THU AI TR

人工智能之知识图谱

Research Report of Knowledge Graph

2019年 第2期

清华大学人工智能研究院

北京智源人工智能研究院

清华-工程院知识智能联合研究中心

2019年1月

目录

图表目录

摘要

知识图谱（Knowledge Graph）是人工智能重要分支知识工程在大数据环境中的成功应用，知识图谱与大数据和深度学习一起，成为推动互联网和人工智能发展的核心驱动力之一。基于此背景，本研究报告对知识图谱这一课题进行了简单梳理，包括以下内容：

知识图谱的概念与研究概况。对知识图谱的概念、分类进行阐述，并分四个阶段对知识工程的发展历程进行介绍。

知识图谱技术。从知识表示与建模、知识获取、知识融合、知识图谱查询推理及知识图谱应用五个子领域来划分，并分别介绍每个领域所应用到的技术。

知识图谱领域专家介绍。依据 AMiner 数据平台信息，对知识图谱领域的 5 个细分领域进行梳理，重点介绍每一细分领域研究学者的研究方向与代表性文章，旨在为学术界、产业界提供知识图谱技术及学者的分析依据，同时面向政府机关、高校、企业等对知识图谱技术感兴趣的机构介绍该领域基本概念、研究与应用方向。包括顶尖学者的全球分布、迁徙概况、学者机构分布、h-index 分析，并依据 AMiner 评价体系，在知识图谱发展过程中近十年的高引学者进行详细介绍。

知识图谱应用。从通用知识图谱应用和领域知识图谱应用两个方面来介绍。以电子商务、图书情报、企业商业、船业投资、生物医疗五个领域，从图谱构建与知识应用两个方面介绍领域知识图谱的技术构建应用与研究现状。

知识图谱趋势研究。对知识图谱的发展趋势特点进行分析。并基于 AMiner 数据平台，对近期知识图谱领域研究热点进行可视化分析，对未来知识图谱研究方向进行预测。

报告（电子版）实时更新，获取请前往：

https://www.aminer.cn/research_report/5c3d5a8709e961951592a49d?download=true&pathname=knowledgegraph.pdf。

1. 概念篇

1.1. 知识图谱概念和分类

知识图谱（Knowledge Graph）以结构化的形式描述客观世界中概念、实体及其之间的关系，将互联网的信息表达成更接近人类认知世界的形式，提供了一种更好地组织、管理和理解互联网海量信息的能力。知识图谱给互联网语义搜索带来了活力，同时也在智能问答中显示出强大威力，已经成为互联网知识驱动的智能应用的基础设施。

知识图谱技术是指知识图谱建立和应用的技术，是融合认知计算、知识表示与推理、信息检索与抽取、自然语言处理与语义 Web、数据挖掘与机器学习等交叉研究，属人工智能重要研究领域知识工程的研究范畴。知识图谱于 2012 年由谷歌提出并成功应用于搜索引擎，是建立大规模知识的一个杀手锏应用。

1.1.1. 知识图谱的概念

1994 年图灵奖获得者、知识工程的建立者费根鲍姆给出的知识工程定义——将知识集成到计算机系统从而完成只有特定领域专家才能完成的复杂任务。在大数据时代，知识工程是从大数据中自动或半自动获取知识，建立基于知识的系统，以提供互联网智能知识服务。大数据对智能服务的需求，已经从单纯的搜集获取信息，转变为自动化的知识服务。我们需要利用知识工程为大数据添加语义/知识，使数据产生智慧（Smart Data），完成从数据到信息到知识，最终到智能应用的转变过程，从而实现对大数据的洞察、提供用户关心问题的答案、为决策提供支持、改进用户体验等目标。知识图谱在下面应用中已经凸显出越来越重要的应用价值：

- 知识融合：当前互联网大数据具有分布异构的特点，通过知识图谱可以对这些数据资源进行语义标注和链接，建立以知识为中心的资源语义集成服务；
- 语义搜索和推荐：知识图谱可以将用户搜索输入的关键词，映射为知识图谱中客观世界的概念和实体，搜索结果直接显示出满足用户需求的结构化信息内容，而不是互联网网页；
- 问答和对话系统：基于知识的问答系统将知识图谱看成一个大规模知识库，通过理解将用户的问题转化为对知识图谱的查询，直接得到用户关心问题的答案；
- 大数据分析与决策：知识图谱通过语义链接可以帮助理解大数据，获得对大数据的洞察，提供决策支持。

当前知识图谱中包含的主要几种节点有：

- 实体：指的是具有可区别性且独立存在的某种事物。如某一个人、某一座城市、某一种植物、某一件商品等等。世界万物有具体事物组成，此指实体。实体是知识图谱中的最基本元素，不同的实体间存在不同的关系。
- 概念：具有同种特性的实体构成的集合，如国家、民族、书籍、电脑等。
- 属性：用于区分概念的特征，不同概念具有不同的属性。不同的属性值类型对应于不同类型属性的边。如果属性值对应的是概念或实体，则属性描述两个实体之间的关系，称为对象属性；如果属性值是具体的数值，则称为数据属性。

1.1.2. 知识图谱的分类

知识图谱的分类方式很多，例如可以通过知识种类、构建方法等划分。从领域上来说，知识图谱通常分为通用（领域无关）知识图谱和特定领域知识图谱：

- 通用知识图谱：通用知识图谱可以形象地看成一个面向通用领域的“结构化的百科知识库”，其中包含了大量的现实世界中的常识性知识，覆盖面极广。
- 特定领域知识图谱：领域知识图谱又叫行业知识图谱或垂直知识图谱，通常面向某一特定领域，可看成是一个“基于语义技术的行业知识库”。

其他分类方式不再一一赘述。

1.2. 知识工程发展历程

回顾知识工程四十年多来发展历程，总结知识工程的演进过程和技术进展，可以将知识工程分成五个标志性的阶段，前知识工程时期、专家系统时期、万维网 1.0 时期，群体智能时期以及知识图谱时期，如图 1 所示。

图 1 知识工程发展历程

- 1950-1970 时期：图灵测试—知识工程诞生前期

人工智能旨在让机器能够像人一样解决复杂问题，图灵测试是评测智能的是手段。这一阶段主要有两个方法：符号主义和连结主义。符号主义认为物理符号系统是智能行为的充要条件，连结主义则认为大脑（神经元及其连接机制）是一切智能活动的基础。这一阶段具有

代表性的工作是通用问题求解程序（GPS）：将问题进行形式化表达，通过搜索，从问题初始状态，结合规则或表示得到目标状态。其中最成功应用是博弈论和机器定理证明等。这一时期的知识表示方法主要有逻辑知识表示、产生式规则、语义网络等。这一时代人工智能和知识工程的先驱 Minsky，Mccarthy 和 Newell 以 Simon 四位学者因为他们在感知机、人工智能语言和通用问题求解和形式化语言方面的杰出工作分别获得了 1969 年、1971 年、1975 年的图灵奖。

- 1970-1990 时期：专家系统—知识工程蓬勃发展期

通用问题求解强调利用人的求解问题的能力建立智能系统，而忽略了知识对智能的支持，使人工智能难以在实际应用中发挥作用。70 年开始，人工智能开始转向建立基于知识的系统，通过“知识库+推理机”实现机器智能，这一时期涌现出很多成功的限定领域专家系统，如 MYCIN 医疗诊断专家系统、识别分子结构的 DENRAL 专家系统以及计算机故障诊断 XCON 专家系统等。斯坦福人工智能实验室的奠基人 Feigenbaum 教授在 1980 年的一个项目报告《Knowledge Engineering：The Applied Side of Artificial Intelligence》中提出知识工程的概念，从此确立了知识工程在人工智能中的核心地位。这一时期知识表示方法有新的演进，包括框架和脚本等。80 年代后期出现了很多专家系统的开发平台，可以帮助将专家的领域知识转变成计算机可以处理的知识。

- 1990-2000 时期：万维网

在 1990 年到 2000 年，出现了很多人工构建大规模知识库，包括广泛应用的英文 WordNet，采用一阶谓词逻辑知识表示的 Cyc 常识知识库，以及中文的 HowNet。Web 1.0 万维网的产生为人们提供了一个开放平台，使用 HTML 定义文本的内容，通过超链接把文本连接起来，使得大众可以共享信息。W3C 提出的可扩展标记语言 XML，实现对互联网文档内容的结构通过定义标签进行标记，为互联网环境下大规模知识表示和共享奠定了基础。这一时期在知识表示研究中还提出了本体的知识表示方法。

- 2000-2006 时期：群体智能

在 2001 年，万维网发明人、2016 年图灵奖获得者 Tim Berners-Lee 在科学美国人杂志中发表的论文《The Semantic Web》正式提出语义 Web 的概念，旨在对互联网内容进行结构化语义表示，利用本体描述互联网内容的语义结构，通过对网页进行语义标识得到网页语义信息，从而获得网页内容的语义信息，使人和机器能够更好地协同工作。W3C 进一步提出万维网上语义标识语言 RDF（资源描述框架）和 OWL（万维网本体表述语言）等描述万维网内容语义的知识描述规范。

万维网的出现使得知识从封闭知识走向开放知识，从集中构建知识成为分布群体智能知识。原来专家系统是系统内部定义的知识，现在可以实现知识源之间相互链接，可以通过关联来产生更多的知识而非完全由固定人生产。这个过程中出现了群体智能，最典型的代表就

是维基百科，实际上是用户去建立知识，体现了互联网大众用户对知识的贡献，成为今天大规模结构化知识图谱的重要基础。

- 2006 年至今：知识图谱—知识工程新发展时期

“知识就是力量”，将万维网内容转化为能够为智能应用提供动力的机器可理解和计算的知识是这一时期的目标。从 2006 年开始，大规模维基百科类富结构知识资源的出现和网络规模信息提取方法的进步，使得大规模知识获取方法取得了巨大进展。与 Cyc、WordNet 和 HowNet 等手工研制的知识库和本体的开创性项目不同，这一时期知识获取是自动化的，并且在网络规模下运行。当前自动构建的知识库已成为语义搜索、大数据分析、智能推荐和数据集成的强大资产，在大型行业和领域中正在得到广泛使用。典型的例子是谷歌收购 Freebase 后在 2012 年推出的知识图谱（Knowledge Graph），Facebook 的图谱搜索，Microsoft Satori 以及商业、金融、生命科学等领域特定的知识库。最具代表性大规模网络知识获取的工作包括 DBpedia、Freebase、KnowItAll、WikiTaxonomy 和 YAGO，以及 BabelNet、ConceptNet、DeepDive、NELL、Probase、Wikidata、XLORE、Zhishi.me、CNDBpedia 等。这些知识图谱遵循 RDF 数据模型，包含数以千万级或者亿级规模的实体，以及数十亿或百亿事实（即属性值和与其他实体的关系），并且这些实体被组织在成千上万的由语义体现的客观世界的概念结构中。

目前知识图谱的发展和应用状况，除了通用的大规模知识图谱，各行业也在建立行业和领域的知识图谱，当前知识图谱的应用包括语义搜索、问答系统与聊天、大数据语义分析以及智能知识服务等，在智能客服、商业智能等真实场景体现出广泛的应用价值，而更多知识图谱的创新应用还有待开发。

在我国知识工程领域研究中，中科院系统所陆汝钤院士、计算所史忠植研究员等老一代知识工程研究学者为中国的知识工程研究和人才培养做出了突出贡献，陆汝钤院士因在知识工程和基于知识的软件工程方面作出的系统和创造性工作，以及在大知识领域的开创性贡献，荣获首届“吴文俊人工智能最高成就奖”。

1.3. 知识图谱的知识图谱

我们根据知识工程生命周期各个阶段的关键技术，利用 AMiner 中近年来知识图谱领域的高水平学术论文，挖掘出了包括知识表示（knowledge representation）、知识获取（knowledge acquisition）、知识推理（knowledge reasoning）、知识集成（knowledge integration）和知识存储（knowledge storage）等相关关键词近年来全球活跃的学术研究。此外，结合知识图谱技术，本报告将以上研究领域表示为三级图谱结构，具体分析和处理的方法如下：

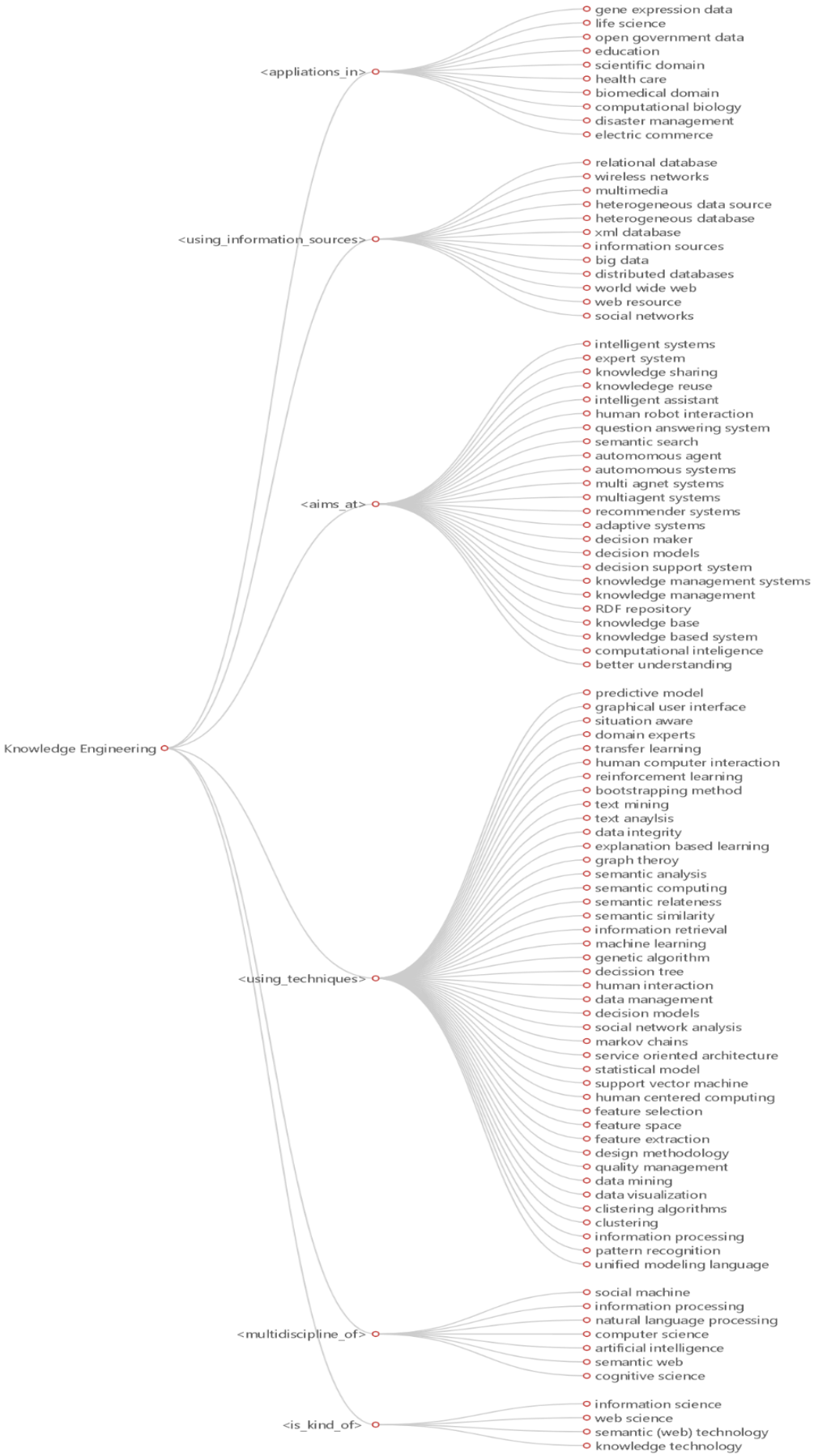
1. 使用自然语言处理技术，提取每篇论文文献的关键词，据此，结合学科领域知识图谱，将文章分配到相应领域；

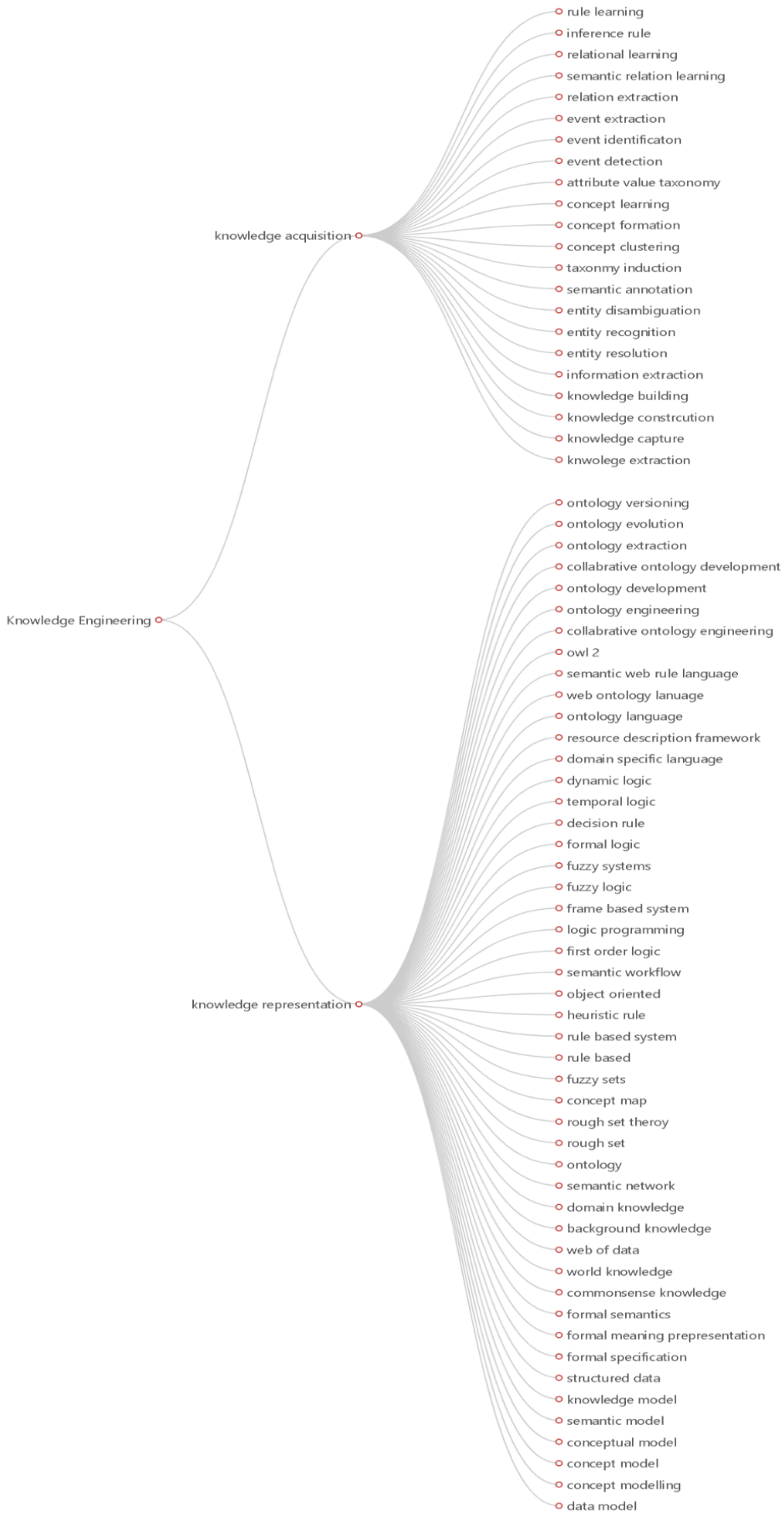
2. 依据学科领域对论文文献进行聚类，并统计论文数量作为领域的研究热度；

3. 领域专家按照领域层级对学科领域划分等级，设计了三级图谱结构，最后根据概念热度定义当前研究热点。

下图是数据挖掘三级知识图谱的可视化表示，详细数据可以参见本报告附录，或到 https://www.aminer.cn/data 中直接下载原始数据。鉴于自动分析技术和论文采集的局限性，图谱还可以进一步完善，欢迎读者批评指正，我们会根据根据读者的反馈定期更新。

注：图中带“<>”的节点表示关系，没有标”<>”的标明的节点关系是上下位关系。


Knowledge Engineering
<appliations_in>
gene expression data
life science
open government data
education
scientific domain
health care
biomedical domain
computational biology
disaster management
electric commerce
<using_information_sources>
relational database
wireless networks
multimedia
heterogeneous data source
heterogeneous database
xml database
information sources
big data
distributed databases
world wide web
web resource
social networks
<aims_at>
intelligent systems
expert system
knowledge sharing
knowledege reuse
intelligent assistant
human robot interaction
question answering system
semantic search
automomous agent
automomous systems
multi agnet systems
multiagent systems
recommender systems
adaptive systems
decision maker
decision models
decision support system
knowledge management systems
knowledge management
RDF repository
knowledge base
knowledge based system
computational inteligence
better understanding
<using_techniques>
predictive model
graphical user interface
situation aware
domain experts
transfer learning
human computer interaction
reinforcement learning
bootstrapping method
text mining
text anaylsis
data integrity
explanation based learning
graph theroy
semantic analysis
semantic computing
semantic relateness
semantic similarity
information retrieval
machine learning
genetic algorithm
decission tree
human interaction
data management
decision models
social network analysis
markov chains
service oriented architecture
statistical model
support vector machine
human centered computing
feature selection
feature space
feature extraction
design methodology
quality management
data mining
data visualization
clistering algorithms
clustering
information processing
pattern recognition
unified modeling language
<multidiscipline_of>
social machine
information processing
natural language processing
computer science
artificial intelligence
semantic web
cognitive science
<is_kind_of>
information science
web science
semantic (web) technology
knowledge technology


Knowledge Engineering
knowledge acquisition
rule learning
inference rule
relational learning
semantic relation learning
relation extraction
event extraction
event identificaton
event detection
attribute value taxonomy
concept learning
concept formation
concept clustering
taxonmy induction
semantic annotation
entity disambiguation
entity recognition
entity resolution
information extraction
knowledge building
knowledge constrcution
knowledge capture
knwolege extraction
knowledge representation
ontology versioning
ontology evolution
ontology extraction
collabrative ontology development
ontology development
ontology engineering
collabrative ontology engineering
owl 2
semantic web rule language
web ontology lanuage
ontology language
resource description framework
domain specific language
dynamic logic
temporal logic
decision rule
formal logic
fuzzy systems
fuzzy logic
frame based system
logic programming
first order logic
semantic workflow
object oriented
heuristic rule
rule based system
rule based
fuzzy sets
concept map
rough set theroy
rough set
ontology
semantic network
domain knowledge
background knowledge
web of data
world knowledge
commonsense knowledge
formal semantics
formal meaning prepresentation
formal specification
structured data
knowledge model
semantic model
conceptual model
concept model
concept modelling
data model

