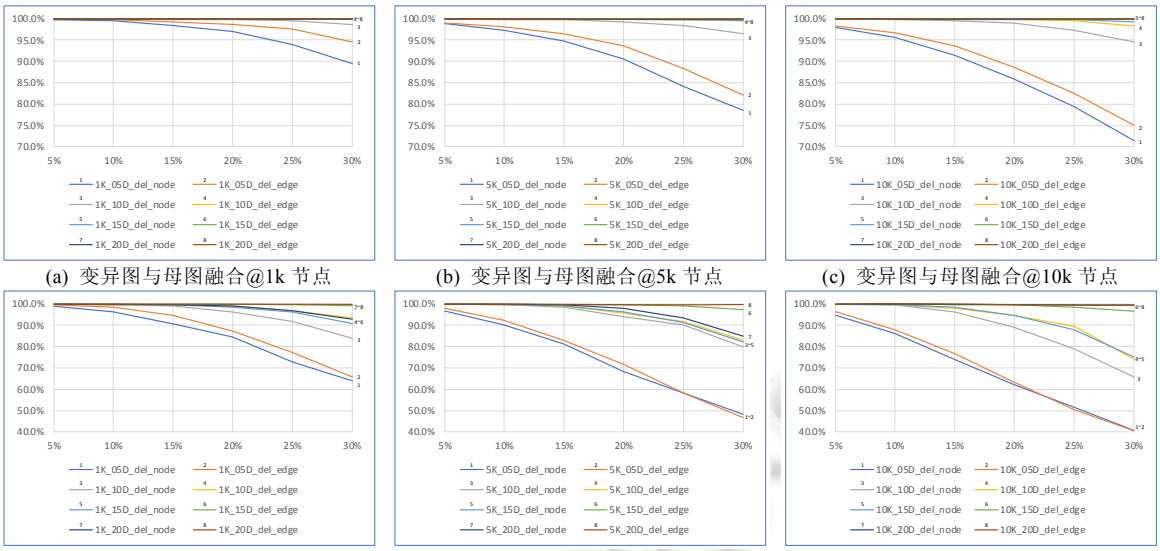
在实验中, 我们将算法的最大迭代轮数 *maxRound* 设置为 5、熵值降低敏感度$\varepsilon$设置为 0、再融合比例 *per* 设置为 10%. 将节点的类型以及节点关联边和点的类型计数作为该节点的局部结构向量, 基于节点局部结构向量生成初始融合方案, 并利用最小熵原则进行再融合得到最终融合方案.

#### 3.1.2 实验结果与分析

图 4 给出了母图节点规模为 1 000、5 000、10 000, 节点平均度规模为 5、10、15、20 时的融合效果. 其中, 横轴表示变异方案中节点/边的删除比例, 纵轴表示在对应变异方案下不同实验组的平均融合准确率. 在图 4(a)–图 4(c)中, 采用每个实验组中的母图与其变异图进行融合, 其准确率为 5 组母图与变异图融合准确率的平均值; 图 4(d)–图 4(f)中, 采用每个实验组中变异图之间相互融合, 其准确率为 10 组变异图相互融合准确率的平均值. 显然, 在同等变异方案下, 图 4(d)–图 4(f)中两图的差异更大.

从图 4 中可以观察到, 当节点平均度较高时, 本文算法具有较好的融合准确率. 在母图与变异图的融合实验中(图 4(a)–图 4(c))中节点度为 10、15 和 20 时融合准确率均接近于 100%. 但随着图规模的增大和变异率的增大, 融合准确率逐渐降低. 在节点平均度为 5 且删除 30%的点这种情况下, 准确率最低: 当节点规模为 1 000、5 000、10 000 时, 准确率分别为 89.5%, 78.5%和 71.4%. 在变异图之间的融合实验中(图 4(d)–图 4(f)), 由于两个变异图的结构差异较大, 融合准确率均小于与母图与变异图的融合准确率. 其中, 当节点平均度较高(大于等于 15)且变异率较低(小于等于 20%)时, 准确率仍接近 100%; 随着节点规模增多, 准确率略有下降; 随着节点平均度的降低和变异率的提升, 准确率明显下降. 在节点平均度为 5 且删除 30%的点这种情况下, 准确率最低: 当节点规模为 1 000、5 000、10 000 时, 准确率分别为 64.0%, 48.2%和 40.6%.