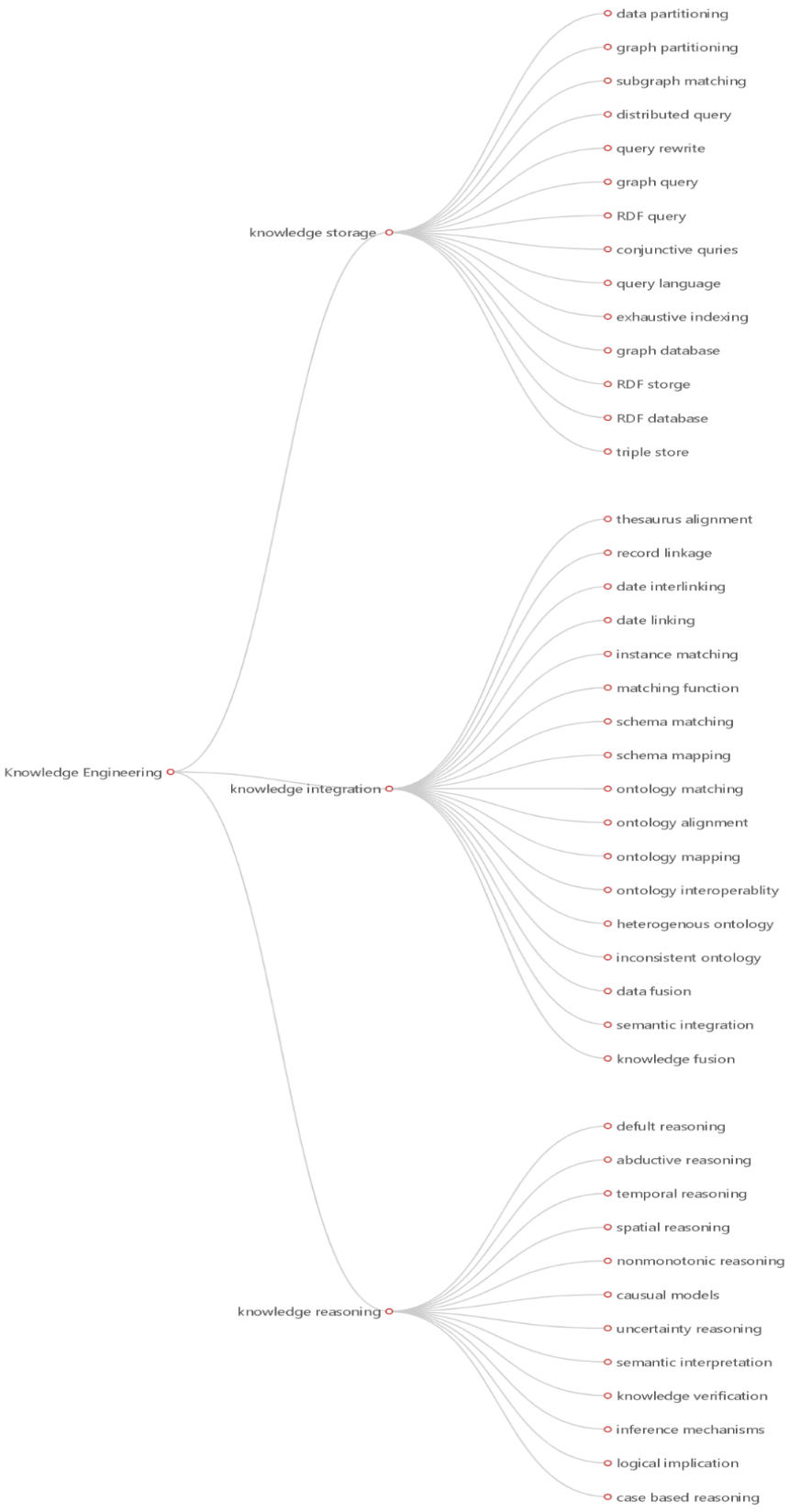


图 2 Knowledge Graph 知识图谱

2. 技术人才篇

知识图谱技术是知识图谱建立和应用的技术，参考中国中文信息学会语言与知识计算专委会发布的《知识图谱发展报告 2018 年版》，我们将知识图谱技术分为知识表示与建模、知识获取、知识融合、知识图谱查询和推理计算及知识应用技术。在大数据环境下，从互联网开放环境的大数据中获得知识，用这些知识提供智能服务互联网/行业，同时通过互联网可以获得更多的知识。这是一个迭代的相互增强过程，可以实现从互联网信息服务到智能知识服务的跃迁。

本报告依据 AMiner 数据平台信息，对知识图谱领域的 5 个细分领域进行梳理，重点介绍每一细分领域研究学者的研究方向与代表性文章，旨在为学术界、产业界提供知识图谱技术及学者的分析依据，同时面向政府机关、高校、企业等对知识图谱技术感兴趣的机构介绍该领域基本概念、研究与应用方向，向科研机构、高技术企业等行业中的专业人士介绍相关领域的前沿技术与发展趋势。

表 1 展示知识图谱领域 10 个相关重要国际学术会议，这些会议为知识图谱领域的研究方向、技术趋势与学者研究成果提供重要信息，为本报告研究学者的选取提供依据。

表 1 知识图谱领域顶级学术会议列表

会议简称	会议全称
ACL	Association of Computational Linguistics
EMNLP	Empirical Methods in Natural Language Processing
WWW	International World Wide Web Conference
ISWC	International Semantic Web Conference
IJCAI	International Joint Conference on Artificial Intelligence
AAAI	National Conference of the American Association for Artificial Intelligence
COLING	International Conference on Computational Linguistics
KR	International Conference on Principles of KR & Reasoning
KDD	ACM International Conference on Knowledge Discovery and Data Mining
CIKM	ACM International Conference on Information and Knowledge Management

图 3 展示的是本报告知识图谱 5 个细分领域的研究学者选取过程：AMiner 选取最近 10 年表 1 展示的重要学术会议为知识图谱领域相关论文作为备选池，在确定细分领域关键词后根据关键词进行细分领域论文的二次分析，最后，按照论文发表数量与工作机构所属国家确定国际、国内知名学者并做简单介绍。

图 3 知识图谱细分领域学者选取流程图

需要说明的是，上述重要学术会议论文发表仅是学者研究水平的一个指标，且报告篇幅有限，难免有遗漏，欢迎指正。另外，为了让读者对了解得更为全面，在进行学者选定时还遵循如下两个原则，1）本报告按照 5 个细分领域进行学者选取，但部分学者可能会在多个领域均有建树，我们仅选择其最突出的领域，其他领域不做重复展示；2）在同一领域内，部分研究机构可能培养多名知名学者，我们仅选择其中一名做介绍。

2.1. 知识表示与建模

2.1.1. 知识表示模型

知识表示将现实世界中的各类知识表达成计算机可存储和计算的结构。机器必须要掌握大量的知识，特别是常识知识才能实现真正类人的智能。本报告 1.2 节在介绍知识工程发展历程的同时也指明了知识表示技术的变化，大致可以分为三个阶段：1）基于符号逻辑进行知识表示和推理，主要包括逻辑表示法（如一阶逻辑、描述逻辑）、产生式表示法和框架表示等。逻辑表示与人类的自然语言比较接近，是最早使用的一种知识表示方法；2）随着语义网概念的提出，万维网内容的知识表示技术逐渐兴起，包括基于标签的半结构置标语言 XML、基于万维网资源语义元数据描述框架 RDF 和基于描述逻辑的本体描述语言 OWL 等，使得将机器理解和处理的语义信息表示在万维网上成为可能，当前在工业界大规模应用的多维基于 RDF 三元组的表示方法；3）随着自然语言处理领域词向量等嵌入（Embedding）技术手段的出现，采用连续向量方式来表示知识的研究（TransE 翻译模型、SME、、SLM、NTN、MLP，以及 NAM 神经网络模型等）正在逐渐取代与上述以符号逻辑为基础知识表示方法相融合，成为现阶段知识表示的研究热点。更为重要的是，**知识图谱嵌入也通常作为一种类型的先验知识辅助输入到很多深度神经网络模型中，用来约束和监督神经网络的训练过程**，如图 4 所示。

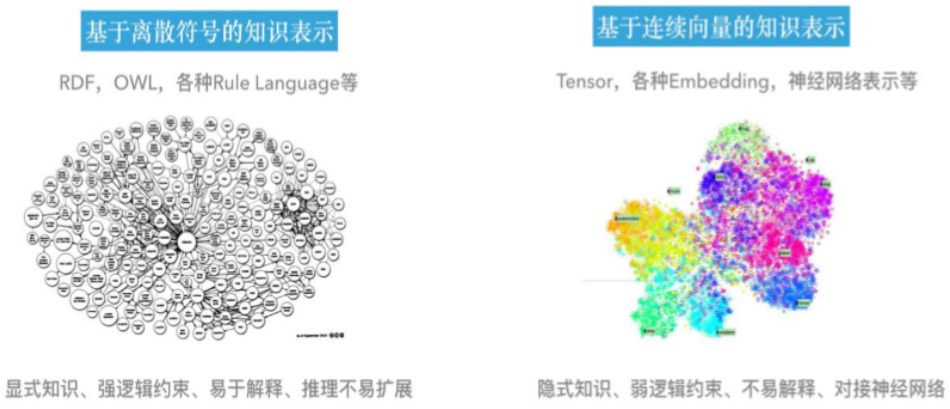


图 4 基于离散符号的知识表示与基于连续向量的知识表示

相比于传统人工智能，知识图谱时代基于向量的知识表示方法不仅能够以三元组为基础

的较为简单实用的知识表示方法满足规模化扩展的要求，还能够作为大数据分析系统的重要数据基础，帮助这些数据更加易于与深度学习模型集成。

2.1.2. 知识表示学习

随着以深度学习为代表的表示学习的发展,面向知识图谱中实体和关系的表示学习也取得了重要的进展。知识表示学习将实体和关系表示为稠密的低维向量实现了对实体和关系的分布式表示已经成为知识图谱语义链接预测和知识补全的重要方法。由于知识表示学习能够显著提升计算效率，有效缓解数据稀疏，实现异质信息融合并有助于实现知识融合，因此对知识库的构建、推理和应用具有重要意义，值得广受关注、深入研究。

知识表示学习是近年来的研究热点，研究者提出了多种模型，学习知识库中的实体和关系的表示。本节将介绍两种代表方法。

（1）复杂关系建模

近期，Bordes 等人受到词向量空间对于词汇语义与句法关系存在有趣的平移不变现象的启发，提出了 TransE 模型，这一模型将知识库中的关系看作实体间的某种平移向量，在大规模知识图谱上效果明显。不过由于 TransE 模型过于简单，导致其在处理知识库的复杂关系时捉襟见肘，为突破 TransE 模型在处理 1-N、N-1、N-N 复杂关系时的局限性，研究学者相继提出了让一个实体在不同关系下拥有不同表示、认为不同关系拥有不同语义空间的 TransH 模型和 TransR 模型，以及针对这两种模型中矩阵参数过多问题再次改进优化的 TransD 模型和 TranSparse 模型。此外，研究学者还提出了利用高斯分布来表示知识库中的实体和关系，可以在表示过程中考虑实体和关系本身语义上不确定性的 TransG 模型和 KG2E 模型。在相关数据集合上的实验表明，这些方法均较 TransE 有显著的性能提升，验证了这些方法的有效性。

（2）关系路径建模

在知识图谱中，多步的关系路径也能够反映实体之间的语义关系。为了突破 TransE 等模型孤立学习每个三元组的局限性，Lin 等人提出考虑关系路径的表示学习方法，以 TransE 作为扩展基础，提出 Path-based TransE（PTransE）模型。几乎同时，其他研究团队在知识表示学习中也成功考虑了关系路径的建模。PTransE 等研究的实验表明，考虑关系路径能够极大提升知识表示学习的区分性，提高在知识图谱补全等任务上的性能。关系路径建模工作较为初步，在关系路径的可靠性计算、语义组合操作等方面还有很多细致的考察工作需要完成。

2.1.3. 知识表示与建模人才介绍

选取 knowledge representation、knowledge modeling、production system、knowledge representation learning、frame language、script representation、knowledge distributed representation、

domain knowledge modeling、taxonomy induction、concept model、knowledge embedding、ontology building 等词作为知识表示与建模领域关键词，按照图 2 所示流程将所选学者定义为该领域知名学者并对其进行统计分析，最终绘制出该领域全球知名学者分布图，分别如图 5、图 6 所示：

图 5 知识表示与建模领域全球知名学者分布图

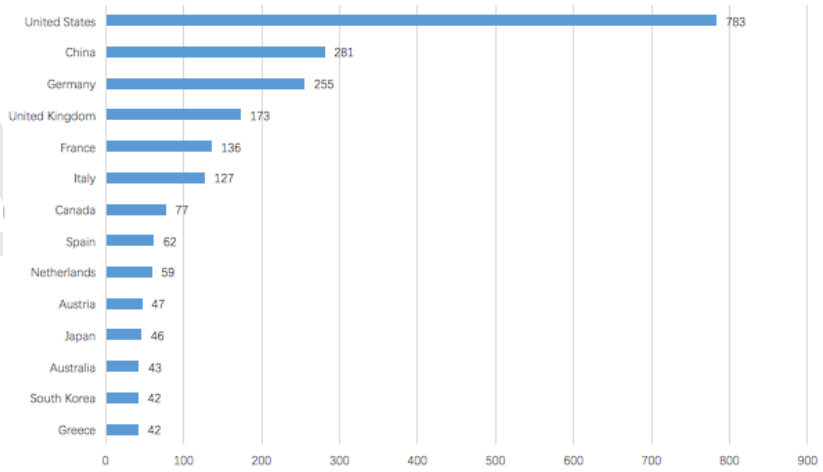


图 6 知识表示与建模领域全球知名学者国家分布统计

由以上两个图可知，全球范围内，北美洲与欧洲是知识表示与建模领域知名学者分布最为集中的地区，亚洲次之，大洋洲、南美洲、非洲等较为匮乏。若按国家进行统计，美国是该领域学者最为集中的国家，境内学者数量多集中分布在东海岸，中国、德国、英国等国家学者数量次之，沙特阿拉伯、斯洛文尼亚等国家人数较少。

对我国知识表示与建模领域知名学者分布进行分析，绘制中国范围内知识表示与建模领域学者分布图，如图 7 所示：

图 7 知识表示与建模领域中国知名学者分布图

由上图可知，中国知识表示与建模领域知名学者人数较为可观，多数学者集中分布在环渤海经济圈、华东以及港澳地区等经济、科研资源相对发达的区域。

对知识表示与建模领域知名学者进行统计，学者工作的科研机构所属国家变更一次即视为迁徙一次，以 0 点所在线为基本线，右侧蓝色表示有学者迁入该国，数值记为正数，左侧绿色表示有学者迁出该国，数值记为负数，红色表示整体人数变化，数值为左右两边数值相加，若为正数则居右显示，反之则居左显示。计算分析后最终绘制出各国知识应用领域人才迁徙图，整体情况图 8 所示：

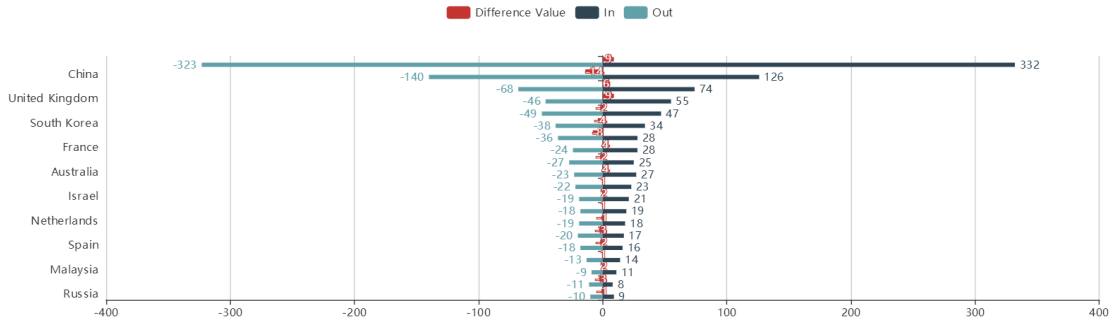


图 8 知识表示与建模领域各国知名学者迁徙图

由上图可知，各国知识表示与建模领域学者流失与引进数量差异较为均衡，美国学者流动幅度大幅度领先，中国、德国、英国等国家紧随其后。这 4 个国家中，3 个发达国家均为学者净流入国家，美国是该领域学者净流入数量最多的国家，中国知识表示与建模领域学者流入量略小于流失量，整体呈现轻微学者流失迹象。

根据 h-index 对知识表示与建模领域全球知名学者进行分析，最终绘制出学者 h-index 分布图，如图 9 所示：

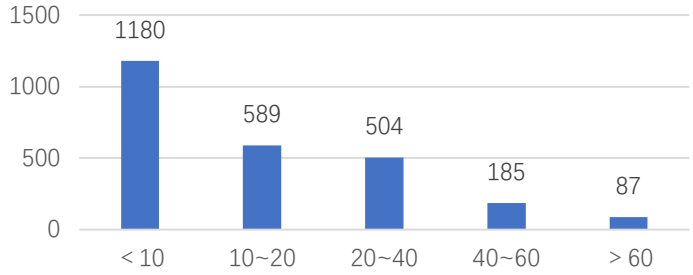
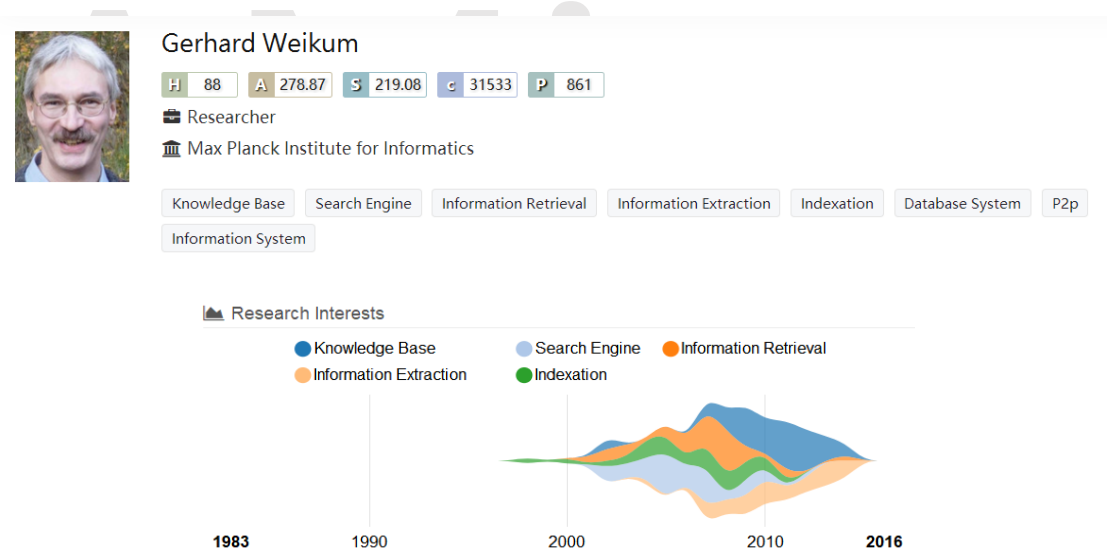


图 9 知识表示与建模领域全球知名学者 h-index 分布图

根据统计信息及上图数据显示可知，知识表示与建模领域学者 h-index 分布呈现金字塔分布结构，大部分学者 h-index 分布在整体的中下区域，其中 h-index 在＜10 区间和 10~20 区间的学者数量最多，h-index＞60 的顶尖学者数量最少，由此可见，知识表示与建模领域学者研究质量差距较大。

受限于本报告篇幅，AMiner 仅选取该领域不同国籍的典型学者做简单介绍，排序不分先后。

- Gerhard Weikum



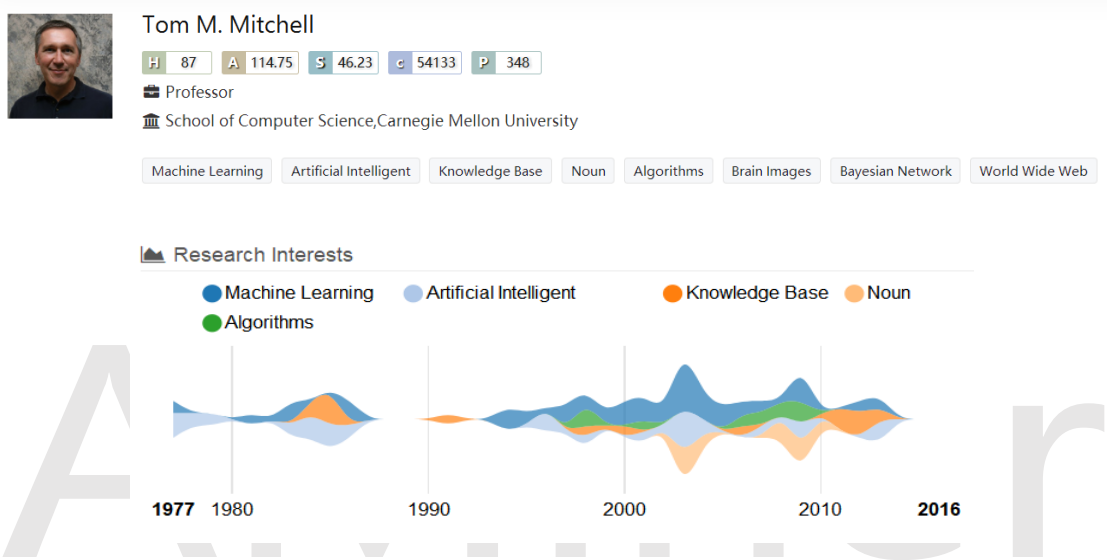
Gerhard Weikum，YAGO 知识库创始人之一，信息抽取与知识挖掘、数据库与信息系统领域著名研究专家。德国萨尔布吕肯 Max-Planck 信息学研究所研究主任，萨尔大学计算机教授，多模式计算与互动卓越集群首席研究员，曾在瑞士苏黎世联邦理工学院、德克萨斯州奥斯汀 MCC 等机构任职。

Gerhard Weikum 的研究涵盖知识获取表示、分布式信息系统、数据库性能优化与自主计算、信息检索与信息提取等方向，2006 年前后侧重于知识库的研究，并在此方向做出了持续性探索。在获得欧洲科学院院士、德国科学与工程院院士、ACM 会士等荣誉的同时，Gerhard Weikum 还曾获得 1998 年 SIGMOD 会议、2006 年 CIKM 会议、2010 年 CIKM 会

议、2018 年 WWW 会议等顶级学术会议最佳论文奖、谷歌聚焦研究奖、Robert Piloty 奖等奖项。

Gerhard Weikum 参与创建的 YAGO 知识库主要集成了 Wikipedia、WordNet 和 GeoNames 三个来源的数据，拥有千万级实体知识，包含超过 1.2 亿条三元组知识，能够将 WordNet 的词汇定义与 Wikipedia 的分类体系进行了融合集成。YAGO 还考虑了时间和空间知识，为很多知识条目增加了时间和空间维度的属性描述，具有更加丰富的实体分类体系，经过人工评估证实确认，准确度达到 95%。

- Tom M. Mitchell



Tom M. Mitchell，NELL 系统、心灵阅读智能计算机系统核心研发成员。美国计算机科学家，卡内基梅隆大学计算机科学学院最高级别 E.Fredkin 讲席教授，曾任卡内基梅隆大学机器学习系首任主席。

Tom M. Mitchell 的研究涵盖知识表示、知识库构建、机器学习、人工智能，机器人和认知神经科学等方向，2000 年至 2010 年间的科研成果较为丰富，长达 40 余年的研究过程中共撰写 130 余篇文章，荣获 NSF 总统青年研究员奖、AAAS 会士、AAAI 会士、美国国家工程院院士以及美国文理科学院院士等荣誉。

Tom M. Mitchell 参与创建的 NELL 系统目标是能够开发用自然语言回答用户提出的问题的方法，而不需要人为干预，自 2010 年初以来，NELL 系统始终保持全天候运行的工作状态，筛选数亿个网页，寻找已知信息与搜索过程中发现的信息之间的联系并建立新的连接，模仿人类学习新信息方式的方式。截至 2010 年 10 月，NELL 系统的准确率已经达到 87%。

- Ian Horrocks

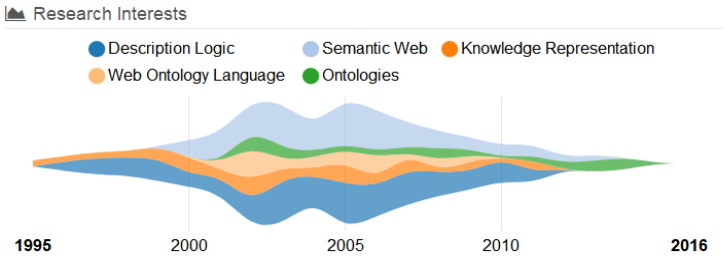
Ian Horrocks

H 107 A 182.90 S 89.77 c 70923 P 494

Professor

Oxford University

Description Logic | Semantic Web | Knowledge Representation | Web Ontology Language | Ontologies | Description Logics | Knowledge Base | Owl

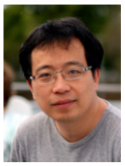


Ian Horrocks，描述逻辑推理系统奠基人。英国牛津大学计算机专业教授，牛津大学奥利尔学院研究员，Web Semantics 期刊主编，国际语义网会议主席。

Ian Horrocks 的研究涵盖描述逻辑、语义网络、知识表达、知识库、网络本体语言等方向，自 1995 年开始，Ian Horrocks 就在知识表达领域开展研究，并在描述逻辑、语义网络领域发表过数量可观的研究成果，现阶段的研究侧重于知识表示和推理，特别是描述逻辑的本体语言和表格决策程序的优化，他所完成的关于描述逻辑的表象推理研究已经成为大多数描述逻辑推理系统的基础。

Ian Horrocks 在网络本体语言 OWL 的开发过程中扮演了奠基人的角色，研究工作构成了本体语言标准的基础，与其他学者共同负责开发的本体交互语言 OIL 和 DAML+OLL 及相关工具已经被开放生物医学本体联合会、美国国家癌症研究所、联合国粮食及农业组织、万维网联合会、一系列大公司和政府机构使用。2005 年荣获英国计算机协会 Roger Needham 奖，2010 年起多次在 AAAI、IJCAI、KR 等学术会议发表相关论文，2011 年当选皇家学会会员。

- 王海勋



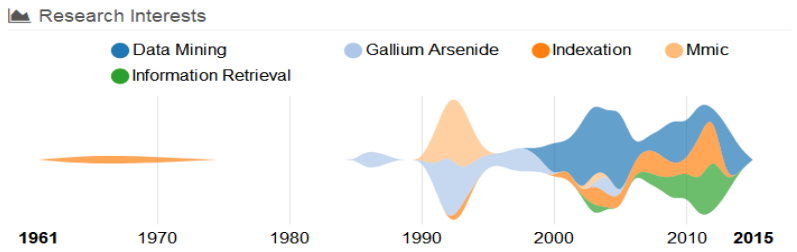
王海勋 (Haixun Wang)

H 77 A 184.46 S 119.04 c 24774 P 588

Research Scientist

We Work

Data Mining | Gallium Arsenide | Indexation | Mmic | Information Retrieval | Ehf | Indexing Terms | Noise Figure

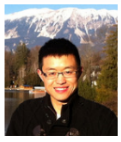


王海勋，WeWork 应用科学研究负责人，首席科学家。IEEE 会士、KAIS、JCST、DAPD、IEEE TKDE 等学术期刊编委，曾任 2013 年 WWW、ICDE、2018 年 CIKM 会议 PC 联合主席，谷歌研究院研究科学家，工程经理，微软亚洲研究院高级研究员，数据管理研究部主任。

王海勋的研究涵盖数据挖掘、信息检索、指数化、单片集成电路等方向，2000 年前后对数据挖掘研究投入较大精力，现阶段研究侧重于语义网络、自然语言处理、数据管理与普适计算等。王海勋在著名学术会议和学术期刊上发表论文 100 余篇，并获得 2008 年 ER 最佳论文奖、2013 年 ICDM 最高影响力奖、2015 年 ICDE 最佳论文奖等奖项。

王海勋被引用量最高的论文是 2003 年在 KDD 会议上发表的“*Mining concept-drifting data streams using ensemble classifiers*”。这篇论文提出了一个使用加权集合分类器挖掘概念漂移数据流的一般框架，经过实验证实该篇论文中所提出的方法在预测精度方面具有优于单分类器方法的显著优势，并且集合框架对于各种分类模型是有效的。

- 唐杰



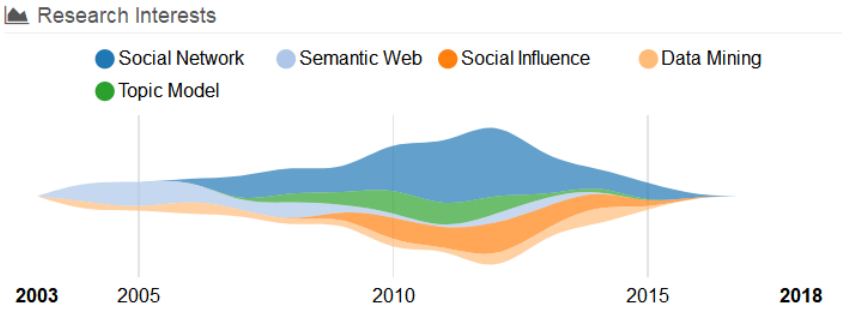
唐杰 (Jie Tang)

H 53 A 167.34 S 60.97 C 10867 P 269

Associate Professor

Department of Computer Science and Technology,Tsinghua University

Social Network | Semantic Web | Social Influence | Topic Model | Data Mining | Predictive Model | Recommender System | Ontology



唐杰，清华大学副教授，AMiner 大数据平台创始人，计算机科学与技术系副主任，CCF YOCSEF 现任主席，国际期刊 ACM TKDD 主编，IEEE TKDE 和 ACM TIST 编委，曾在康奈尔大学、香港科技大学、南安普顿大学等地进行学术访问并多次担任国际顶级学术会议协同主席、副主席等职务。

唐杰提出多项创新性研究并在这些研究的基础上研发出研究者社会网络 ArnetMiner 系统，该系统曾在国际顶级会议 WWW、KDD、ISWC、ICDM 中进行演示并获得一致好评，系统数据被广泛应用于科学研究，在国际上具有较高的影响力。

唐杰的高引用论文是 2008 年在 KDD 会议上发表的“*ArnetMiner: extraction and mining of academic social networks*”对其负责的知识工程实验室 ArnetMiner 系统关键问题进行讨论，整合来自在线 Web 数据库的出版物并提出一个概率框架来处理名称歧义问题，除此之外，该篇论文还描述了系统的体系结构和专家画像的主要特征，提出系统应用方法的实证评估。

2.2. 知识获取

2.2.1. 实体识别与链接

实体识别与链接是海量文本分析的核心技术，为解决信息过载提供了有效手段。实体识别是文本理解意义的基础，也就是识别文本中指定类别实体的过程，可以检测文本中的新实体，并将其加入到现有知识库中。实体链接是识别出文本中提及实体的词或者短语并与知识库中对应实体进行链接的过程，通过发现现有实体在文本中的不同出现，可以针对性的发现关于特定实体的新知识。实体识别与链接是知识图谱构建、知识补全与知识应用的核心技术，为计算机类人推理和自然语言理解提供知识基础。

本章节介绍三种统计模型方法中的实体识别与链接：

（1）传统统计模型方法

实体识别：自 90 年代以来，统计模型一直是实体识别的主流方法。最大熵分类模型、SVM 模型、隐马尔可夫模型、条件随机场模型等统计方法都曾被用来抽取文本中的实体识别，其中的条件随机场模型作为实体识别的代表性统计模型能够将实体识别问题转化为序列标注问题。

实体链接：实体链接在传统模型中的核心在于挖掘可用于识别提及目标实体相互关联的证据信息，目前主要使用的证据信息包括实体统计信息、名字统计信息、上下文词语分布、实体关联度、文章主题等信息。同时，考虑到一段文本中实体之间的相互关联，相关的全局推理算法也被提出用来寻找全局最优决策。

（2）深度学习方法

实体识别：目前存在两类用于命名实体识别的典型深度学习架构，一种是 NN-CRF 架构，在该架构中，CNN/LSTM 被用来学习每一个词位置处的向量表示，基于该向量表示 NN-CRF 解码该位置处的最佳标签，第二种是采用滑动窗口分类的思想，使用神经网络学习句子中的每一个 N-Gram 的表示，然后预测该 N-Gram 是否是一个目标实体。

实体链接：实体链接在深度学习的核心是构建多类型多模态上下文及知识的统一表示，并建模不同信息、不同证据之间的相互交互，通过将不同类型的信息映射到相同的特征空间，并提供高效的端到端训练算法。

（3）文本挖掘方法

文本挖掘方法应用于半结构 Web 数据源上的语义知识获取，工作核心是从特定结构（如列表、Infobox）构建实体挖掘的特定规则，代表性文本挖掘抽取系统包括 DBPedia、YAGO、BabelNet、NELL 和 Kylin 等。由于规则本身可能带有不确定性和歧义性，同时目标结构可能会有一定的噪音，文本挖掘方法往往基于特定算法来对语义知识进行评分和过滤。此外，人们发现结构化数据源只包含有限类别的实体，对长尾类别覆盖不足，另一方面，实体获取技术往往采用 Bootstrapping 策略，充分利用大数据的冗余性，开放式的从 Web 中获取指定类型的实体。该部分的代表性的工作包括 TextRunner 系统和 Snowball 系统。开放式实体集合扩展的主要问题是语义漂移问题，近年来的主要工作集中在解决该问题。具体技术包括互斥 Bootstrapping 技术、Co-Training 技术和 Co-Bootstrapping 技术。

2.2.2. 实体关系学习

实体关系定义为两个或多个实体间的某种联系，用于描述客观存在的事物之间的关联关系。实体关系学习就是自动从文本中检测和识别出实体之间具有的某种语义关系，也称为关系抽取。实体关系抽取分为预定义关系抽取和开放关系抽取。预定义关系抽取是指系统所抽取的关系是预先定义好的，如上下位关系、国家—首都关系等。开放式关系抽取不预先定义抽取的关系类别，由系统自动从文本中发现并抽取关系。实体关系识别是知识图谱自动构建和自然语言理解的基础。

本章节从不同维度对现有关系抽取的技术方法和研究现状进行介绍：

（1）限定域关系抽取和开放域关系抽取

限定域关系抽取是指系统所抽取的关系是预先定义好的，预定义关系个数有限。这类抽取可以抽取语义化的实体关系三元组，方便用于辅助其它任务。

开放域关系抽取是指不预先定义关系，由系统自动从文本中发现、抽取关系。由于开放域关系抽取难以抽取语义化三元组，近年来，越来越多的研究者关注限定域关系抽取。

（2）基于规则的关系抽取和基于机器学习的关系抽取

所谓基于规则的关系抽取方法是指首先由通晓语言学知识的专家根据抽取任务的要求设计出一些包含词汇、句法和语义特征的手工规则（或称为模式），然后在文本分析的过程中寻找与这些模式相匹配的实例，从而推导出实体之间的语义关系。

按照机器学习方法对语料库的不同需求大致可分成三大类：无监督关系抽取，有监督关系抽取、弱监督关系抽取。无监督关系抽取希望把表示相同关系的模版聚合起来，不需要人工标注的数据。有监督关系抽取使用人工标注的训练语料进行训练。有监督关系抽取目前可以取得最好的抽取效果，但是由于其需要费时费力的人工标注，难以应用到大规模场景。因此有学者提出了利用知识库回标文本来自动获得大量的弱监督数据，目前弱监督关系抽取是关系抽取领域的一大热点。

2.2.3. 事件知识学习

事件是促使事物状态和关系改变的条件，是动态的、结构化的知识。目前已存在的知识资源（如谷歌知识图谱）所描述多是实体以及实体之间的关系，缺乏对事件知识的描述。事件知识学习，就是将非结构化文本中自然语言所表达的事件以结构化的形式呈现，对于知识表示、理解、计算和应用意义重大。知识图谱中的事件知识隐含互联网资源中，包括已有的结构化的语义知识、数据库的结构化信息资源、半结构化信息资源以及非结构化资源，不同性质的资源有不同的知识获取方法。

考虑到事件识别和抽取、事件检测和追踪两个任务的处理对象、着眼点和技术路线的差异，本章节对其主流方法和现状分别进行梳理。

（1）事件识别和抽取

根据抽取方法，事件抽取可以分为基于模识匹配的事件抽取和基于机器学习的事件抽取。

基于模式匹配的事件抽取方法是指对某种类型事件的识别和抽取是在一些模式的指导下进行的，模识匹配的过程就是事件识别和抽取的过程。采用模式匹配的方法进行事件抽取的过程一般可以分为两个步骤：模式获取和模式匹配。模式准确性是影响整个方法性能的重要因素，按照模式构建过程中所需训练数据的来源可细分为基于人工标注语料的方法和弱监督的方法。

基于机器学习的事件抽取方法建立在统计模型基础上，一般将事件抽取建模成多分类问题，因此研究的重点在于特征和分类器的选择。根据利用信息的不同可以分为基于特征、基于结构和基于神经网络三类主要方法。

基于特征的方法：研究重点在于如何提取和集成具有区分性的特征，从而产生描述事件实例的各种局部和全局特征，作为特征向量输入分类器。该类方法多用于阶段性的管道抽取，即顺序执行事件触发词识别和元素抽取。

基于结构的方法：将事件结构看作依存树，抽取任务则相应地转化为依存树结构预测问题，触发词识别和元素抽取可以同时完成。

神经网络的方法：利用卷积神经网络模型抽取特征来完成两阶段的识别任务以便更好地

考虑事件内部结构和各个元素间的关系。将联合抽取模型与 RNN 相结合，利用带记忆的双向 RNN 抽取句子中的特征，并联合预测事件触发词和事件元素，进一步提升了抽取效果。

（2）事件检测和追踪

基于相似度的方法首先需要定义相似度度量，而后基于此进行聚类或者分类。Yang 等提出在 TDT 中用向量空间模型（Vector Space Model，VSM）对文档进行表示，并提出了组平均聚类（Group Average Clustering，GAC）和单一通过法（Single Pass Algorithm，SPA）两种聚类算法。GAC 只适用于历史事件发现，它利用分治策略进行聚类。SPA 可以顺序处理文档并增量式产生聚类结果，能同时应用于历史事件发现和在线事件发现。

概率统计方法通常使用生成模型，由于需要大量数据的支持，所以这种方法更加适用于历史事件检测。对比基于相似度聚类的模型，这类模型虽然复杂，但当数据量充足时，通常可以取得更高的准确率。基于概率的方法是目前 TDT 中的研究热点，主要分成两个方向，一是针对新闻等比较正式的规范文档，另一个则用于不规则或没有规律的非规范文档。

（3）事件知识库构建

已有知识图谱，如 DBpedia，YAGO 和 Wikidata 等均侧重于实体的客观属性及实体间的静态关联，缺乏结构化的事件数据。事件知识学习的最终目的就是从非结构化的文本数据中抽取结构化的事件表示，构建事件知识库弥补现有知识图谱的动态事件信息缺失问题。目前事件知识库构建的研究处于起步阶段，基础就是上述两方面研究，基于句子级的事件抽取和文档级的事件发现。

2.2.4. 知识获取人才介绍

选取 knowledge extraction、knowledge discovery、knowledge acquisition、knowledge learning、ontology learning、ontology extraction、ontology generation、ontology acquisition、event knowledge、domain knowledge、common sense、name entity recognition、entity linking、relation extraction、relation classification、event learning、event extraction、event detection and tracking、semantic annotation 等词作为知识获取领域关键词，按照图 2 所示流程将所选学者定义为该领域知名学者并对其进行统计分析，最终绘制出该领域全球知名学者分布图，分别如图 10、图 11 所示：

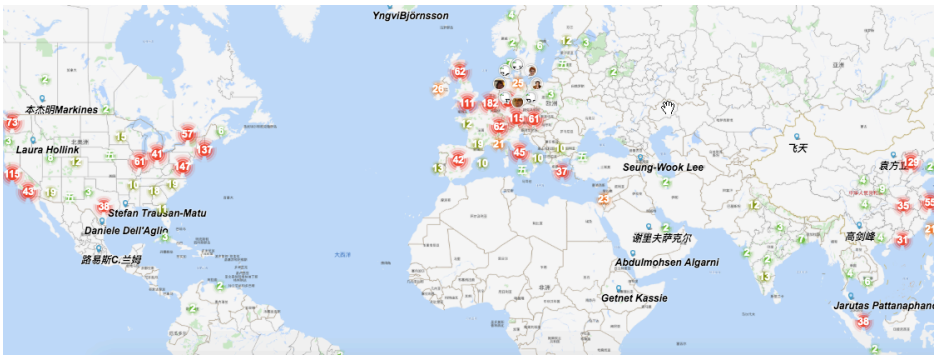


图 10 知识获取领域全球知名学者分布图

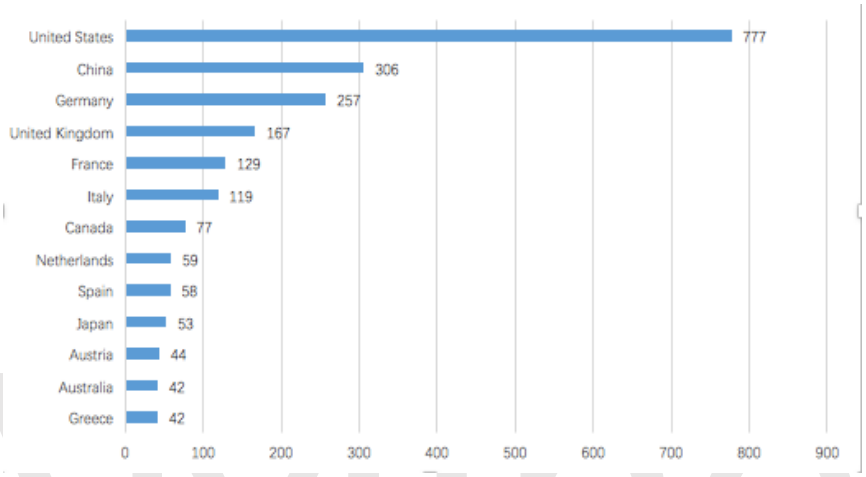


图 11 知识获取领域全球知名学者分布统计

由以上两图可知，全球范围内，符合筛选条件的知识获取领域学者集中分布在北美洲，欧洲、亚洲次之，大洋洲、南美洲、非洲等较为匮乏。若按国家进行统计，美国是该领域学者最为集中的国家，境内学者数量多集中分布在东海岸，中国、德国、英国等国家学者数量次之，其他国家人数较少。

对符合上述条件的我国知识获取领域学者分布进行分析，绘制中国范围内知识获取领域知名学者分布图，如图 12 所示：

图 12 知识获取领域中国知名学者分布图

中国知识获取领域知名学者人数较为可观，其中多数学者集中分布在环渤海经济圈、东南沿海以及港澳地区等经济、科研资源相对发达的城市。

对知识获取领域知名学者进行计算分析，最终绘制出该领域各国人才迁徙图，整体情况图 13 所示：

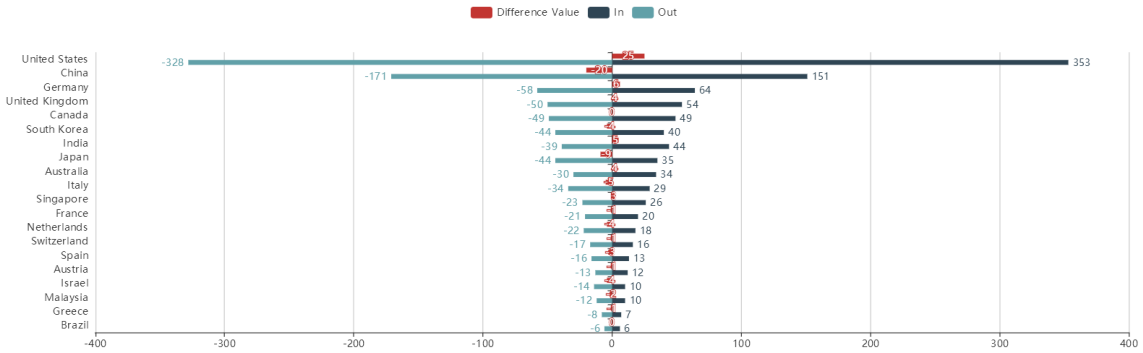


图 13 知识获取领域各国知名学者迁徙图

由上图可知，各国知识获取领域学者流失与引进数量差异较为均衡，美国作为全球该领域学者数量最多的国家，学者流动幅度大幅度领先，也是全球学者流入量最多的国家，中国的学者流入量略小于流失量，整体呈现出轻微的学者流失迹象。