实验结果说明, 本文算法能够有效利用图的结构信息进行融合: 图的结构信息越丰富(即节点平均度越高), 算法的准确率越高. 但如果图的结构信息不够丰富, 由于节点/边的类型信息有限, 限制了节点的表达能力, 对本文算法的准确率造成了较大的影响. 在真实知识图谱中, 节点通常会具有名称、属性等语义描述信息. 我们认为, 在图的结构信息的基础上, 进一步考虑节点自身的语义信息, 会进一步提高本文算法的融合准确率.




(a) 变异图与母图融合@1k 节点 (b) 变异图与母图融合@5k 节点 (c) 变异图与母图融合@10k 节点

(d) 变异图与变异图融合@1k 节点 (e) 变异图与变异图融合@5k 节点 (f) 变异图与变异图融合@10k 节点

图 4 仿真图融合的准确率

### 3.2 知识图谱融合实验

#### 3.2.1 实验设计与数据

我们选取了两组真实的知识图谱数据[39]进行知识图谱融合实验. 该数据来源于通过对 DBpedia、Wikidata 和 YAGO3 这 3 个真实大规模知识图谱采样形成的两对知识图谱: DBP-WD 和 DBP-YG. 其中, 每一个知识图谱包含 10 万对实体、上百万条属性和关系. 在这两对知识图谱上, 我们将本文提出的知识图谱融合算法与传统的知识图谱融合算法(LogMap)以及基于嵌入(embedding)的知识图谱融合算法(MTransE、BootEA、GCN-Align、TransD、JAPE 以及 MultiKE)进行了对比分析.

基于嵌入的融合算法, 一般采用实体的名称、属性和关系中的一类或多类信息进行训练并构建实体向量; 然后通过实体向量之间的相似性得到一组最相似的实体节点, 作为对齐关系的候选节点. 这些算法通常采用 Hit@$N$ 指标(即前 $N$ 个候选节点中包含正确对齐节点的概率)作为对齐质量的评判标准. 而传统算法和本文算法直接生成两个图之间的一个对齐结果(即对每个节点给出与其对齐的唯一一个节点). 为了保证公平性, 我们采用 Hit@1 指标对这 3 类算法的效果进行对比.

实验采用的硬件平台为深腾 X8800 服务器(2×Intel Xeon E5-2697A V4, 256 GB CPU). 实验中将算法的最大迭代轮数 *maxRound* 设置为 5、熵值降低敏感度$\varepsilon$设置为 0、再融合比例 per 分别设置为 5%和 10%. 利用知识图谱中实体名称的相似性生成初始融合方案. 并利用最小熵原则进行再融合得到最终融合方案.

#### 3.2.2 实验结果与分析

表 1 给出了 3 类融合算法在 DBP-WD 和 DBP-YG 这两对知识图谱上的融合效果.

表 1 不同知识图谱融合方法的比较 (%)

| | 传统方法 | 基于 Embedding 方法 | | | | | | 本文方法 | |
|---|---|---|---|---|---|---|---|---|---|
| | LogMap | MTransE | BootEA | GCN-Align | TransD | JAPE | MultiKE | EIR&5% | EIR&10% |
| DBP-WD@Hit1 | 80.68 | 28.12 | 74.79 | 47.70 | 36.20 | 31.84 | 91.45 | **93.20** | **93.69** |
| DBP-YG@Hit1 | 83.73 | 28.12 | 76.10 | 60.05 | 33.49 | 23.57 | 88.03 | **99.33** | **99.33** |