根据 h-index 对知识获取领域全球知名学者进行分析，最终绘制出学者 h-index 分布图，如图 14 所示：

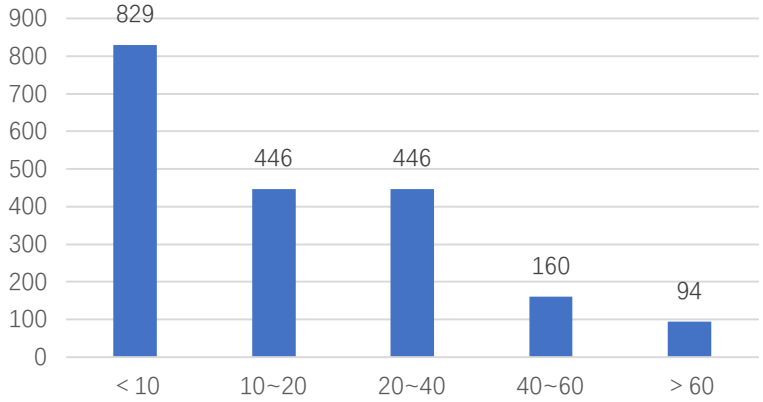
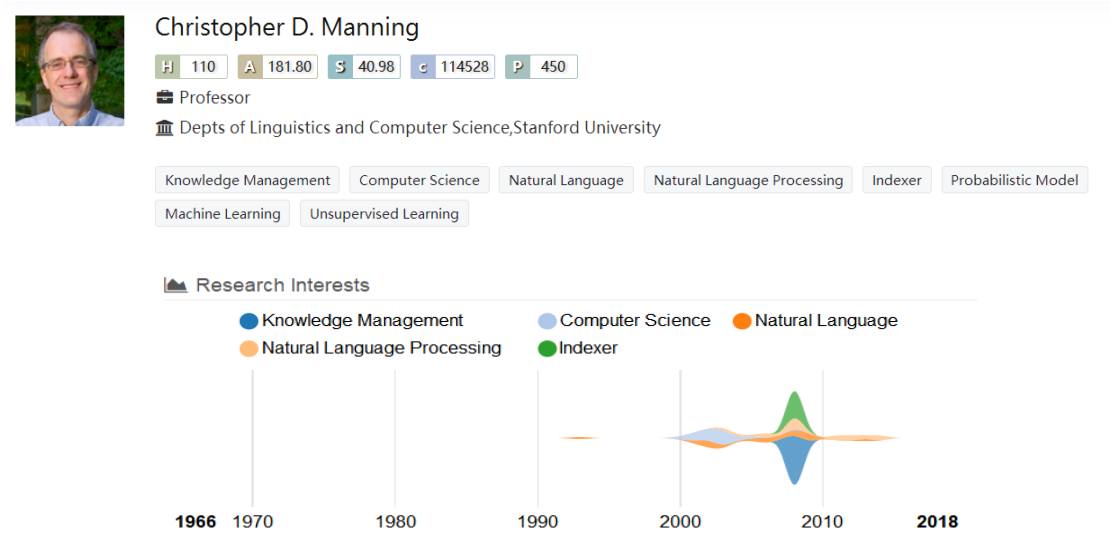


图 14 知识获取领域全球知名学者 h-index 分布图

根据统计信息及上图数据显示可知，知识获取领域学者 h-index 分布呈现金字塔分布结构，大部分学者 h-index 分布在整体的中下区域，其中 h-index 在＜10 区间和 10~20 区间的学者数量最多，h-index＞60 的顶尖学者数量最少，由此可见，知识获取领域学者研究质量差距较大。

受限于本报告篇幅，AMiner 仅选取该领域不同国籍的典型学者做简单介绍，此次排序不分先后。

- Christopher D.Manning

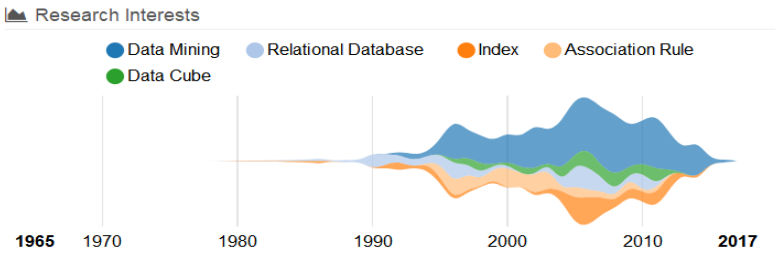


Christopher D.Manning，斯坦福 NLP 实验室创始人。斯坦福大学计算机科学与语言学系教授，Thomas M. Siebel 机器学习首席教授，计算机语言学协会主席。

Christopher D.Manning 的研究涵盖知识管理、计算机科学、自然语言处理等方向，2000 年前几乎未在知识获取领域做出过相关研究，2000 年后，Christopher D.Manning 从 Computer Science 领域入手展开对知识应用领域的研究工作，并在短时间内取得突破，在 2008 年前后发表了数量可观研究成果。荣获 ACM 会士、AAAI 会士、ACL 会士等荣誉的同时还获得了 ACL、COLING、EMNL 等顶级学术会议最佳论文奖等奖项。

斯坦福 NLP 小组包括计算机科学系与语言学系成员，工作范围从计算语言学的基础研究到人类语言技术的关键应用，涵盖句子翻译、句法分析与标记、自动问答、机器翻译、文本及视觉场景模拟等领域。将复杂深入的语言建模、数据分析与创新的概率以及机器学习和 NLP 深度学习方法有效结合是斯坦福 NLP 集团的显著特点，能够为用户提供最先进的词性标注器、高速、高性能的神经网络依赖解析器以及能够处理阿拉伯文、中文、法文、德文和西班牙文文本的算法。

- 韩家炜（Jiawei Han）



韩家炜（Jiawei Han），知识获取领域著名专家。美国伊利诺伊大学厄巴纳-香槟分校计算机教授，美国陆军研究实验室网络科学协作技术联盟计划成员，信息网络学术研究中心主任。曾在 100 多个国际会议和研讨会计划委员会中担任主席或任职，包括 2005 年 IEEE 的 PC 联合主席，国际数据挖掘会议 ICDM 主席，2006 年美国大型数据库国际会议协调员等职务。

韩家炜（Jiawei Han）的研究涵盖知识获取、数据挖掘、数据库系统、关联规则、时空数据挖掘、Web 数据及信息网络数据等方向，1990 年前后侧重于数据挖掘的研究并一直延续至今。在获得 ACM 会士、IEEE 会士等荣誉的同时，韩家炜还曾获得 2004 年 ACM SIGKDD 创新奖、2005 年 IEEE 技术成就奖以及 2009 年 IEEE 最高技术奖 McDowell 奖等奖项。

韩家炜的高引用论文是 2000 年在 SIGMOD 会议上发表的“*Mining frequent patterns without candidate generation*”提出了一种全新的、可以用于数据的存储压缩的、关于频繁模式关键信息的频繁模式树结构，并开发一种有效的基于频繁模式树结构（FP 树结构）的挖掘方法 FP-growth，用于通过模式片段增长挖掘整套频繁模式。采用将大型数据库压缩成高度精简结构、基于 FP 树的挖掘采用模式片段、基于不同分区分而治之等方法的实验结果最终证实论文提出的 FP 树结构能够减少搜索空间，对于挖掘长和短频繁模式是有效且可扩展的，相比于其他算法更为高效。

- Ralph Grishman

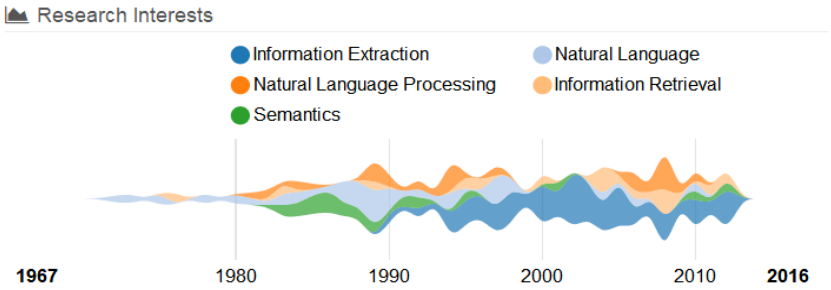
Ralph Grishman

H 63 A 45.72 S 12.30 c 18684 P 325

Professor

Department of Computer Science,Courant Institute of Mathematical Sciences,New York University

Information Extraction | Natural Language | Natural Language Processing | Information Retrieval | Semantics | Syntax | Machine Translation | Event Extraction

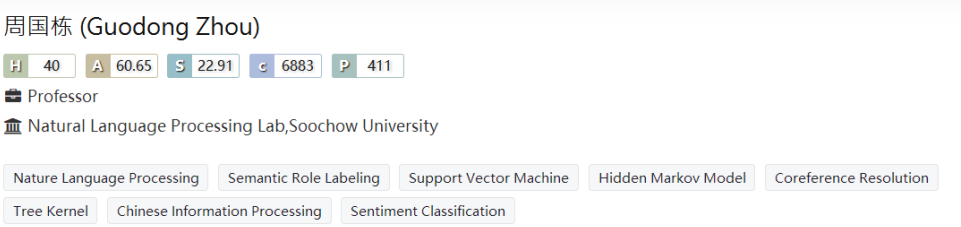


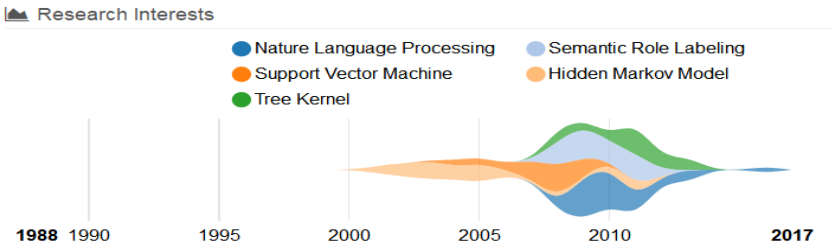
Ralph Grishman，纽约大学数学科学院计算机科学教授，主持创建的 Proteus 项目对自然语言处理领域展开了广泛的研究。

Ralph Grishman 的研究涵盖自然语言处理、信息检索、信息抽取、语义学、知识获取、机器翻译等方向，早在 1967 年便在这一领域开始研究工作，自 1990 年逐渐加大对信息抽取的研究力度并在此领域建树颇多，连续多年在 ACL、CIKM、COLING 等学术会议上发表相关专业论文并担任 2000 年 ACL 北美分会执行委员会成员、2010 年-2015 年美国国家标准与技术研究院文本分析会议组织委员会成员等职务。

Ralph Grishman 的高引用论文“*A maximum entropy approach to named entity recognition*”介绍了一种新的统计命名实体（即“专有名称”）识别系统，称为“MENE”。命名实体（NE）识别是一种信息提取形式，将文档中的每个单词分类为人名、组织、位置、日期、时间、货币价值、百分比或“以上都不是”。对互联网搜索引擎、机器翻译、文档的自动索引以及作为更复杂的工作的基础具有特别重要的意义。

- 周国栋





周国栋，苏州大学计算机科学与技术学院特聘教授，苏州大学自然语言处理实验室创建人，中国人工智能学会自然语言理解专委会和 CCF 中文信息技术专委会副主任委员。

周国栋的研究涵盖自然语言处理、知识获取、信息抽取、隐马尔科夫模型研究等方向，主持完成多项国家级科研项目，曾任国际顶级期刊《Computational Linguistics》杂志编委，NSFC 信息学部会评专家以及许多著名的国际杂志和会议的评审和委员会委员。近年来在 ACL、COLING、IJCAI 等国际顶级会议发表相关专业论文超过 80 篇，并获得 IEEE 会士、ACM 会士、ACL 会士等多项荣誉。

周国栋高引用论文是 2002 年在 ACL 上发表的“*Named entity recognition using an HMM-based chunk tagger*”提出了一种隐马尔科夫模型和一种基于该模型的模块标记器，从中建立了一个命名实体识别系统用于识别并分类名称、时间与数量。通过这一模型系统能够应用和整合四种类型的内部和外部证据，从而有效地解决 NER 问题，基于该系统对 MUC-6 和 MUC-7 英语 NE 任务的系统评估分别达到 96.6%和 94.1%，性能明显优于任何其他机器学习系统。

- 黄萱菁

黄萱菁，复旦大学计算机科学技术学院教授，中国中文信息学会理事，《中文信息学报》编委，中国计算机学会中文信息技术专委会委员，中国人工智能学会自然语言理解专业委员会委员。

黄萱菁的研究涵盖问答系统、自然语言处理、中文信息编译等方向，2005 年至 2010 年间有多项研究成果产出，曾多次在人工智能、自然语言处理和信息检索的国际学术会议 IJCAI、ACL、SIGIR、WWW、EMNLP、COLING、CIKM、WSDM 担任程序委员会委员、资深委员、竞赛主席等职务。在 SIGIR、ACL、ICML、IJCAI、AAAI、NIPS、CIKM、ISWC、EMNLP、WSDM 和 COLING 等多个国际学术会议上发表论文数十篇的同时收获 ACM 会士、ACL 会士等多项荣誉。

黄萱菁的高引用论文是 2009 年在 EMNLP 上发表的“P*hrase dependency parsing for*

opinion mining”提出了一种从产品评论中挖掘挖掘意见的新方法，这种方法将意见挖掘任务转换为识别产品特征、意见表达和它们之间关系，通过多种产品的特征是基于短语所观察的特点引入了短语依赖性解析的概念，将传统的依赖性解析扩展到短语级别，然后实现了该概念以提取产品特征和意见表达之间的关系，经过实验评估表明挖掘任务可以受益于该种方法的短语依赖性解析。

2.3. 知识融合

知识图谱可以由任何机构和个人自由构建，其背后的数据来源广泛、质量参差不齐，导致它们之间存在多样性和异构性。语义集成的提出就是为了能够将不同的知识图谱融合为一个统一、一致、简洁的形式，为使用不同知识图谱的应用程序间的交互建立操作性。常用的技术包括本体匹配（也称为本体映射）、实力匹配（也称为实体对齐、对象公指消解）以及知识融合等。

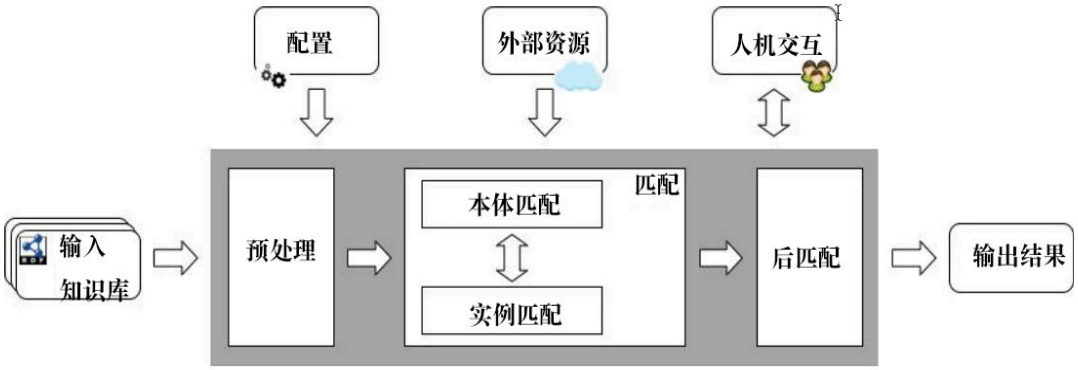


图 15 语义集成的常见流程

一个语义集成的常见流程，主要包括：输入、预处理、匹配、知识融合和输出 5 个环节。语义集成的输入包括待集成的若干个知识库以及配置、外部资源等，如图 15 所示。

待集成的知识库格式一般为 RDF/OWL 数据文件或 SPARQL 端点（endpoint）。外部资源是语义集成过程中使用到的背景知识，例如字/辞典背景知识（例如 WordNet）、常识背景知识（例如 Cyc）、实时背景知识（例如搜索引擎）等。

预处理主要包括对输入知识库进行清洗和后续步骤的准备。清洗主要是为了解决输入质量问题，与自有文本不同，知识库通常基于 RDF/OWL 语言构建，质量较好。

后续步骤的准备分为配置和数据两方面。根据匹配对象的不同，匹配一般分为本体匹配和实力匹配两方面。文本相似性度是发现匹配的最基础方法，大致可分为四种类型：基于字符的（例如 Leven-shtein 编辑距离）、基于单词的（例如 Jaccard 系数）、混合型（例如 soft TF-IDF）和基于语义的（例如 WordNet）。

在匹配的基础上，知识融合一般通过冲突检测、真值发现等技术消解知识集成过程中的

冲突，再对知识进行关联与合并，最终形成一个一致的结果。语义集的输出是一个统一的、一致的、简洁的知识库。

2.3.1. 本体匹配

伴随链接数据的蓬勃发展，本体的数量越来越多。现有大多数本体匹配方法处理的是成对的本体，但是成对匹配方法在同时匹配多个本体时会产生一些问题，最主要的问题是它们得到的结果从全局看可能存在冲突。LPHIM 是一种多文本全体匹配方法，能够在匹配多个本体的同时保证结果是全局最优解。

随着多语言知识库的发展，跨语言本体匹配方法的重要性已经凸显。由于语言不同，跨语言本体匹配相较一般本体匹配更为困难，特别是影响文本相似性度量的准确性。较有代表性的工作包括：EAFG 和双语主题模型。

2.3.2. 实例匹配

众包和主动学习等人机协作方法是目前实例匹配的研究热点。这些方法雇佣普通用户，通过付出较小的人工代价来获得丰富的先验数据，从而提高匹配模型的性能。

随着表示学习技术在诸如图像、视频、语言、自然语言处理等领域的成功，一些研究人员开始着手研究面向知识图谱的表示学习技术，将实体、关系等转换成一个低维空间中的实质向量（即分布式语义表示），并在知识图谱补全、知识库问答等应用中取得了不错的效果。

近年来，强化学习取得了一些列进展，如何在语义集成中运用强化学习逐渐成为新的动向。ALEX 是一个通过利用用户提供的查询答案反馈来提高实例匹配质量的系统，它将每个匹配视作一个状态，用户反馈被转换为行为奖励，通过最大化收集到的行为奖励改善策略。

2.3.3. 知识融合人才介绍

选取 knowledge integration、knowledge linking、knowledge fusion、semantic integration heterogeneous knowledge、ontology matching、ontology alignment、linked data、linked-data、linked open data，instance matching，instance mapping，ontology mapping，entity matching，schema matching 等词作为知识融合领域关键词，按照图 2 所示流程将所选学者定义为该领域知名学者并对其进行统计分析，最终绘制出该领域全球知名学者分布图，分别如图 16、图 17 所示：

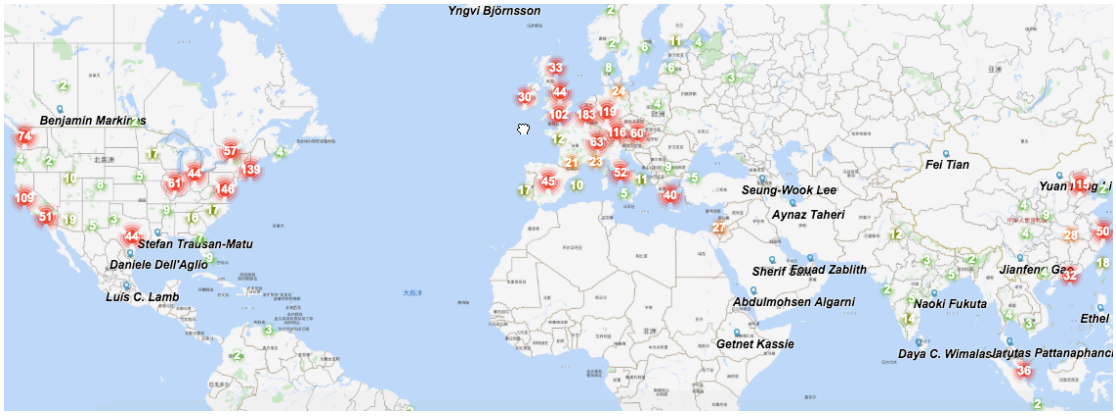


图 16 知识融合领域全球知名学者分布图

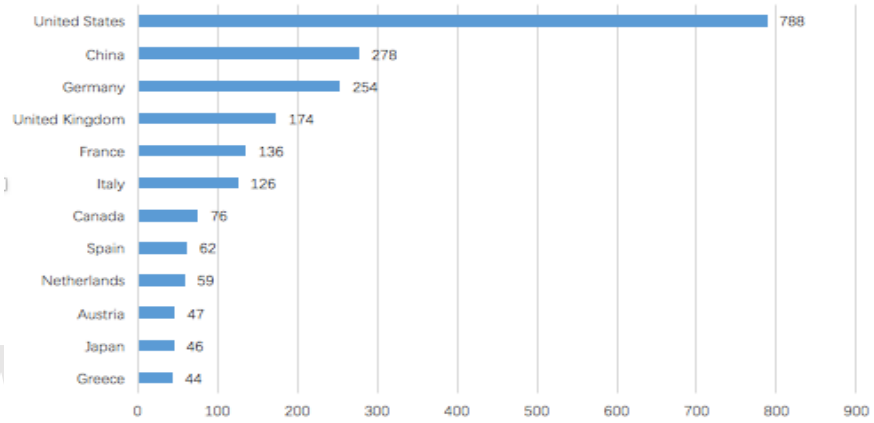


图 17 知识融合领域全球知名学者分布统计

由以上两图可知，全球范围内，符合筛选条件的知识融合领域知名学者集中分布在欧洲及北美洲，亚洲次之，大洋洲、南美洲等较为匮乏。若按国家进行统计，美国是该领域学者最为集中的国家，境内学者数量多集中分布在东海岸，德国、中国、英国等国家学者数量次之，其他国家人数较少。

对符合上述条件的我国知识融合领域学者分布进行分析，绘制中国范围内知识融合领域知名学者分布图，如图 18 所示：

图 18 知识融合领域中国知名学者分布图

中国知识融合领域知名学者人数较少，其中多数学者集中分布在环渤海经济圈以及华东地区，地域性明显。

对知识融合领域知名学者进行计算分析，最终绘制出该领域各国人才迁徙图，整体情况图 19 所示：

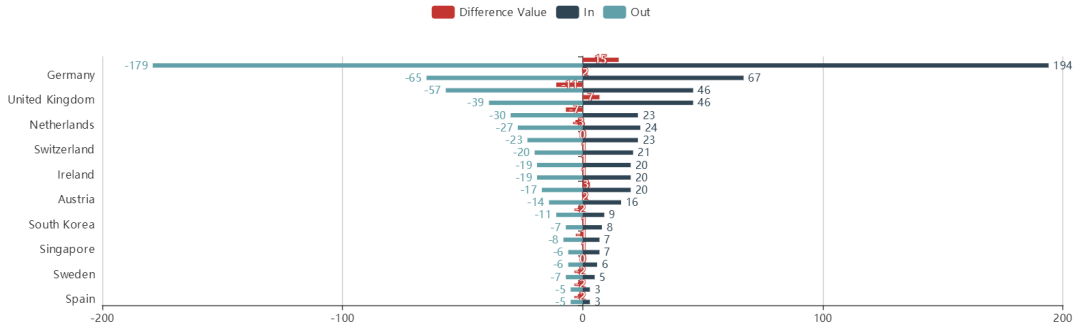


图 19 知识融合领域各国知名学者迁徙图

各国知识融合领域学者流失与引进数量差异较为均衡，美国作为全球该领域学者数量最多的国家，学者流动幅度大幅度领先，同样也是该领域知名学者流入量最多的国家，德国、中国、英国等国家紧随其后，学者流动幅度较高的国家中，中国的学者流入量略小于流失量，整体呈现出轻微的学者流失迹象。

根据 h-index 对知识融合领域全球知名学者进行分析，最终绘制出学者 h-index 分布图，如图 20 所示：

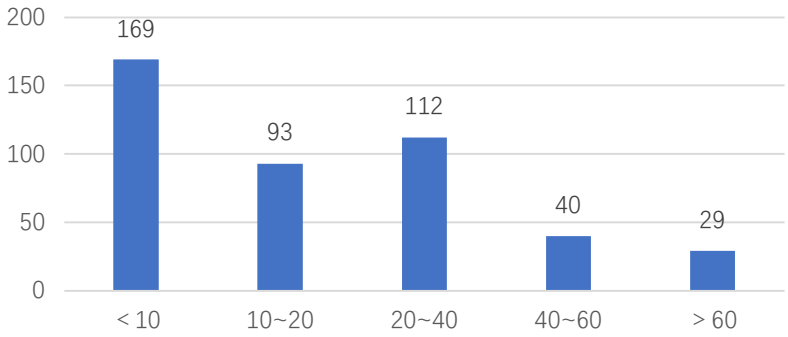
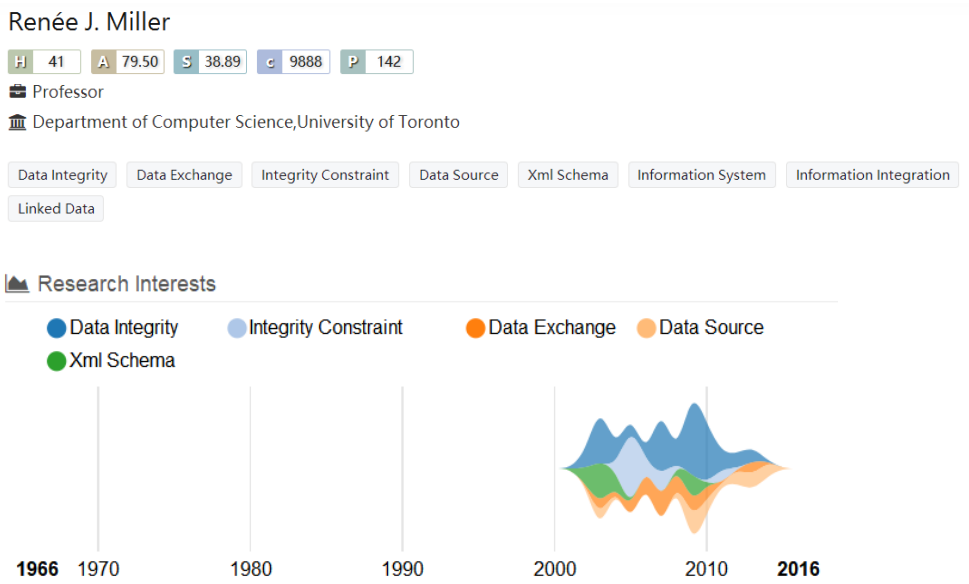


图 20 知识融合领域全球知名学者 h-index 分布图

根据统计信息及上图数据显示可知，知识融合领域学者 h-index 分布并不均衡，大部分学者 h-index 分布在整体的中下区域，其中 h-index 在＜10 区间和 20~40 区间的学者数量最多，h-index＞60 的顶尖学者数量最少，由此可见，知识融合领域学者研究质量差距较大。

受限于本报告篇幅，AMiner 仅选取该领域不同国籍的典型学者做简单介绍，此次排序不分先后。

- Renée J. Miller

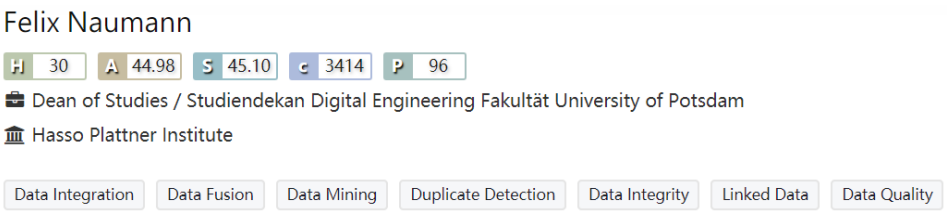


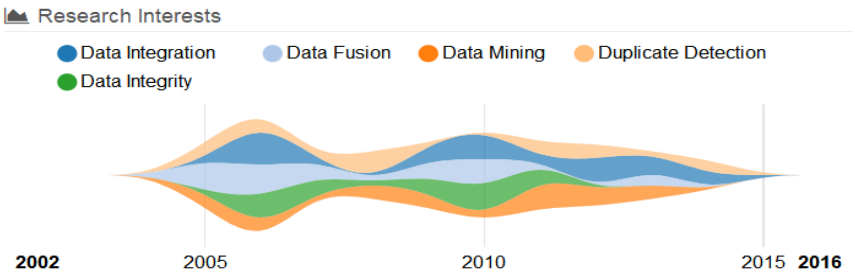
Renée J. Miller，NSERC 商业智能战略网络领导人。多伦多大学计算机科学教授，美国东北大学特聘教授，非营利性国际超大型数据库基金会主席。

Renée J. Miller 自 2000 年涉足知识图谱及相关领域，研究涵盖数据交换、知识融合、数据集成、知识管理和数据共享等方向，曾任 2011 年 ACM SIGMOD 计划主席，在数据整合及完整性领域建树颇多，并因此荣获总统青年研究员奖、2003 年 ICDT 时间测试奖、ACM 会士等奖项与荣誉。

Renée J. Miller 的代表性论文是 2003 年在 ICDT 上发表的“*Data Exchange: Semantics and Query Answering*”给出了一个代数规范，这种规范代表了可能解决方案的整个空间从而使其在数据交换问题的所有解决方案中能够选择通用的特殊解决方案作为问题的解决方法。论文研究了在这种情况下计算某些答案的计算复杂性，并通过在规范的通用解决方案上评估它们来研究计算目标查询的某些答案，分析并解决了数据交换语义相关的基础算法以及在数据交换环境中查询答案的问题，被学术界认可为数据交换的奠基性文章。

- Felix Naumann



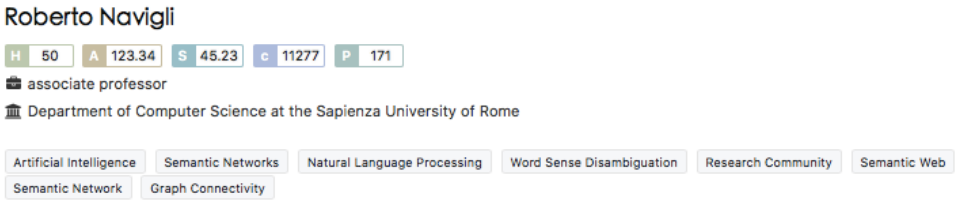


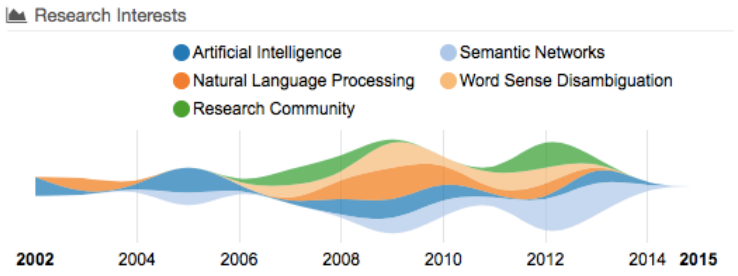
Felix Naumann，是哈索·普拉特钠数字工程研究院（Hasso-Plattner-Institut für Digital Engineering gGmbH,HPI）的教授。隶属于公立波茨坦大学的 HPI，是 IT 领域的德国大学卓越中心。Felix Naumann 是数字工程系主任，信息系统教授，德国计算机科学协会数据库部分发言人。

Felix Naumann 的研究涵盖数据挖掘、数据完整性、知识融合等方向，2005 年前后研究成果增长迅速。曾任 QCRI 访问首席科学家、2012 年 EDBT 演示主席、2017 年 VLDB 行业联合主席、2018 年 VLDB 副主编等职务。多次受邀出席国际顶级学术会议，指导学生获得 2008 年 IEEE 服务杯第一名、2013 年 USEWOD 最佳论文奖、2014 年 PROFILES 和 KnowLOD 研讨会最佳论文奖、2014 年 CIKM 最佳学生论文奖等奖项。因其在领域内的杰出研究，Felix Naumann 荣获 ACM 会士、GI 会士等荣誉。

Felix Naumann 的高引用论文是 2005 年在 ICDE 上发表的“*Schema Matching Using Duplicates*”展示了利用数据集中重复项的存在来自动识别匹配的属性，论文中介绍的算法能够通过比较重复记录中的数据识别相应的属性，经过验证已经证实了该方法的有效性。

- Roberto Navigli



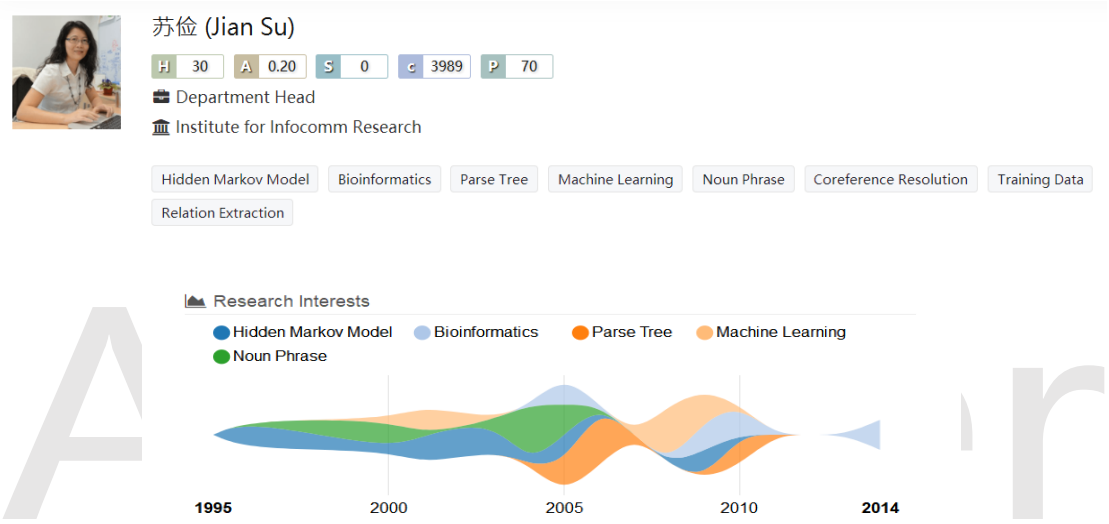


Roberto Navigli 是罗马大学计算机科学系的教授，语言计算实验室的成员，他是 ELEXIS Sapienza 部门的协调员，CINI 人工智能和智能系统国家实验室的指导委员会成员，同时是

欧洲为数不多的获得欧洲研究理事会（ERC）两项奖学金的研究人员之一。

Roberto Navigli是BabelNet的创始人，BabelNet是最大的高质量多语言百科全书计算机辞典。他2013年在ACM上发表的论文的高引论文“*BabelNet: The automatic construction, evaluation and application of a wide-coverage multilingual semantic network*”，提出了构建BabelNet的自动方法，一个覆盖广泛的大型多语言语义网络。该网络通过从WordNet和维基百科中整合词典性与百科式知识，自动构建资源。此外，机器翻译也被用于丰富所有语言的词汇信息资源。我们在新的和现有的标准数据集上进行的实验证明了这一资源的高品质与覆盖范围。

- 苏俭

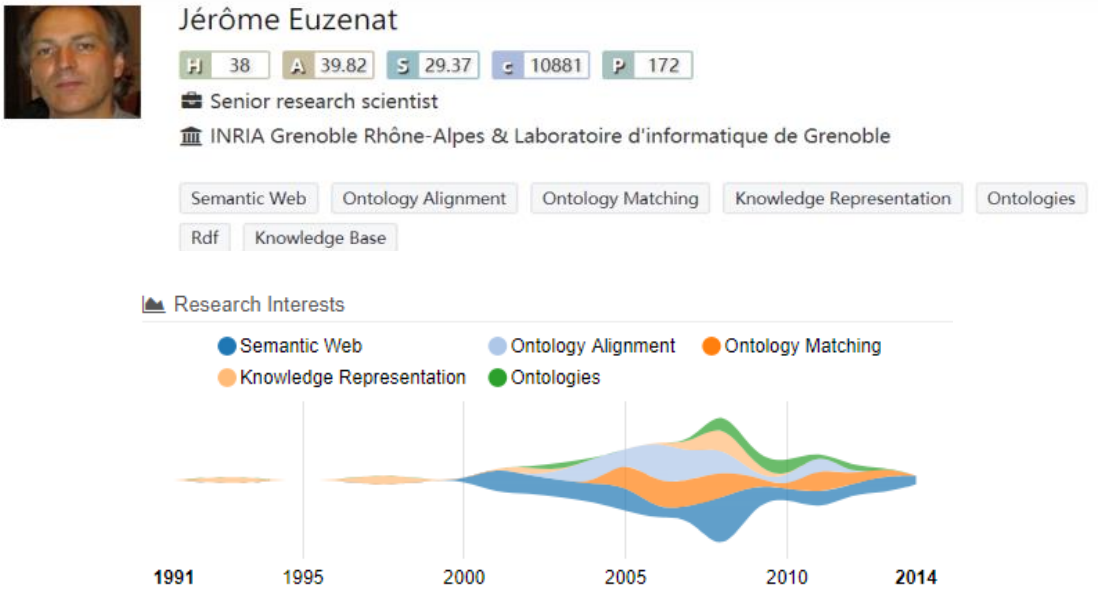


苏俭，大规模技术部署首席专家、BIRC 自然语言处理部门主管、联合主任，SIGDAT 总裁及顾问委员会成员，2018-2020 年 ACL 亚太分会创始执行董事会成员。

苏俭的研究涵盖机器学习、信息提取、情感分析，文本挖掘、机器翻译、自然语言处理等方向，2012 年前后开始专注研究生物信息。曾任 ACL 等国际会议、期刊编委会成员，2015-2016 年 EMNLP 项目主席。2010 年、2011 年、2012 年分别在 COLING、IJCAI 发表 3 篇文章，除此之外，她还获得 2000 年 CONLL 最佳个人系统奖、2004 年 CONLING 最佳表现奖等奖项。

苏俭 2002 年发表在 ACL 上的高引用论文“*Named entity recognition using an HMM-based chunk tagger*”提出了一种隐马尔科夫模型和一种基于该模型的标记器，系统基于上述模型能够有效解决 NER 问题，且性能明显优于任何其他机器学习系统。

- Jérôme Euzenat



Jérôme Euzenat，法国国家计算机科学与控制研究中心（NIRIA）主任研究员，结构化知识共享（EXMO）实验室负责人，欧洲人工智能学会（ECCAI）会士，主要研究领域包括语义知识表示、本体匹配等。

Jérôme Euzenat 曾多次担任国际语义网大会（ISWC）本体匹配国际评测 OAEI 的负责人，他和 Pavel Shvaiko 于 2007 年合著的《Ontology Matching》一书系统性地介绍了本体匹配技术，引用引用高达 3200 余次，已经成为本体语义集成领域的必读书。

2.4. 知识图谱查询和推理计算

2.4.1. 知识推理

知识推理从给定的知识图谱推导出新的实体跟实体之间的关系。知识图谱推理可以分为基于符号的推理和基于统计的推理。在人工智能的研究中，基于符号的推理一般是基于经典逻辑（一阶谓词逻辑或者命题逻辑）或者经典逻辑的变异（比如说缺省逻辑）。基于符号的推理可以从一个已有的知识图谱推理出新的实体间关系，可用于建立新知识或者对知识图谱进行逻辑的冲突检测。基于统计的方法一般指关系机器学习方法，即通过统计规律从知识图谱中学习到新的实体间关系。知识推理在知识计算中具有重要作用，如知识分类、知识校验、知识链接预测与知识补全等。

（1）基于符号的并行知识推理

基于多核、多处理器技术的大规模推理：单机环境下的并行技术以共享内存模型为特点，侧重于提升本体推理的时间效率，适用于对于实时性要求较高的应用场景，这种方法成为首

选。对于表达能力较低的语言，比如 RDFS、OWL EL，单机环境下的并行技术显著地提升了本体推理效率。

基于分布式技术的大规模推理：基于分布式技术可以突破大规模数据的处理界限，这种方法利用多机搭建集群来实现本体推理，很多工作基于 MapReduce 的开源实现设计提出了大规模本体的推理方法，其中较为成功的一个尝试是 Urbani 等人在 2010 年公布的推理系统 WebPIE，在大集群上可以完成上百亿的 RDF 三元组的推理，利用 MapReduce 来实现 OWL EL 本体的推理算法证明 MapReduce 技术同样可以解决大规模的 OWL EL 本体推理并在后续工作中进一步扩展，从而使得推理可以在多个并行计算平台完成。

（2）链接预测

基于表示学习的方法：知识图谱表示学习旨在于将知识图谱中的实体与关系统一映射至低维连续向量空间，以刻画它们的潜在语义特征。通过比较实体与关系在该向量空间中的分布式表示，可以推断出实体和实体之间潜在的关系。

基于图特征的方法：基于图特征的方法借助从知识图谱中抽取出的图特征来预测两个实体间可能存在的不同类型的边（关系）。例如，根据两个实体“姚明”和“叶莉”在知识图谱中的联通路径可以预测出他们之间大概率具备“配偶”关系。

（3）模式归纳方法

基于 ILP 的模式归纳方法：基于 ILP 的方法进行本体学习的早期工作给出了很好的综述。Jens Lehmann 等提出用向下精化算子学习 ALC 的概念定义公理的方法，并在后续工作中将原有方法扩展到处理大规模知识库上。相关的算法都在本体学习工具 DL-Learner 中得到实现，并且在工作中得到进一步扩展，涉及到框架的设计和可扩展性的提升等方面。

基于关联规则挖掘的模式归纳方法：利用谓词偏好因子度量方法以及谓词语义相似度学习相反和对称公理；利用模式层信息给规则的挖掘提供更多的语义；对传统关联规则挖掘技术进行了改进，事务表中用 0 到 1 之间的一个实数代替原来的 0 或者 1，使得提出的方法更符合语义数据开放的特点。

基于机器学习的模式归纳方法：利用聚类的算法学习关系的定义域和值域；应用统计的方法过滤属性的使用，并找出准确、健壮的模式，用于学习属性的数量约束公理。

2.4.2. 知识存储和查询

知识图谱以图（Graph）的方式来展现实体、事件及其之间的关系。知识图谱存储和查询研究如何设计有效的存储模式支持对大规模图数据的有效管理，实现对知识图谱中知识高效查询。

（1）基于关系数据模型的 RDF 数据存储和查询

简单三列表：系统通过维护一张巨大的三元组表来管理 RDF 数据。这张三元组表包含三列，对应存储主体、谓词和客体（或者主体、属性和属性值）。当系统接收到用户输入的 SPARQL 查询时，这些系统将 SPARQL 查询转化为 SQL 查询。然后根据所得 SQL 查询，这些系统通过对三元组表执行多次自连接操作以得到最终解。

水平存储：将知识图谱中的每一个 RDF 主体（subject）表示为数据库表中的一行。表中的列包括该 RDF 数据集合中所有的属性。这种的策略的好处在于设计简单，同时很容易回答面向某单个主体的属性值的查询，即星状查询。存储方法的缺点也是很明显的：其一，表中存在大量的列；其二，表的稀疏性问题；其三，水平存储存在多值性的问题；其四，数据的变化可能带来很大的更新成本。

属性表：为降低自连接操作次数，Jena 和 Oracle 在单张大三元组表之外还支持利用属性表进行 RDF 数据管理。具体而言，Jena 通过聚类的方式将一些类似的三元组聚类到一起，然后将每一个聚类的三元组统一到一张属性表中进行管理，这种方式下的属性表也被称之为聚类属性表；而 Oracle 利用 RDF 资源的类型信息将三元组进行分类，相同类的三元组放到同一张表中，这种方式下的属性表也被称之为分类属性表。

垂直划分策略：SW-Store 提出了对 RDF 数据按照谓词（或属性）分割成若干表的方法。具体而言，SW-Store 将 RDF 三元组按照谓词（或属性）的不同分成不同的表，每张表能保存在谓词（或属性）上相同的三元组。SW-Store 称这种方法为垂直分割。

全索引策略：简单的三列表存储的缺点在于自连接次数较多。为了提高简单三列表存储的查询效率，目前一种普遍被认可的方法是“全索引（exhaustive indexing）”策略。

（2）基于图模型的 RDF 数据存储和查询

RDF 数据的图模型可以最大限度的保持 RDF 数据的语义信息，也有利于对语义信息的查询。在这种情况下，SPARQL 查询就可以视为在 RDF 数据图上进行子图匹配运算。子图匹配运算是图数据库中一个比较经典的问题：其问题定义在于给定一个数据图和一个查询图，找出数据上所有与查询图子图同态的位置。这个问题已被证明是一个 NP 难问题。针对 RDF 数据的 SPARQL 查询已经有一些基于图模型的查询处理系统，如 gStore、和 TurboHOM++。它们都是利用 RDF 数据图的特点来构建索引。

2.4.3. 知识查询与推理人才介绍

选取 knowledge management、knowledge storage、knowledge storing、graph database、triple store、knowledge query、knowledge graph query、knowledge validation、knowledge evaluation、knowledge conflict、knowledge consistency、ontology evaluation、ontology refinement、description

logic、rule extraction、rule learning、knowledge inference、knowledge reasoning、patterning learning、reasoner 等词作为知识查询与推理领域关键词，按照图 2 所示流程将所选学者定义为该领域知名学者并对其进行统计分析，最终绘制出该领域全球知名学者分布图，分别如图 21、图 22 所示：