本文算法在迭代过程中需要对一定比例的节点进行再融合; 我们选择了两种再融合比例(5%、10%)分别对本文算法进行实际运行. 实验结果表明, 这两种再融合比例对于 Hit@1 指标的影响并不大; 但与其他算法相比, 在这两个实验数据上, 本文算法都取得了最高的 Hit@1 指标. 与之前效果最好的算法相比, 在 DBP-WD

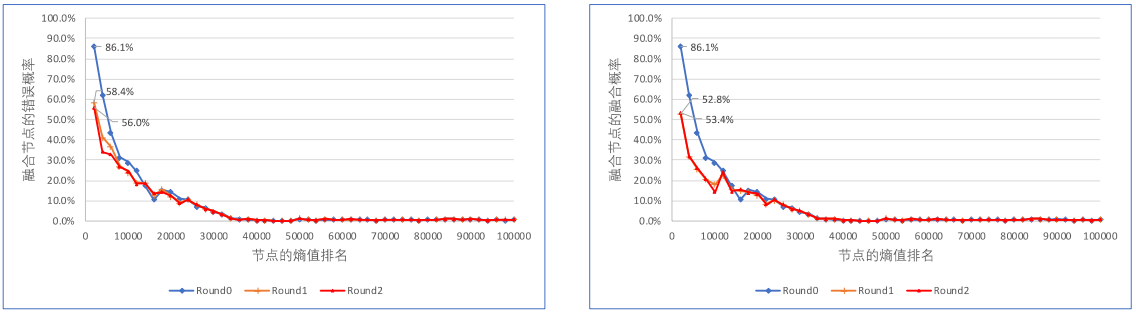
数据上, 本文算法的 Hit@1 指标提高了 1.75%–2.24%; 在 DBP-YG 数据上, 本文算法的 Hit@1 指标提高了 11.3%.

表 2 给出了在这两个实验数据上, 本文算法分别采用 5%和 10%再融合率条件时 Hit@1 指标和融合图熵值的变化过程. 可以看到, 在迭代式融合过程中, Hit@1 指标和融合图熵值在两个数据集上都得到了快速收敛. 其中, 在 DBP-WD 数据上, 本文算法在经过 3 次迭代后进入收敛状态(融合图熵值和 Hit@1 指标都不再发生变化); 且在迭代过程中, 熵值持续下降, Hit@1 指标持续提升. 在 DBP-YG 数据上, 本文算法在经过一次迭代后即进入收敛状态. 实验结果表明, 以熵值最小化为优化目标, 可以形成一个具有较高质量的融合结果.

表 2 Hit@1 指标与融合图熵值在迭代中的变化

| | DBP-WD@5% | | DBP-WD@10% | | DBP-YG@5% | | DBP-YG@10% | |
|---|---|---|---|---|---|---|---|---|
| | Hit@1 (%) | Entropy | Hit@1 (%) | Entropy | Hit@1 (%) | Entropy | Hit@1 (%) | Entropy |
| Round0 | 91.47 | 9 435.10 | 91.47 | 9 435.10 | **99.33** | **518.10** | **99.33** | **518.10** |
| Round1 | 92.92 | 8 991.86 | 92.92 | 8 881.72 | 99.33 | 518.15 | 99.33 | 518.15 |
| Round2 | **93.20** | **8 939.41** | **93.69** | **8 876.72** | 99.33 | 518.15 | 99.33 | 518.15 |
| Round3 | 93.20 | 8 940.76 | 93.68 | 8 878.53 | 99.33 | 518.15 | 99.33 | 518.15 |
| Round4 | 93.19 | 8 940.24 | 93.67 | 8 878.07 | 99.33 | 518.15 | 99.33 | 518.15 |
| Round5 | 93.20 | 8 940.24 | 93.68 | 8 878.13 | 99.33 | 518.15 | 99.33 | 518.15 |

图 5 给出了在 DBP-WD 数据上, 每轮迭代中融合节点的错误率和融合节点熵值排名之间的关系. 其中, 在 5%再融合条件下, 熵值排名最高的前 2 000 个融合节点的错误率由 86.1%逐步下降为 58.4%和 56.0%; 在 10%再融合条件下, 熵值排名最高的前 2 000 个融合节点的错误率由 86.1%下降为 52.8%和 53.4%. 实验结果表明, 融合节点熵值排名与错误率存在较强的正相关性, 且通过选择熵值较高的融合节点进行再融合能够有效提升融合质量.