图 21 知识查询与推理领域全球知名学者分布图

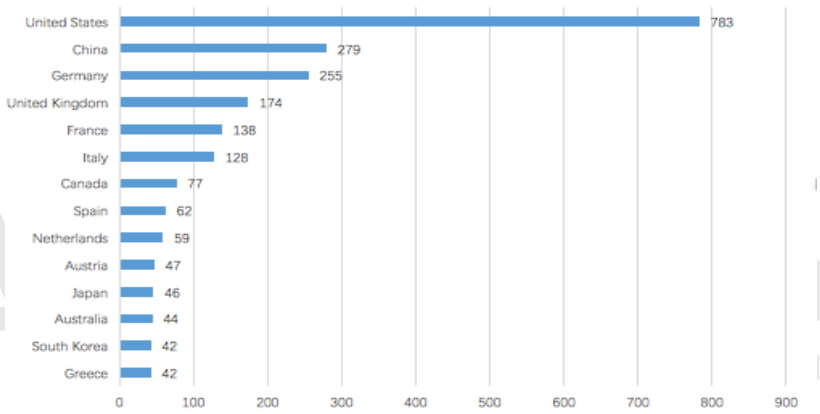


图 22 知识查询与推理领域全球知名学者分布统计

由以上两图可知，全球范围内，符合筛选条件的知识查询与推理领域知名学者集中分布在欧洲及北美洲，亚洲次之，大洋洲、南美洲较为匮乏。若按国家进行统计，美国是该领域学者最为集中的国家，境内学者数量多集中分布在东海岸，德国、中国、英国等国家学者数量次之，其他国家人数较少。

图 23 知识查询与推理领域中国知名学者分布图

对符合上述条件的我国知识查询与推理领域学者分布进行分析，绘制中国范围内知识查询与推理领域知名学者分布图，如图 23 所示。

由图可知，中国知识查询与推理领域知名学者人数较少，境内学者在东北地区、环渤海经济圈、华东地区以及港澳地区均有分布，整体分布较为均匀。

对知识查询与推理领域知名学者进行计算分析，最终绘制出该领域各国人才迁徙图，整体情况图 24 所示：

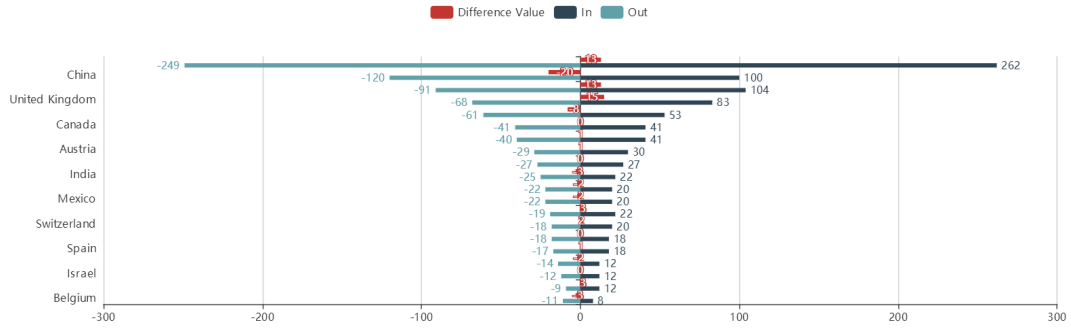


图 24 知识表示与推理领域各国知名学者迁徙图

由上图可知，各国知识查询与推理领域知名学者流失与引进数量差异较为均衡，流动幅度较大的国家分别是美国、中国、德国、英国与意大利等国家。美国是全球该领域学者数量最多的国家，也是该领域知名学者流入量最多的国家，中国的学者流入量略小于流失量，整体呈现出轻微的学者流失迹象。

根据 h-index 对知识查询与推理领域全球知名学者进行分析，最终绘制出学者 h-index 分布图，如图 25 所示：

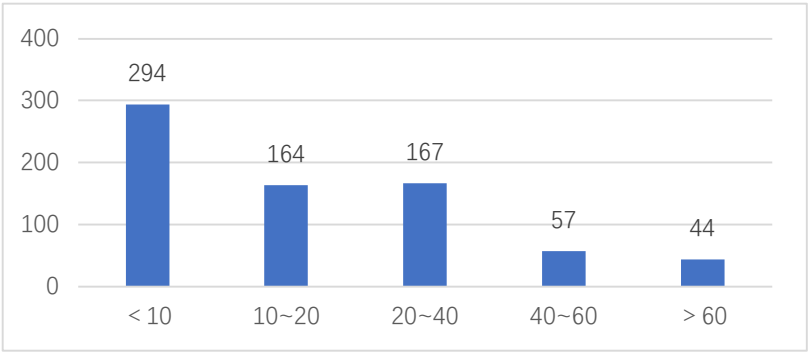


图 25 知识查询与推理领域全球知名学者 h-index 分布图

根据统计信息及上图数据显示可知，知识融合领域学者 h-index 分布呈现金字塔结构，大部分学者 h-index 分布在整体的中下区域，其中 h-index 在＜10 区间的数量最多，位于整体中部的 10~20 区间、20~40 区间学者数量相差不大，h-index＞60 的顶尖学者数量最少，由此可见，知识查询与推理领域顶尖学者与知名学者无论是数量还是研究质量均存在较大差距。

受限于机器挖掘算法、原始数据信息与本报告篇幅，AMiner 仅选取该领域不同国籍的典型学者做简单介绍，此次排序不分先后。

- Frank Wolter

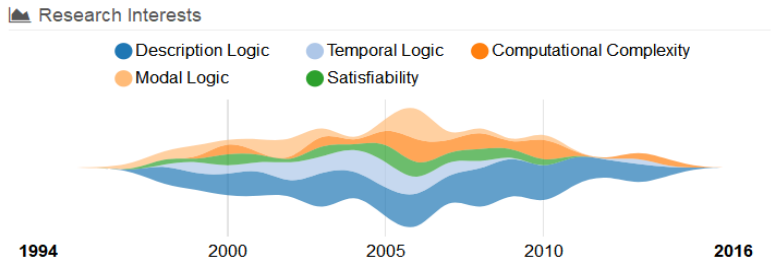
Frank Wolter

H 60 A 54.28 S 30.23 C 10449 P 183

Professor

Department of Computer Science,University of Liverpoo

Description Logic | Temporal Logic | Computational Complexity | Modal Logic | Satisfiability | Decidability | Expressive Power | Conjunctive Query



Frank Wolter，利物浦大学计算机教授，EPSRC 学院成员，描述逻辑研讨会指导委员会成员，AIML 指导委员会成员。

Frank Wolter 的研究涵盖模态逻辑、语义、逻辑推理、人工智能、知识表示与推理等方向，自 1994 年起在知识查询与推理领域的研究从未间断并屡次获奖，其中包括 2000 年、2008 年、2010 年 KR 会议、2013 年 ISWC 会议、2017 年 ACM PODS 会议、2018 年 IJCAI 会议等顶级学术会议最佳论文奖等奖项以及 2018 年 KR 主席、2019 年 AAAI 高级会士等荣誉。

Frank Wolter 的高引用论文是 2003 年发表在 IEEE 上的“*E-connections of abstract description systems*”认为优秀的 AI 应用程序包含了现实世界中的不同方面，因此需要对每一个方面进行建模的可用形式组合，论文提出了一种新的根据抽象描述系统描述逻辑通用概括的组合方法，允许组合之间进行非常规交互，不仅如此，该论文还定义了 E 连接的几种自然变体，并深入研究了从组件系统到其 E 连接的可判定性的转移。

- Diego Calvanese

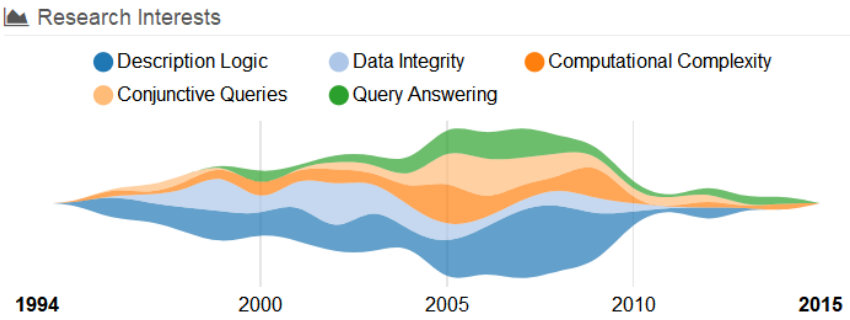
Diego Calvanese

H 74 A 181.74 S 85.65 C 30127 P 431

professor

Free University of Bozen-Bolzano

Description Logic | Data Integrity | Computational Complexity | Conjunctive Queries | Query Answering | Knowledge Base | Satisfiability | Query Language

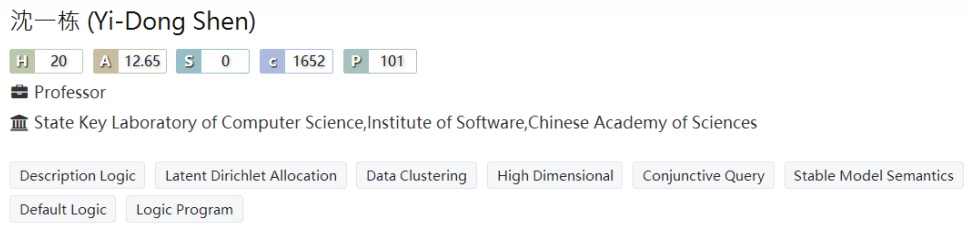


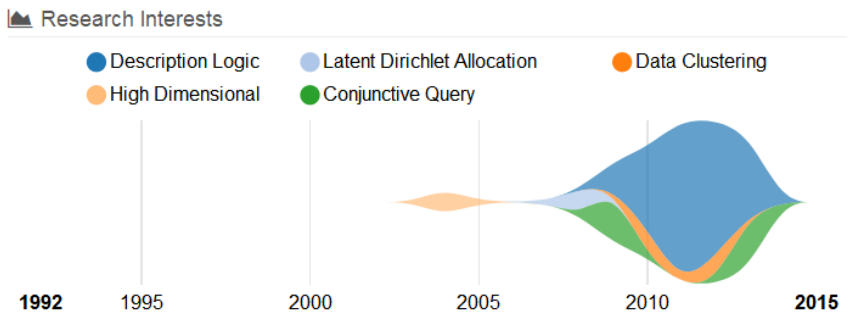
Diego Calvanese，KRDB 知识与数据研究中心教授、意大利波尔扎诺自由大学计算机科学学院副研究员。

Diego Calvanese 的研究涵盖知识表示和推理、本体语言、描述逻辑、概念数据建模、数据集成、图形数据等方向，在 2005 年前后有大量研究成果产出，主要为逻辑描述与数据完整性方向，现阶段负责 Euregio 知识运营支持、SMartDF 等科研项目。曾任维也纳技术大学访问研究员，受邀出席 2015 年 FOFAI 会议、2016 年 DL 会议、2016 年 AMW 会议等学术会议并开展主题演讲，担任 2019 年 AAAI 区域主席等职务并在 IJCAI、KR、AAAI 等会议上发表多篇文章。

Diego Calvanese 的高引用论文是 2007 年发表的“*Tractable Reasoning and Efficient Query Answering in Description Logics: The DL-Lite Family*”提出了一个新的专门用于捕捉本体语言同时保持低推理复杂度的描述逻辑系列 DL-Lite，论文中的推理不仅意味着计算概念之间的包含和检查整个知识库的可满足性，还意味着在 DL 知识库的实例级上回答复杂查询、特别是联合查询的联合，最终结果亦表明 DL-Lite 是支持在大量实例上进行高效查询回答的最大逻辑描述产品。

- 沈一栋

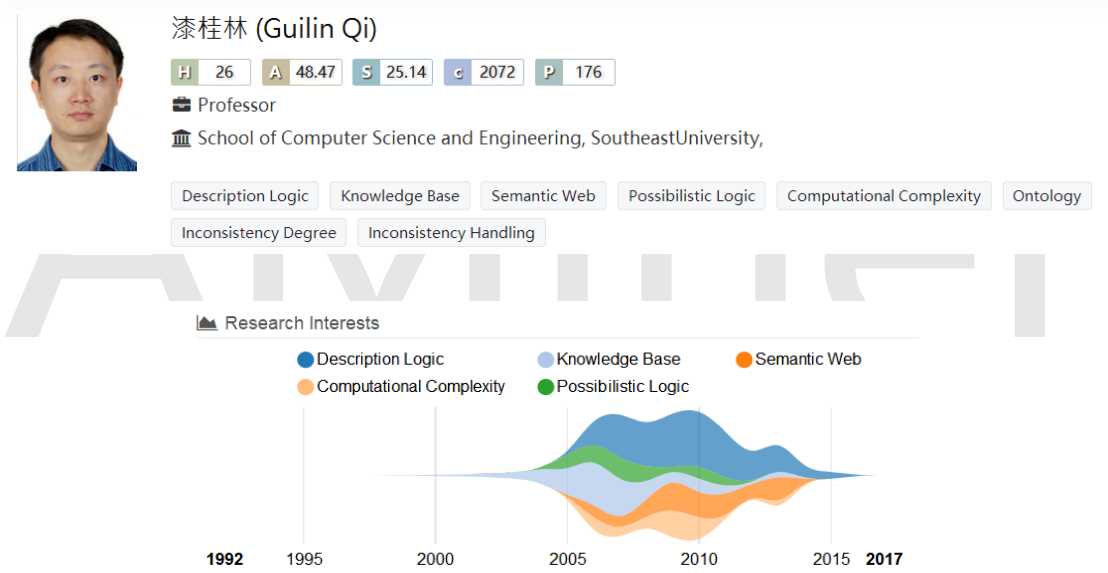




沈一栋，中国科学院软件研究所计算机科学国家重点实验室主任、中国科学院软件研究所研究员、中国科学院大学计算机科学系教授、中国国家人事部百千万人才工程第一、二层次人选、中国教育部跨世纪优秀人才培养计划人选。沈一栋的研究涵盖逻辑描述、逻辑程序设计、数据挖掘、联合查询、知识推理与查询等方向，主持负责国家 973、863 科研计划、国家自然科学基金课题多项，曾担任 AAAI、IJCAI、KDD、KR、ISWC 等重要国际学术会议程序委员会委员且在 AAAI、KDD、ICDM、WWW 等顶级学术会议发表多篇论文。

沈一栋的高引用论文是 2003 年发表在 ICDM 上的“*Mining High Utility Itemsets*”，开发重点挖掘直接支持给定业务目标的 top-K 高实用性封闭模式以解决传统关联挖掘算法只能生成大量高频率规则不能提供有效答案的问题，为了保证关联效率，作者添加了实用程序的概念来捕获非常理想的统计模式并提出逐级项目设置算法，最终的实验结果表明论文中提出的算法在实践中是有效的。

- 漆桂林

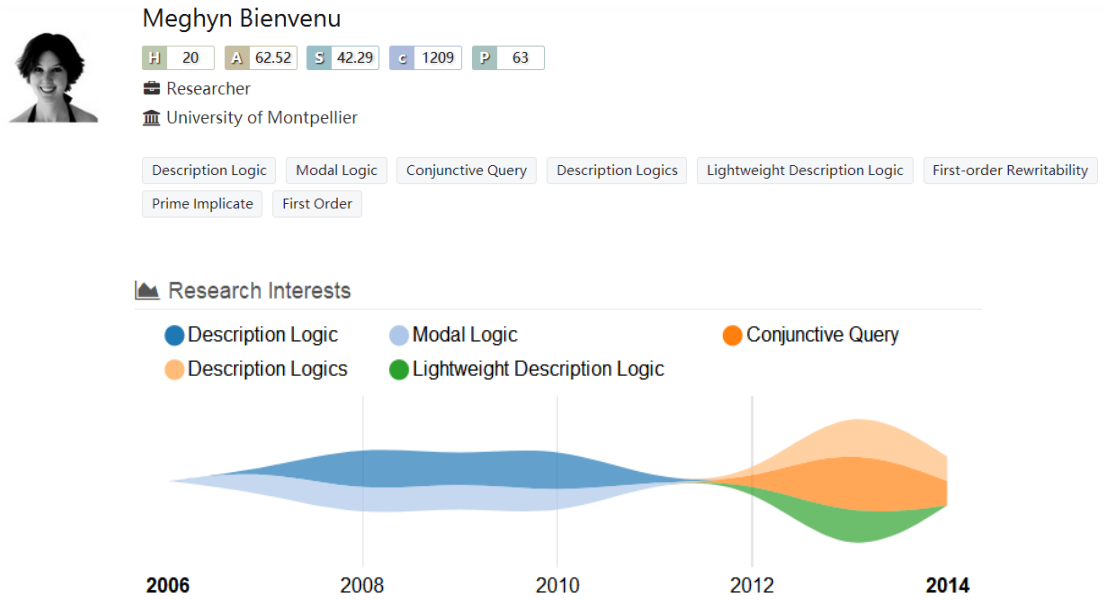


漆桂林，东南大学计算机与科学教授，中国计算机协会会员，Journal of Web Semantics 等学术杂志编委。漆桂林的研究涵盖知识库构建与清理、知识挖掘、语义 Web、深度学习等方向，2005 年至今在知识图谱领域从事长期研究，作为主要参与人参加 863 大数据“类人类智能”方向课题，曾任 2011-2012 年 AAAI 会议、2009-2011 年 ISWC 会议、2009 年 IJCAI 会议、2012 年 KR 会议等顶级学术会议审稿人，Griffith 大学访问教授，法国图卢兹第一大学访问教授，获得亚洲语义 Web 会议最佳论文提名奖并在 IJCAI、AAAI、KR、UAI、ISWC、ESWC 等顶级学术会议上发表多篇文章。

漆桂林的高引用论文是 2008 年发表在 ASWC 上的“*A Modularization-Based Approach to Finding All Justifications for OWL DL Entailments*”提出了一个全新的方法：通过将搜索空间限制为较小的模块来查找 OWL DL 本体中的所有隐藏的最小公理集，最终的实验结果表明

这种方法在OWL DL本体中找到所有隐藏最小公理集的效率和可扩展性提高了几个数量级，具有实用性。

- Meghyn Bienvenu



Meghyn Bienvenu，法国国家科学研究中心研究员、波尔多大学 LaBRI 研究实验室首席科学家，蒙彼利埃大学研究员。Meghyn Bienvenu 的研究涵盖逻辑模型、知识表示和推理、逻辑描述、联合查询等方向，2011 年是其研究方向的分水岭，现阶段的主要研究方向围绕描述逻辑本体及其在查询数据中的应用展开，曾受邀出席 2016 年 IJCAI 会议演讲、担任 2017 年 DL 会议主席、2018 年 IJCAI-ECAI 会议程序委员会联合主席、AAAI 高级 PC 成员等职务并荣获 2009 年 AFIA 论文奖、2016 年 RR 会议最佳论文奖、CNRS 铜奖等荣誉与奖项。

Meghyn Bienvenu 的高引用论文是 2014 年 ACM 上的“*Ontology-based data access: a study through disjunctive datalog, CSP, and MMSNP*”研究了几类本体介导的查询，将数据库查询作为某种形式的联合查询给出并保证本体在描述逻辑或一阶逻辑的其他相关片段中制定，该论文共解决了 3 个问题：用析取数据的碎片来表征本体介导查询的表达能力；在本体介导的查询和约束满足问题（CSP）及其逻辑推广以及利用这些连接获得关于（i）本体介导查询的一阶可重写性和数据记录可重写性的新结果。

2.5. 知识应用

2.5.1. 典型应用

知识应用能够将知识图谱特有的应用形态与领域数据与业务场景相结合并助力领域业务转型。知识图谱的典型应用包括语义搜索、智能问答以及可视化决策支持三种。如何针对

业务需求设计实现知识图谱应用，并基于数据特点进行优化调整，是知识图谱应用的关键研究内容。

（1） 语义搜索

知识图谱是对客观世界认识的形式化表示，将字符串映射为客观事件的事务。当前基于关键词的搜索技术在知识图谱的知识支持下可以上升到基于实体和关系的检索，称之为语义搜索。语义搜索可以利用知识图谱可以准确地捕捉用户搜索意图，进而基于知识图谱中的知识解决传统搜索中遇到的关键字语义多样性及语义消歧的难题，通过实体链接实现知识与文档的混合检索。语义检索需要考虑如何解决自然语言输入带来的表达多样性问题，同时需要解决语言中实体的歧义性问题。同时借助于知识图谱，直接给出满足用户搜索意图的答案，而不是包含关键词的相关网页的链接。

（2） 智能问答

问答系统（Question Answering，QA）是信息服务的一种高级形式，能够让计算机自动回答用户所提出的问题。不同于现有的搜索引擎，问答系统返回用户的不再是基于关键词匹配的相关文档排序，而是精准的自然语言形式的答案。华盛顿大学图灵中心主任 Etzioni 教授 2011 年曾在 Nature 上发表文章“*Search Needs a Shake-Up*”，其中明确指出：“以直接而准确的方式回答用户自然语言提问的自动问答系统将构成下一代搜索引擎的基本形态”。因此，智能问答系统被看作是未来信息服务的颠覆性技术之一，亦被认为是机器具备语言理解能力的主要验证手段之一。

智能问答需要针对用户输入的自然语言进行理解，从知识图谱中或目标数据中给出用户问题的答案，其关键技术及难点包括准确的语义解析、正确理解用户的真实意图、以及对返回答案的评分评定以确定优先级顺序。

（3） 可视化决策支持

可视化决策支持是指通过提供统一的图形接口，结合可视化、推理、检索等，为用户提供信息获取的入口。例如，决策支持可以通过图谱可视化技术对创投图谱中的初创公司发展情况、投资机构投资偏好等信息进行解读，通过节点探索、路径发现、关联探寻等可视化分析技术展示公司的全方位信息，通过知识地图、时序图谱等形态对地理分布、发展趋势等进行解读，为投融资决策提供支持。