(a) 基于 DBP-WD 的融合, 再融合率 5% (b) 基于 DBP-WD 的融合, 再融合率 10%

图 5 融合节点的错误率与融合节点熵值排名之间的关系

### 3.3 知识图谱构造实验

#### 3.3.1 实验设计与数据

我们开发了一个支持“探索-融合-反馈”回路的群体协同知识图谱在线构造支撑环境(访问网址 http://www.tupu.fun). 图 6 给出了该支撑环境的基本架构与核心模块. 在服务器端, 群体知识图谱是核心制品. 融合模块不断地将每一参与者的构建操作(包括新建的个体知识图谱片段和对反馈信息的响应)融入群体知识图谱, 使得群体知识图谱持续演化. 在群体知识图谱的基础上, 反馈模块将其中的信息有针对性的反馈给不同的参与者, 提高其构造知识图谱的效率. 在客户端, 通过全局和局部两类视图对参与者的个体知识图谱以及相关的反馈信息进行展示, 并支持相应的知识图谱构造操作. 图 7 给出了反馈信息在两类视图中的展示方式. 在全局视图中, 通过高亮方式对存在反馈的实体进行提示; 在局部视图下, 对当前实体的反馈信息进行更详细展示(灰色区域部分). 参与者可以对反馈信息进行查看并引用. 引用后, 对应的知识图谱片段会被加入到当前个体知识图谱中.

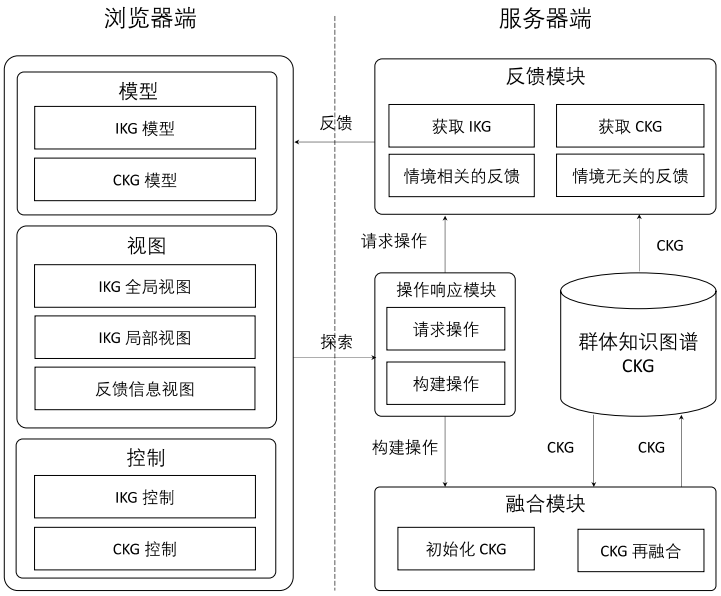
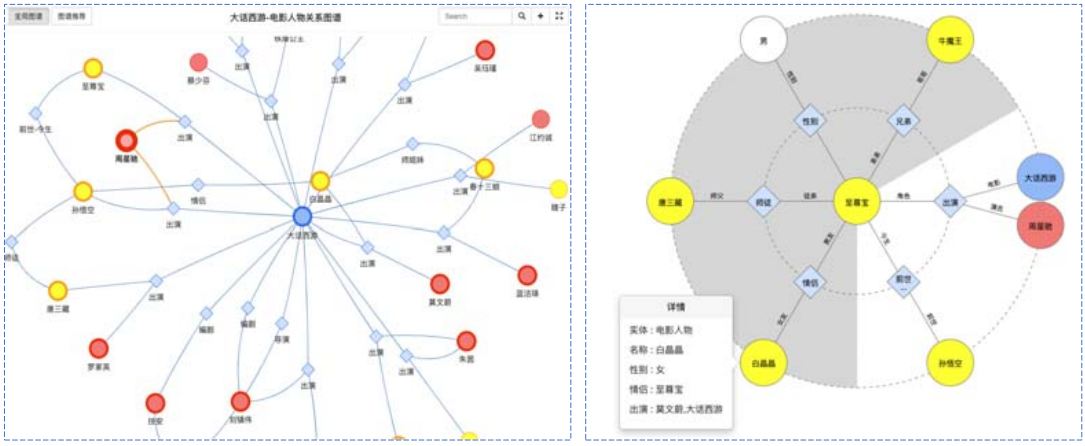



图 6 基于“探索-融合-反馈”回路的知识图谱构造支撑环境的体系结构



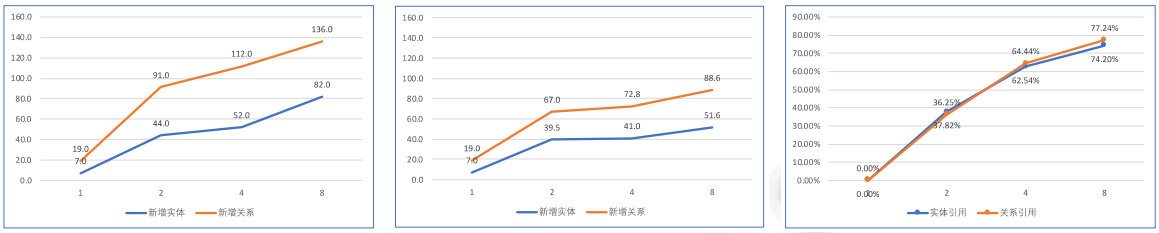
(a) 全局视图 (b) 局部视图

图 7 基于“探索-融合-反馈”回路的知识图谱构造支撑环境截图

我们通过针对真实知识图谱的构造实验, 验证基于“探索-融合-反馈”回路的多人协同构建知识图谱的可行性. 我们将开放知识图谱 openkg 上新冠事件相关数据(数据来源 http://openkg.cn/dataset/covid-19)中的 140 项政府政策作为原始知识图谱; 然后让参与者通过阅读这些政策的文字内容, 识别并添加政策之间的两类关系. 第 1 类关系为政策依据, 即当前政策是依据哪些其他政策而提出的. 第 2 类关系为前期政策, 即当前政策的前一版本的政策. 在关系的添加过程中, 如果发现涉及的政策不在当前知识图谱内, 则要求参与者将该政策添加到知识图谱中. 我们征集到 15 名实验参与者, 将这些参与者分为 4 组, 每一组分别包含 1、2、4、8 位参与者. 每一组参与者作为一个独立群体进行上述知识图谱的构造任务, 时间限制为 2 小时. 在每一组的实验过程中, 在支撑环境的支持下, 每一个参与者构造的知识图谱片段会被实时融合在一起, 并通过自动反馈机制推荐给同组中的其他参与者; 除此之外, 参与者之间不存在其他形式的信息交流.

3.3.2 实验结果与分析

图8展示了不同人数规模下, 进行协同构建知识图谱的实验效果. 横坐标表示每组实验中参与人数, 分别为 1、2、4、8. 其中 1 人组实验由于不存在同组其他实验者无法获得相关反馈信息, 可以视为对比实验组. 对于每组实验中的个体知识图谱与群体知识图谱, 通过人工方式筛选出其中有效的实体和关系, 即仅对参与者创建的与原知识图谱中政府政策直接关联的关系及实体进行统计.