可视化决策支持需要考虑的关键问题包括通过可视化方式辅助用户快速发现业务模式、提升可视化组件的交互友好程度、以及大规模图环境下底层算法的效率等。

2.5.2. 通用和领域知识图谱

知识图谱分为通用知识图谱与领域知识图谱两类，两类图谱本质相同，其区别主要体现

在覆盖范围与使用方式上。通用知识图谱主要强调知识的广度，可以形象地看成一个面向通用领域的结构化的百科知识库，其中包含了大量的现实世界中的常识性知识，覆盖面广，通常运用百科数据进行自底向上（Top-Down）的方法进行构建；领域知识图谱又被称为行业知识图谱或垂直知识图谱，可看成是一个面向某一特定领域的基于语义技术的行业知识库，有着严格而丰富的数据模式，应用需求各不相同，因此没有一套通用的标准和规范来指导构建，需要基于特定行业通过工程师与业务专家的不断交互沟通与定制来实现，所以对该领域知识的深度、知识准确性有着更高的要求。

2.5.3. 知识应用人才介绍

选取 knowledge application、knowledge based application、knowledge-based application、knowledge based system、knowledge service、knowledge based service、question answering、decision making、semantic search、knowledge recommendation 等词作为知识应用领域关键词，按照图 2 所示流程将所选学者定义为该领域知名学者并对其进行统计分析，最终绘制出该领域全球知名学者分布图，分别如图 26、图 27 所示：

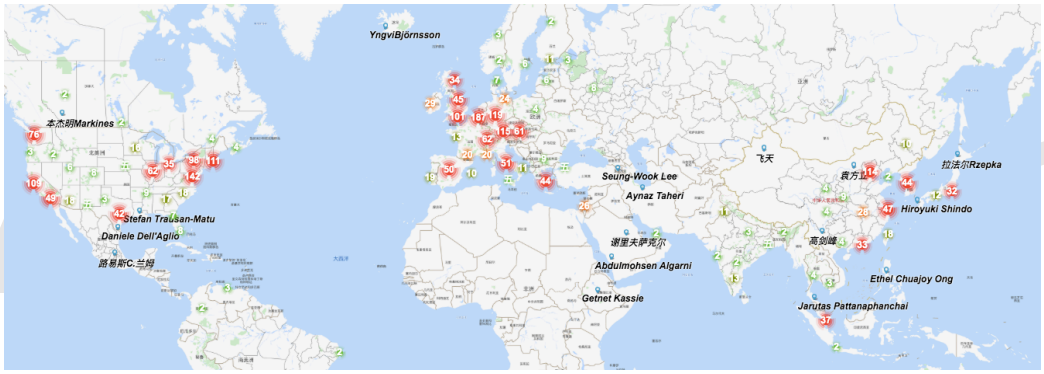


图 26 知识应用领域全球知名学者分布图

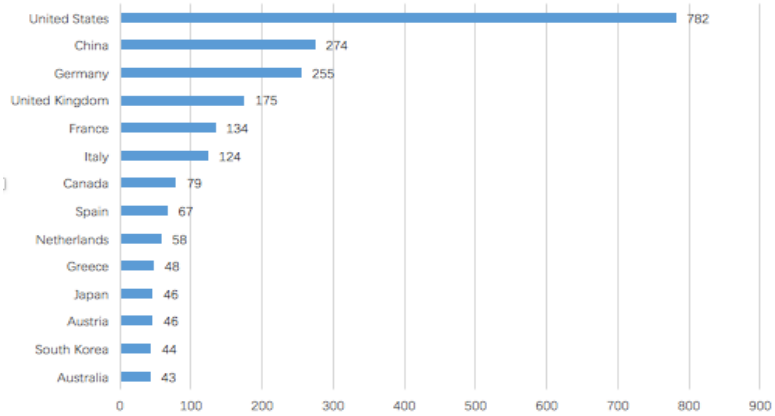


图 27 知识应用领域全球知名学者分布统计

由以上两图可知，全球范围内，符合筛选条件的知识应用领域知名学者集中分布在欧洲及北美洲，亚洲次之，大洋洲、南美洲、非洲较为匮乏。若按国家进行统计，美国是该领域学者最为集中的国家，境内学者在东、西海岸均有分布，中国、德国、意大利、英国等国家

学者数量次之，其他国家人数较少。

对符合上述条件的我国知识应用领域学者分布进行分析，绘制中国范围内知识应用领域知名学者分布图，如图 28 所示：

图 28 知识应用领域中国知名学者分布图

由上图可知，中国知识应用领域知名学者人数较少，境内学者多数集中分布在环渤海经济圈以及东南沿海地区等经济、科研资源相对发达的城市。

对知识应用领域知名学者进行计算分析，最终绘制出该领域各国人才迁徙图，整体情况图 29 所示：

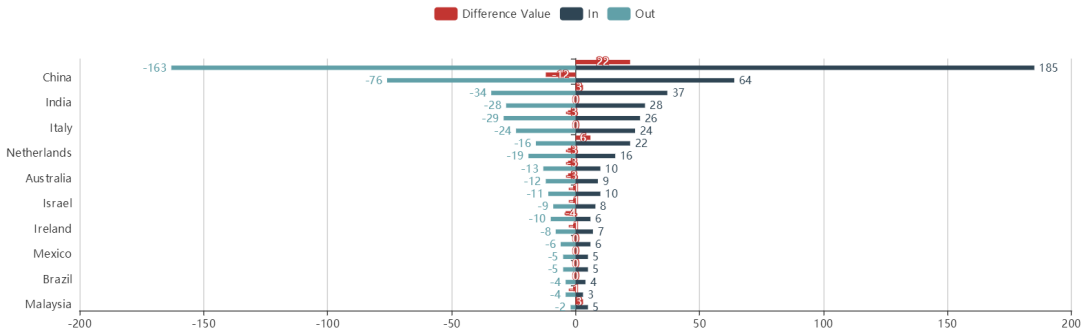


图 29 知识应用领域各国知名学者迁徙图

由上图可知，各国知识应用领域知名学者流失与引进数量差异较为均衡，流动幅度较大的国家分别是美国、中国、德国等国家。美国是全球该领域学者数量最多的国家，也是该领域知名学者流入量最多的国家，中国的学者流入量略小于流失量，整体呈现出学者流失迹象。

根据 h-index 对知识应用领域全球知名学者进行分析，最终绘制出学者 h-index 分布图，如图 30 所示：

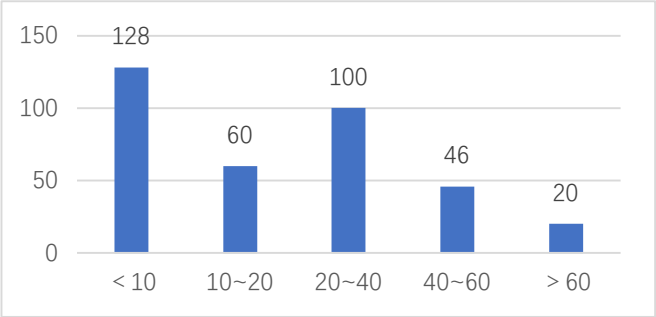
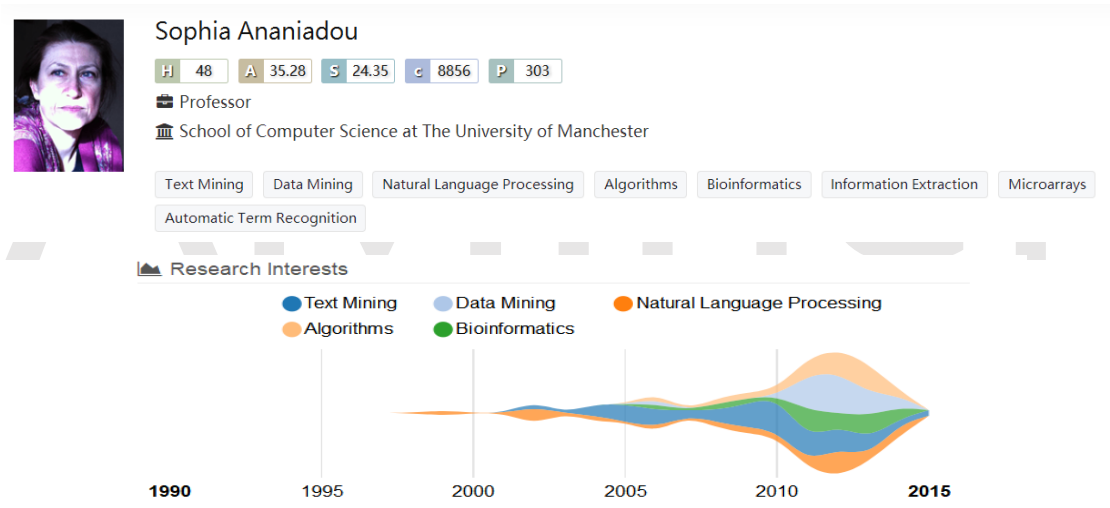


图 30 知识应用领域全球知名学者 h-index 分布图

根据统计信息及上图数据显示可知，知识应用领域学者 h-index 分布并不均衡，大部分学者 h-index 分布在整体的中下区域，其中 h-index 在＜10 区间和 20~40 区间的学者数量最多，h-index＞60 的顶尖学者数量最少，由此可见，知识应用领域学者研究质量差距较大。

受限于机器挖掘算法、原始数据信息与本报告篇幅，AMiner 仅选取该领域不同国籍典型学者做简单介绍，此次排序不分先后。

- Sophia Ananiadou



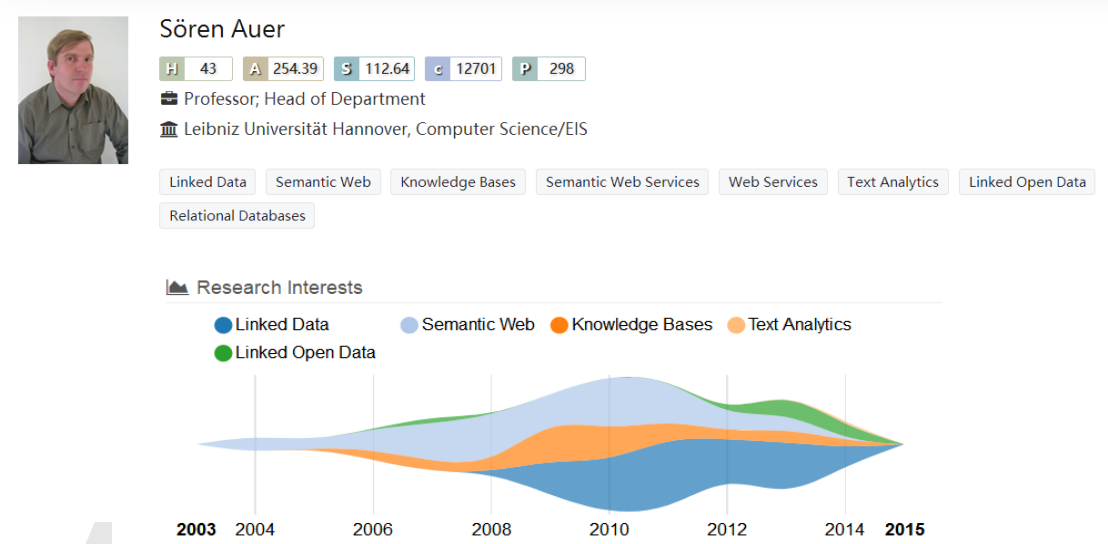
Sophia Ananiadou，英国国家文本挖掘中心（NaCTeM）主任、计算机科学家、曼彻斯特大学计算机教授。

Sophia Ananiadou 的研究涵盖信息提取、文本挖掘、数据挖掘、自然语言处理、生物信息、算法等方向，其中的文本挖掘方向贡献显著，为生物医学领域提供了工具、资源、系统及基础设施。现阶段的研究侧重于提高知识发现速度，上述研究使得 Sophia Ananiadou 发表了 350 余篇同行评审文章，并获得 2004 年大和奖、2006、2007、2008 联续三年 IBM UIMA 创新奖等奖项。

Sophia Ananiadou 的高引用论文是 2005 年 PCL 上的“*Developing a robust part-of-speech tagger for biomedical text*”介绍了一种专门针对生物医学文本进行调整的词性标注器。作者

使用最大熵建模和最先进的标记算法构建了标记器并将其置于包含报纸文章和生物医学文档的语料库上进行训练，以便标记器可以在各类型生物医学文本上顺利工作。最终的实验数据表明标记器的工作精度达到 98%，添加来自不同领域的训练数据也不会影响标记器的性能。

- Sören Auer



Sören Auer，汉诺威大学计算机科学家，德国国家科学技术图书馆-莱布尼兹科技信息中心和汉诺威大学图书馆主任，数据科学与数字图书馆负责人。Sören Auer 的研究涵盖关联数据、知识库、文本分析、语义网络、开放数据等方向，对语义网络、关联数据的研究较为深入。曾任莱比锡大学 AKSW 研究组创建人，波恩大学企业信息系统主席，2010 年 OKCON 会议、ESWC 会议、2011 年 ICWE 会议、2012 年 WWW 会议联合主席且在 2010 年前后在 IJCAI、WWW 等顶级学术期刊发表多篇文章，同时获得 ESWC 最佳论文奖、OpenCourseWare 创新奖等奖项。为语义 Web、知识工程、软件工程可用性以及数据库和信息系统做出了重大贡献。

Sören Auer2007 年发表于 ISWC 的“*DBpedia: a nucleus for a web of open data*”和 2009 年发表的“*DBpedia - A crystallization point for the Web of Data*”介绍了 DBpedia 的工作原理：作为一项社区工作程序，DBpedia 允许用户从维基百科派生的数据集请求复杂查询并在维基百科中提取结构化信息，再使这些信息能够在 Web 上进行访问。经过验证，越来越多的数据发布者开始设置数据级链接到 DBpedia 资源，使得 DBpedia 成为新兴 Web 数据的中心互连中心。目前，围绕 DBpedia 的相互关联的数据源 Web 提供了大约 47 亿条信息，涵盖了地理信息，人员，公司，电影，音乐，基因，药物，书籍和科学出版物等领域。

- 周明

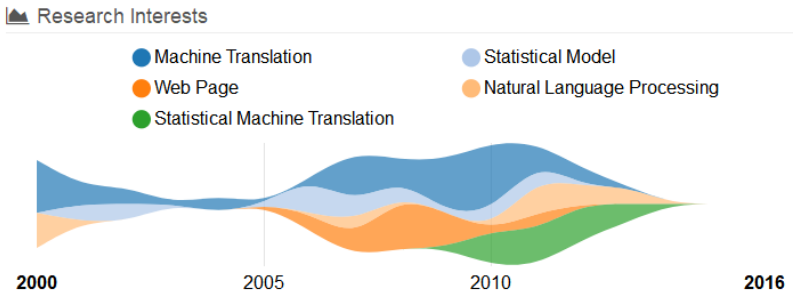
周明 (Ming Zhou)

H 35 A 53.90 S 17.07 C 6052 P 138

微软亚洲研究院副院长;Principal Research Manager

Microsoft Research China

Machine Translation | Web Page | Statistical Model | Statistical Machine Translation | Natural Language Processing | Sentiment Analysis | Semantic Role Labeling | Language Model



周明，中国首个中英翻译系统 CEMT-I 和最知名中日机器翻译产品 J-北京研制者、亚太地区自然语言处理技术推动人物。微软亚洲研究院副院长、ACL 候任主席、中国计算机学会理事、中文信息技术专委会主任、术语工作委员会主任、中国中文信息学会常务理事、哈尔滨工业大学、天津大学、南开大学、山东大学等多所学校博士导师。

周明的研究涵盖机器翻译、知识应用、统计模型、自然语言处理等方向，1999 年加入微软亚洲研究院以来，带领团队进行微软输入法、英库（必应）词典、中英翻译、微软中国文化系列（微软对联、微软字谜、微软绝句）等重要产品和项目的研发，同时优化微软其他产品中的自然语言技术，在机器翻译领域做出杰出贡献，近年来他所领导的研究团队与微软产品组合作开发微软小冰、Rinna、Zo 等聊天机器人系统。前后共发表重要会议与期刊论文 120 余篇，拥有国际专利 40 余项，对推动自然语言处理在中国和亚太地区的卓越发展做出了杰出贡献。

- 赵军

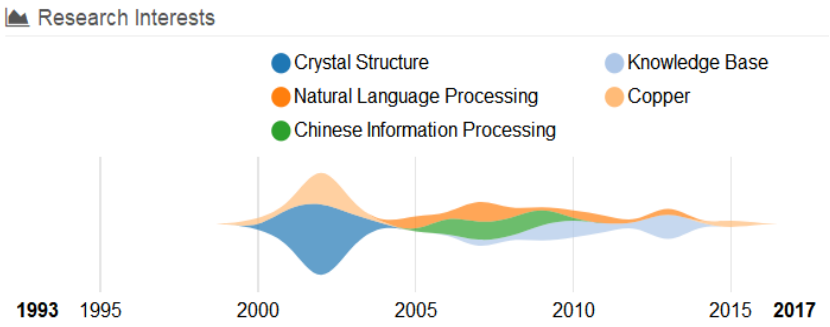
赵 军 (Jun Zhao)

H 36 A 41.26 S 16.27 C 5345 P 293

Professor

Natural Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences

Crystal Structure | Knowledge Base | Natural Language Processing | Computer Application | Copper | Chinese Information Processing | Microrna | Self Assembly



赵军，中国科学院研究员，1998 年至 2002 年香港科技大学博士后，访问学者，2002 年起在中国科学院自动化所模式识别国家重点实验室工作。

赵军的研究涵盖问答系统、信息提取、知识库构建、自然语言处理、中文信息处理等方向，2005 年后在知识库构建领域有持续性研究。主持国家 973 科研计划、国家自然科学基金重点项目多项，在 IEEE TKDE、JMLR 等顶级国际期刊和 ACL、SIGIR、EMNLP、COLING 等顶级国际会议上发表论文 60 余篇，荣获 2011 年 KDD-CUP 亚军、2014 年 COLING 最佳论文奖等多项荣誉。

赵军在获得 COLING2014 Best Paper 的论文“*Relation Classifiction via Convolutional Deep Neural Network*”里提出利用卷积深层神经网络（Convolutional Deep Neural Network）自动学习表征实体关系的词汇特征、上下文特征以及实体所在的句子特征等，相对于以往的分类方法，这个方法不需要利用 NLP 处理工具抽取特征，极大的改善了特征抽取过程中多个处理环节带来的误差累积问题。

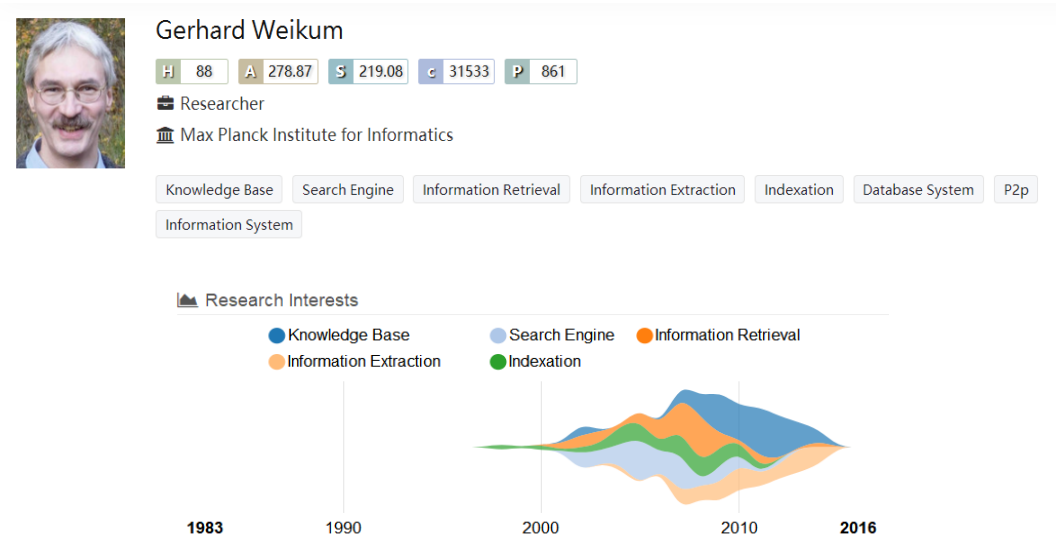
总结本章节所分析的内容，大数据时代的知识图谱已经能够从大数据中自动或半自动获取知识，以结构化的形式描述客观世界中的概念，将互联网信息表达成更接近人类认知世界的方式，技术日趋成熟，智能服务的概念也从单纯的收集获取信息转变为自动化知识服务。就现阶段而言，我国的知识图谱技术已居于世界前列并不断取得进步，但仍需在此基础上利用知识工程为大数据添加语义知识，是数据产生智慧，进一步完成从数据到信息到知识最终到智能应用的转变过程，从而实现对大数据的洞察，为用户提供答案，为决策提供提供支持，改进用户体验等目标。

2.6. 高引学者及论文介绍

2.6.1. 高引学者介绍

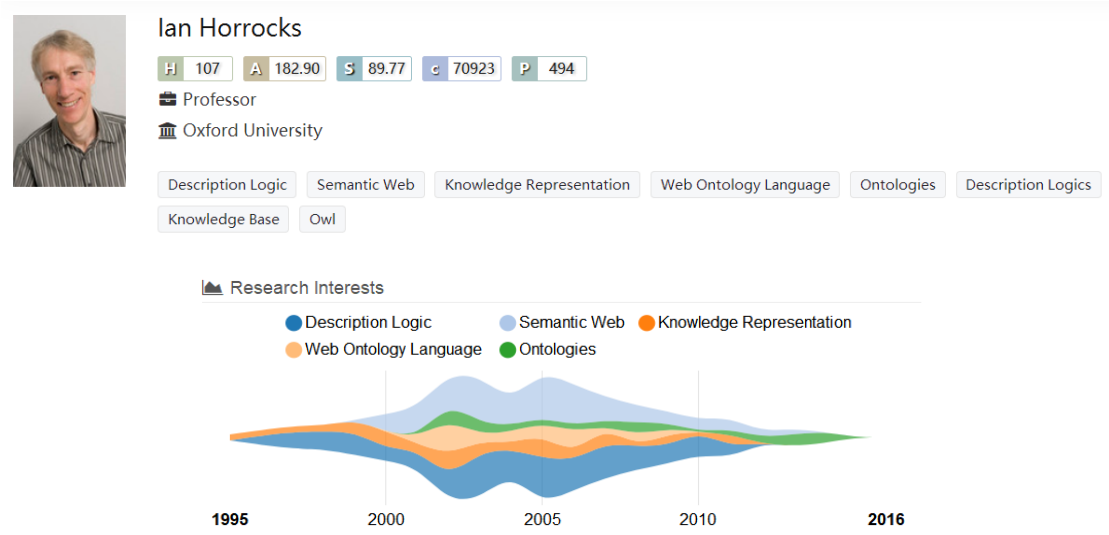
按照本章前 5 节所覆盖的知识图谱领域，挖掘近 10 年在 10 个主要相关会议上发表论文的引用量排名前十的学者如下：