(a) 群体知识图谱的新建数量 (b) 个体知识图谱中的新建数量 (c) 协同知识图谱构建中的引用率

图 8 多人协同构建知识图谱的实验结果

图 8(a)显示了不同人员规模下, 群体知识图谱中新增实体和关系的数量. 其中, 新增实体数量分为 7、44、52 和 82 个; 新增关系数量分别为 19、91、112 和 136 个. 可以看到, 随着人员规模的增加, 群体知识图谱的规模不断增加. 图 8(b)显示了不同人员规模下, 实验组中每一个体知识图谱的平均新增实体和关系的数量. 其中, 新增实体数量分为 19、39.5、41.0 和 51.6 个; 新增关系数量分别为 19.0、67.0、72.8 和 88.6 个. 可以看到, 随着人员规模的增加, 通过“探索-融合-反馈”回路, 群体中每一个体的工作效率也在不断提高. 图 8(c)显示了不同人员规模下, 协同环境中每一个体通过引用反馈信息新建的实体和关系, 在其构建的所有实体和关系中的占比. 除单人实验外, 对实例的引用操作占比为 37.82%、62.54%、74.20%; 对关系的引用操作占比为 36.25%、64.44%、77.24%. 可以看到, 随着群体规模的增加, 通过“探索-融合-反馈”回路, 群体对个体的建模工作形成了更加有效的支持.

上述实验结果在一定程度上表明, 基于“探索-融合-反馈”回路的知识图谱构造环境能够有效支撑多人协同构建知识图谱, 提升群体构建知识图谱的规模和个体构建知识图谱的效率, 并具备较好的规模可扩展性. 需要指出的是, 上述知识图谱构造支撑环境目前还是一个原型系统, 其目的是对本文提出的知识图谱构造方法进行初步验证. 但我们认为, 随着该支撑环境易用性和性能的不断提升, 其将具有广泛的应用前景.

## 4 总结与展望

本文提出了一种基于互联网群体智能的知识图谱构造方法, 其目标是通过基于互联网的人类群体协同构造高质量的知识图谱. 该方法的核心是一个包含探索、融合、反馈 3 个活动的回路. 该回路将参与者构建的个体知识图实时融合形成群体知识图谱, 并将群体知识图谱中的知识片段有针对性地反馈给不同的参与者, 为群体的有效协同提供了一种基于环境(群体知识图谱)的间接交互机制. 我们开发了一个支持该方法的多人在线知识图谱构造环境, 设计并实施了 3 种不同类型的实验, 对所提方法和关键技术的可行性进行了初步验证. 后继研究工作计划从两个方面展开. 一方面, 我们计划招募更多的参与者在更多领域中进行更大规模的知识图谱构建实验, 进一步考察本文方法的群体规模可扩展性, 并对支撑环境进行持续改进. 另一方面, 我们计划将更细粒度的用户行为考虑到方法中, 实现更为精准和个性化的信息反馈, 进一步提升群体协同构造知识图谱的效率和效果.

长远而言, 我们希望以本文提出的知识图谱构造方法为基础, 促成一种基于群体智能的结构化知识建模生态系统. 在微观层次上, 该生态系统允许任何具有自主性的人类个体或智能算法在其中按照自己的意愿建立和维护面向特定问题领域的个体知识图谱, 并在“探索-融合-反馈”回路的支持下实现与其他参与者之间的有效协同. 在宏观层次上, 该生态系统将涌现形成一种大规模、多样性、结构化、持续演化的复杂信息制品, 其

中记录了人类文明发展过程中形成的各类知识和信息. 在比喻的意义上, 或许可以将这种信息制品称为“网络空间中的巴别塔(a tower of Babel in cyberspace)”.

**References**:


[1] Zhang W, Mei H. A constructive model for collective intelligence. National Science Review, 2020, 7(8): 1273−1277.

[2] Berners-Lee T, Hendler J, Lassila O. The semantic Web. Scientific American, 2001, 284(5): 34−43.

[3] Lenat DB. CYC: A large-scale investment in knowledge infrastructure. Communications of the ACM, 1995, 38(11): 33−38.

[4] Miller GA. WordNet: A lexical database for English. Communications of the ACM, 1995, 38(11): 39−41.

[5] Bollacker K, Evans C, Paritosh P, *et al*. Freebase: A collaboratively created graph database for structuring human knowledge. In: Proc. of the 2008 ACM SIGMOD Int'l Conf. on Management of Data. 2008. 1247−1250.

[6] Bizer C, Lehmann J, Kobilarov G, *et al*. DBpedia—A crystallization point for the Web of data. Journal of Web Semantics, 2009, 7(3): 154−165.

[7] Ayers P, Matthews C, Yates B. How Wikipedia Works: And How You Can Be a Part of It. San Francisco: No Starch Press, 2008.

[8] Paulheim H. Knowledge graph refinement: A survey of approaches, evaluation methods. Semantic Web, 2017, 8(3): 489−508.

[9] Aone C, Ramos-Santacruz M. REES: A large-scale relation, event extraction system. In: Proc. of the 6th Applied Natural Language Processing Conf. 2000. 76−83.

[10] Han X, Sun L, Zhao J. Collective entity linking in Web text: A graph-based method. In: Proc. of the 34th Annual ACM SIGIR Conf. 2011. 765−774.

[11] Lao N, Mitchell T, Cohen W. Random walk inference, learning in a large scale knowledge base. In: Proc. of the 2011 Conf. on Empirical Methods in Natural Language Processing. 2011. 529−539.

[12] Rau LF. Extracting company names from text. In: Artificial Intelligence Applications. 1991. 29−30.

[13] Bordes A, Usunier N, Garcia-Duran A, *et al*. Translating embeddings for modeling multi-relational data. In: Advances in Neural Information Processing Systems. 2013. 2787−2795.

[14] Ganea OE, Hofmann T. Deep joint entity disambiguation with local neural attention. In: Proc. of the 2017 Conf. on Empirical Methods in Natural Language Processing. 2017. 2619−2629.

[15] Jiang XT, Wang Q, Li P, *et al*. Relation extraction with multi-instance multi-label convolutional neural networks. In: Proc. of the COLING 2016, the 26th Int'l Conf. on Computational Linguistics: Technical Papers. 2016. 1471−1480.

[16] Xu M, Jiang H, Watcharawittayakul S. A local detection approach for named entity recognition, mention detection. In: Proc. of the 55th Annual Meeting of the Association for Computational Linguistics. 2017. 1237−1247.

[17] Etzioni O, Banko M, Cafarella MJ, *et al*. Open information extraction from the Web. Communications of the ACM, 2008, 51(12): 68−74.

[18] Mintz M, Bills S, Snow R, *et al*. Distant supervision for relation extraction without labeled data. In: Proc. of the Joint Conf. of the 47th Annual Meeting of the ACL and the 4th Int'l Joint Conf. on Natural Language Processing of the AFNLP. 2009. 1003−1011.

[19] Dong X, Gabrilovich E, Heitz G, *et al*. Knowledge vault: A Web-scale approach to probabilistic knowledge fusion. In: Proc. of the 20th ACM SIGKDD Int'l Conf. on Knowledge Discovery and Data Mining. 2014. 601−610.

[20] Vrandečić D, Krötzsch M. Wikidata: A free collaborative knowledgebase. Communications of the ACM, 2014, 57(10): 78−85.

[21] Chakrabarti K, Chaudhuri S, Cheng T, *et al*. A framework for robust discovery of entity synonyms. In: Proc. of the 18th ACM SIGKDD Int'l Conf. on Knowledge Discovery and Data Mining. 2012. 1384−1392.

[22] Liu A, Soderland S, Bragg J, *et al*. Effective crowd annotation for relation extraction. In: Proc. of the 2016 Conf. of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. 2016. 897−906.

[23] Piscopo A, Phethean C, Simperl E. What makes a good collaborative knowledge graph: Group composition, quality in Wikidata. In: Proc. of the Int'l Conf. on Social Informatics. 2017. 305−322.

[24] Collins AM, Quillian MR. Retrieval time from semantic memory. Journal of Verbal Learning, Verbal Behavior, 1969, 8(2): 240−247.

[25] Liao SH. Expert system methodologies, applications—a decade review from 1995 to 2004. Expert Systems with Applications, 2005, 28(1): 93−103.

[26] Brachman RJ, Schmolze JG. An overview of the KL-ONE knowledge representation system. In: Readings in Artificial Intelligence, Databases. 1989. 207−230.

[27] Wang Z, Zhang J, Feng J, *et al*. Knowledge graph embedding by translating on hyperplanes. In: Proc. of the 28th AAAI Conf. on Artificial Intelligence. 2014. 1112−1119.

[28] Lin Y, Liu Z, Sun M, *et al*. Learning entity, relation embeddings for knowledge graph completion. In: Proc. of the 29th AAAI Conf. on Artificial Intelligence. 2015. 2181−2187.

[29] Bonabeau E, Theraulaz G, Dorigo M. Swarm Intelligence: From Natural to Artificial Systems. Oxford University Press, 1999.

[30] Woolley A, Chabris C, Pentland A, *et al*. Evidence for a collective intelligence factor in the performance of human groups. Science, 2010, 330(6004): 686−688.

[31] Theraulaz G, Bonabeau E. A brief history of stigmergy. Artificial Life, 1999, 5(2): 97−116.

[32] Levy P. Collective Intelligence: Mankind's Emerging World in Cyberspace. Basic Books, 1999.

[33] Zhang W, Mei H. Software development based on Internet collective intelligence: Feasibility, state-of-the-practice, challenges. Science China Information Sciences, 2017, 47(12): 1601−1622 (in Chinese with English abstract).

[34] Rosenberg LB. Human Swarms, a real-time method for collective intelligence. In: Proc. of the 13th Artificial Life Conf. 2015. 658−659.

[35] Lee J, Kladwang W, Lee M, *et al*. RNA design rules from a massive open laboratory. Proc. of the National Academy of Sciences, 2014, 111(6): 2122−2127.

[36] Doyle MJ, Marsh L. Stigmergy 3.0: From ants to economies. Cognitive Systems Research, 2013, 21: 1−6.

[37] Lewis TD, Leslie M. Human stigmergy: Theoretical developments, new applications. Cognitive Systems Research, 2016, 38: 1−3.

[38] Wang SJ. The design and implementation of a stigmergy-based tool for conceptual modeling [MS. Thesis]. Beijing: Peking University, 2016 (in Chinese with English abstract).

[39] Zhang Q, Sun Z, Hu W, *et al*. Multi-view knowledge graph embedding for entity alignment. In: Proc. of the 28th Int'l Joint Conf. on Artificial Intelligence. 2019. 5429−5435.

**附中文参考文献**:

[33] 张伟, 梅宏. 基于互联网群体智能的软件开发: 可行性、现状与挑战. 中国科学: 信息科学, 2017, 47(12): 1601−1622.

[38] 王诗君. 基于环境激发效应的概念建模工具的设计与实现 [硕士学位论文]. 北京: 北京大学, 2016.

**蒋逸**(1989－), 男, 博士, 主要研究领域为群体智能, 知识图谱融合, 知识图谱建模.

**张馨月**(1995－), 女, 硕士, 主要研究领域为群体智能, 知识图谱建模.

**张伟**(1978－), 男, 博士, 副教授, CCF 专业会员, 主要研究领域为软件复用, 需求工程, 基于群体智能的软件开发.

**梅宏**(1963－), 男, 博士, 教授, 博士生导师, CCF 会士, 主要研究领域为软件工程, 系统软件.

**王佩**(1995－), 男, 硕士, 主要研究领域为群体智能, 知识图谱融合.