- Gerhard Weikum



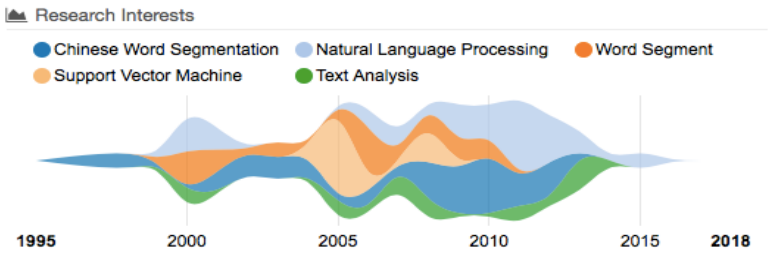
Gerhard Weikum 前文已介绍，不再赘述。

- Ian Horrocks



Ian Horrocks 前文已介绍，不再赘述。

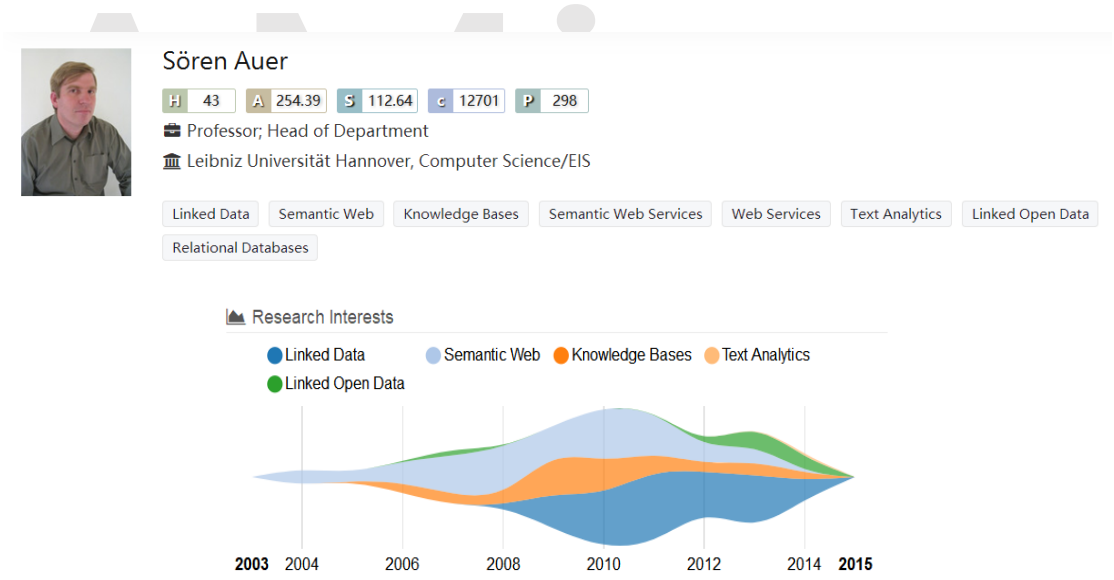
- 孙茂松



孙茂松，清华大学信息科学与技术学院教授，博士生导师。2007-2010 年任该系系主任，现为教育部在线教育研究中心副主任、清华大学计算机系党委书记、清华大学大规模在线开放教育研究中心主任。

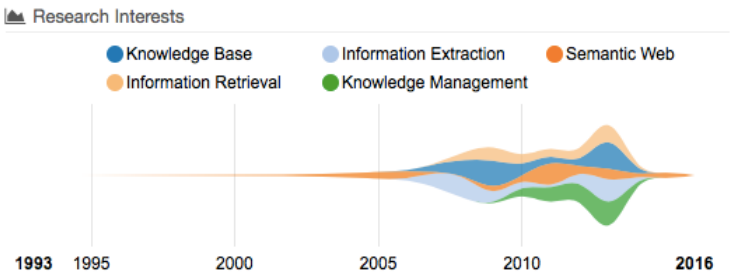
孙茂松的研究涵盖自动图像标注、Web 智能、社会计算、自然语言处理、机器学习等领域。在 ACM、AAAI、ACL 等国内外一流学术期刊和会议上发表论文数十篇。孙茂松的高引论文是 2018 年发表于 AAAI 的“*Neural Knowledge Acquisition via Mutual Attention Between Knowledge Graph and Text*”，提出了一个关于知识获取的通用联合表示框架，用于知识图完成（KGC）和文本关系提取（RE）这两个任务。

- Sören Auer



Sören Auer 前文已介绍，不再赘述。

- Fabian M.Suchanek



Fabian M.Suchanek 是巴黎高等信息学院（Télécom ParisTech）的教授，德国萨尔布吕肯马克斯普朗克信息学研究所 the Otto Hahn“Ontologies”小组的领导者。

Fabian M.Suchanck 获得了 WWW2018 的“Test of Time Award”；提名了 ISWC2018 的“Best research paper”；在 2013 年，他还曾荣获 WWW2013 的“Bset student paper award”。

Fabian M.Suchanck 的研究涵盖了信息提取、知识库、语义网、信息检索等方向。他的高引论文是 2007 年发表于 WWW 的“*Yago：a core of semantic knowledge*”，展示了一种轻量级和可扩展的本体—YAGO。YAGO 以实体和关系为基础，包括 Is-A 层次结构以及实体之间的非分类关系。

- Bernardo Cuenca Grau

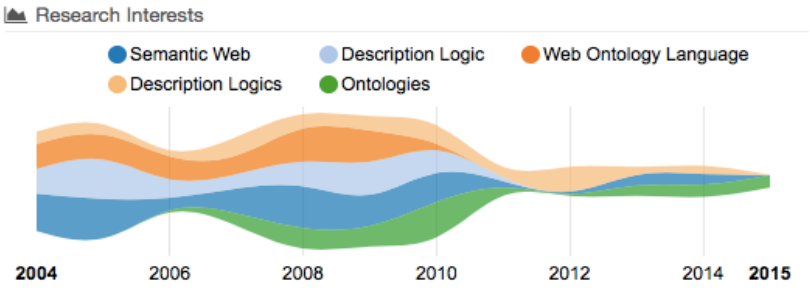
Bernardo Cuenca Grau

H 42 | A 157.95 | S 95.27 | C 11569 | P 197

Professor

University of Oxford

Semantic Web | Description Logic | Web Ontology Language | Description Logics | Ontologies | Real-world Ontology | Owl | Ontology



- Boris Motik

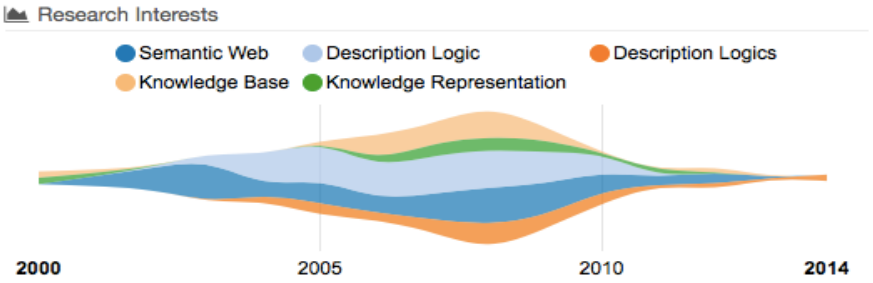
Boris Motik

H 59 | A 127.57 | S 51.31 | C 14941 | P 194

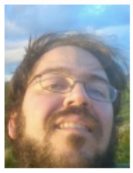
Professor

Computing Laboratory University of Oxford

Semantic Web | Description Logic | Description Logics | Knowledge Base | Knowledge Representation | Owl | Web Ontology Language | Satisfiability



- Bijan Parsia



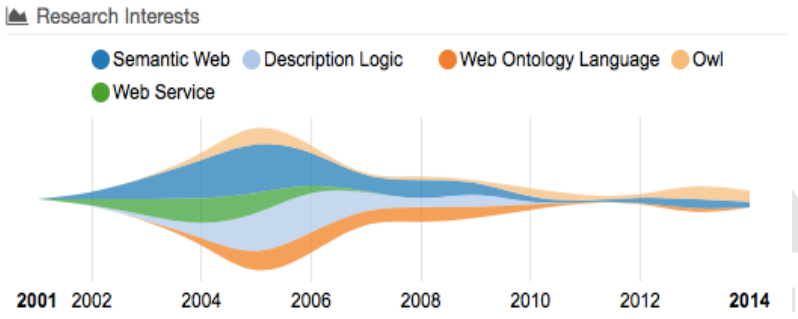
Bijan Parsia

H 50 A 43.43 S 31.68 C 17453 P 238

Reader

University of Manchester (UK)

Semantic Web | Description Logic | Web Ontology Language | Owl | Web Service | Owl Ontology | Ontology | Internet



- Peter Mika

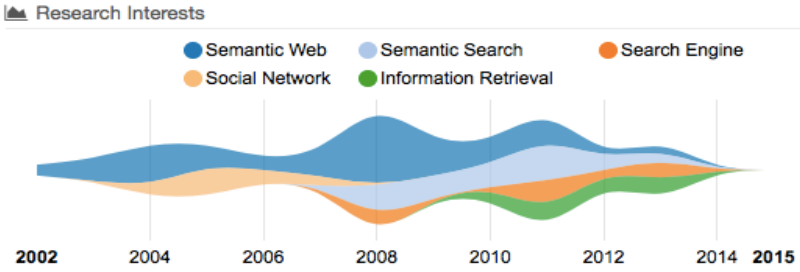
Peter Mika

H 26 A 3.54 S 0 C 4798 P 77

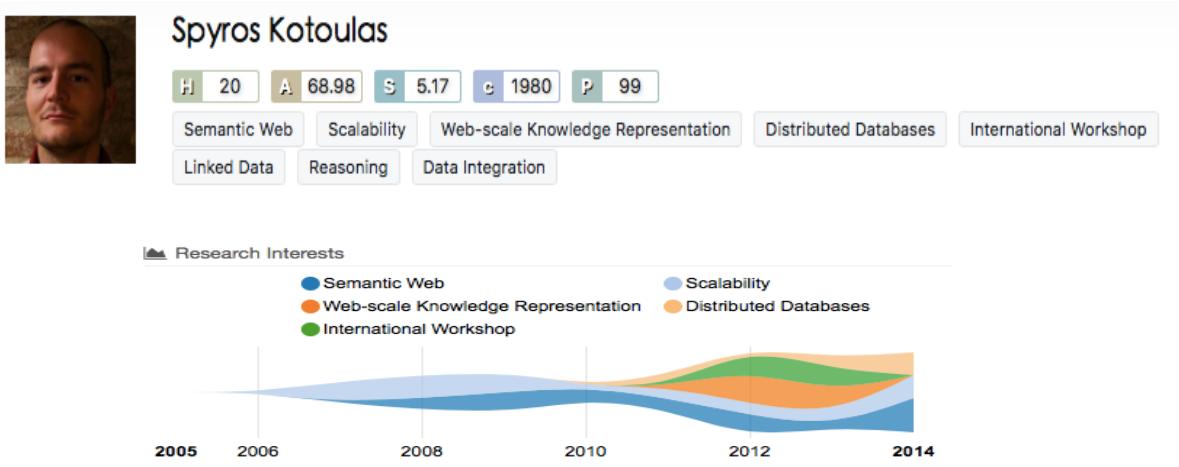
Director

Yahoo

Semantic Web | Semantic Search | Search Engine | Social Network | Information Retrieval | Semantic Technologies | Web Pages | Web Search Engine



- Spyros Kotoulas



2.6.2. 高引论文介绍

有关知识图谱所有论文引用量最高的前十篇论文为：

表 2 知识图谱引用量前十论文

序号	论文题目
1	*Distant supervision for relation extraction without labeled data* Mike Mintz, Steven Bills, Rion Snow, and Dan Jurafsky.ACL/IJCNLP,2009.
2	*You are where you tweet: a content-based approach to geo-locating twitter users* Zhiyuan Cheng, James Caverlee, and Kyumin Lee.CIKM,2010.
3	*YAGO2: a spatially and temporally enhanced knowledge base from wikipedia* Johannes Hoffart, Fabian M. Suchanek, Klaus Berberich, and Gerhard Weikum.IJCAI,2013.
4	*Knowledge vault: a web-scale approach to probabilistic knowledge fusion* Xin Dong 0001, Evgeniy Gabrilovich, Geremy Heitz, Wilko Horn, Ni Lao, Kevin Murphy, Thomas Strohmann, Shaohua Sun, and Wei Zhang.KDD,2014.
5	*Robust disambiguation of named entities in text* Johannes Hoffart, Mohamed Amir Yosef, Ilaria Bordino, Hagen Fürstenau, Manfred Pinkal, Marc Spaniol, Bilyana Taneva, Stefan Thater, and Gerhard Weikum.EMNLP,2011.
6	*BabelNet: building a very large multilingual semantic network* Roberto Navigli, and Simone Paolo Ponzetto.ACL,2010.
7	*Driving with knowledge from the physical world* Jing Yuan, Yu Zheng, Xing Xie, and Guangzhong Sun.KDD,2011.
8	*Open domain event extraction from twitter* Alan Ritter, Mausam, Oren Etzioni, and Sam Clark.KDD,2012.
9	*Sentiment analysis of blogs by combining lexical knowledge with text classification*

	Prem Melville, Wojciech Gryc, and Richard D. Lawrence. KDD,2009.
10	*Open information extraction: the second generation* Etzioni, Oren and Fader, Anthony and Christensen, Janara and Soderland, Stephen and Mausam, Mausam. IJCAI, 2011

2.7. 会议奖项介绍

本报告对人工智能领域 IJCAI、CIKM、ACL、AAAI、ISWC 顶级学术会议最近 3 年（2016-2018）所公布的获奖论文、学者信息进行整合，参考论文主题、学者研究方向与知识图谱领域的契合程度，对所选取的知识图谱领域的获奖论文、学者进行简单介绍。

2.7.1. IJCAI 奖项介绍

通过整理 IJCAI 会议最近 3 年在知识图谱领域获奖论文、学者名单，汇总 IJCAI 知识图谱领域论文、奖项如下：

- 2018 年杰出论文

From Conjunctive Queries to Instance Queries in Ontology-Mediated Querying

从本体论调查查询中的联合查询到实例查询

论文作者：Cristina Feier，Carsten Lutz，Frank Wolter

论文摘要：基于 ALC 的表达描述逻辑和联合查询的本体介导查询（OMQ），研究实例查询（IQ）到本体介导查询（OMQ）中的可重写性。结果包括这种重写的精确特征以及决定可重写性的严格复杂性限制。我们还对决定给定 MMSNP 句子（换句话说：monadic 析取 Datalog 程序的补充）是否等同于约束满足问题的相关问题给出了严格的复杂性约束。

论文地址：https://www.ijcai.org/proceedings/2018/0250.pdf

- 2017 年最佳杰出论文

Foundations of Declarative Data Analysis Using Limit Datalog Programs

使用限制数据记录程序进行声明性数据分析的基础

论文作者：Mark Kaminski，Bernardo Cuenca Grau1，Egor V. Kostylev，Boris Motik，Ian Horrocks

论文摘要：受声明性数据分析应用的启发，我们研究了 DatalogZ，一个带有整数运算功能的实际数据记录（positive Datalog）的扩展。这一语言被认为是不可判定的，因此我们提出了两个分段（fragment）。在 limit DatalogZ 中谓语被公理化以保持最小、最大数值，允许

我们表明事实蕴含（fact entailment）是结合中的完整 ConexpTime 数据复杂性中的完整 Conp。此外，额外的稳定性需求致使复杂性分别降至 ExpTime 和 PTime。最终，我们证明稳定的 DatalogZ 能够表达很多有用的数据分析任务，因此我们的研究成果为高级信息系统的发展打下了坚实的基础。

论文地址：https://www.ijcai.org/proceedings/2017/0156.pdf

- 2017 年卓越论文

BabelNet: The automatic construction, evaluation and application of a wide-coverage multilingual semantic network

BabelNet：广泛覆盖的多语言语义网络的自动构建，评估和应用

论文作者：Roberto Navigli、Simone Paolo Ponzetto

论文摘要：我们在本论文中提出了 BabelNet，一个覆盖广泛的大型多语言语义网络。该网络通过从 WordNet 和维基百科中整合词典性与百科式知识，自动构建资源。此外，机器翻译也被用于丰富所有语言的词汇信息资源。我们在新的和现有的标准数据集上进行的实验证明了这一资源的高品质与覆盖范围。

论文地址：http://www.aclweb.org/anthology/P10-1023

YAGO2：A spatially and temporally enhanced knowledge base from Wikipedia

YAGO2：来自维基百科的空间和时间增强的知识库

论文作者：Gerhard Weikum、Johannes Hoffart、Fabian M. Suchanek、Klaus Berberich

论文摘要：我们提出了 YAGO2——、YAGO 知识库的扩展，该知识库中实体、事实和事件按照时间和空间的顺序排列。YAGO2 从维基百科、GeoNames 和 WordNet 中自动构建而成，涵盖了 980 万实体的 4.47 亿事实。人类评估已经确认其中 95%的事实属实。在本论文中，我们展示了抽取方法、时空维度的整合，以及我们的知识表征 SPOTL（原始的三合一 SPO 模型在时空上的扩展版）。

论文地址：https://www.sciencedirect.com/science/article/pii/S0004370212000719

- 2016 年杰出学生论文

Using Task Features for Zero-Shot Knowledge Transfer in Lifelong Learning

在终身学习中使用任务特征做 Zero-Shot 知识迁移

论文作者：David Isele，Eric Eaton，Mohammad Rostami

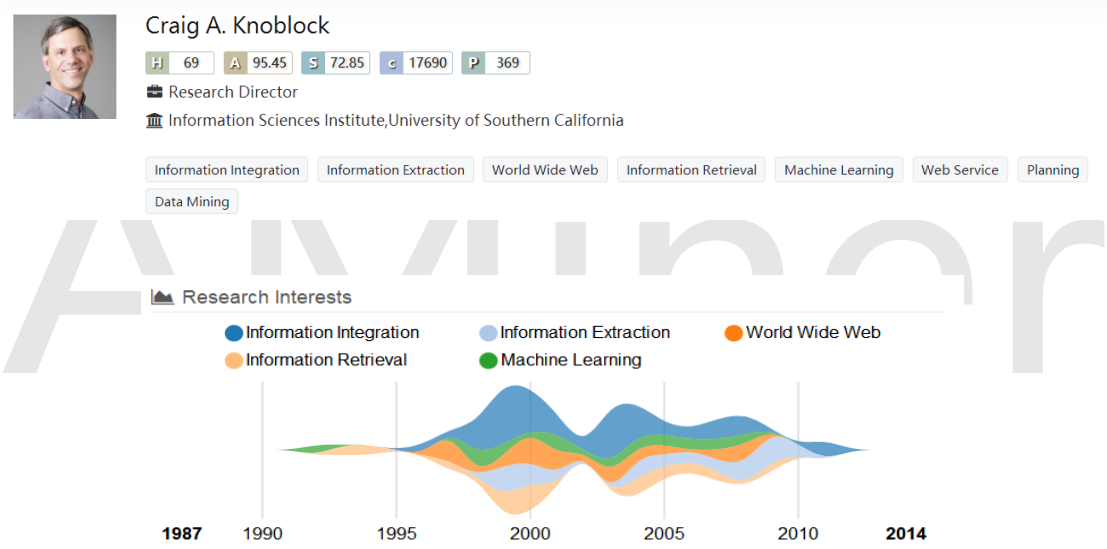
论文摘要：任务间的知识迁移可以改善学习模型的表现，但需要任务间关系的准确评估，

从而识别需要迁移的相关知识。这些任务间的关系一般是基于每个任务的训练数据而进行评估的，它们在以从少量数据中快速学习每个连续任务为目标的终身学习设定中是无效的。为了减轻负担，我们基于耦合词典学习（coupled dictionary learning）开发了一个终身强化学习方法。耦合词典学习将高层任务描述符（descriptors）合并到了任务间关系建模中。我们的结果表明，使用任务描述符能改善学习到的任务策略的性能，既提供了我们方法有益之处的理论证明，又实证展示了在一系列动态控制问题上的进步。在只给描述符一个新任务的情况下，这一终身学习器也能够通过 zero-shot learning 使用耦合词典准确预测任务策略，不再需要在解决任务之前暂停收集训练数据了。

论文地址：https://www.ijcai.org/Proceedings/16/Papers/232.pdf

- 2018 年杰出服务奖

2018 年 IJCAI 杰出服务奖由 Craig A. Knoblock 获得。



Craig A. Knoblock，现任南加州大学信息科学研究所（USC）执行主任，计算机科学与空间科学研究教授，ISI 知识图谱中心的研究主任和信息学计划副主任。AAAI 会士，ACM 会士，IJCAI 前任主席兼受托人。

Craig A. Knoblock 的研究侧重于信息集成、信息抽取、信息检索、描述获取和利用数据语义技术等方向。在源建模、模式和本体对齐、实体和记录链接、数据清理和规范化、从 Web 提取数据以及组合上述技术以构建知识图谱等研究方向开展了深入工作。

2.7.2. CIKM 奖项介绍

通过整理 CIKM 会议最近 3 年知识图谱领域获奖论文、学者名单，汇总 CIKM 知识图谱领域论文、奖项如下：

- 2017 年最佳论文

Hike: A Hybrid Human-Machine Method for Entity Alignment in Large-Scale Knowledge Bases

一种基于人机协作的大型知识图谱对齐方法

论文作者：Yan Zhuang，Guoliang Li，Jianzhong Zhuo，Jianhua Feng

论文摘要：为了实现利用开放的众包平台提高对齐质量这一目标，作者们提出了一种用于大规模 KB 集成的新型混合人机框架。首先根据它们的关系将不同 KB 的实体划分为许多较小的模块，然后在这些分区上构建一个部分顺序并开发一个推理模型，该模型将一组任务集中到人群中，并根据众包任务推断出其他任务的答案。接下来，作者们用公式表示问题选择问题，给定一个货币预算 B，选择 B 众包任务最大化，推断任务的数量。这一研究证明了这个问题是 NP 难度问题，并提出了贪婪算法来解决这个问题，其近似比为 1--1/e。我们在实际数据集上的实验表明，我们的方法提高了质量，并且优于最先进的方法。

论文地址：http://dbgroup.cs.tsinghua.edu.cn/ligl/papers/CIKM2017-Hike.pdf

2.7.3. ACL 奖项介绍

通过整理 ACL 会议最近 3 年知识图谱领域获奖论文、学者名单，汇总 ACL 知识图谱领域论文、奖项如下：

- 2018 年最佳长论文

Learning to Ask Good Questions: Ranking Clarification Questions using Neural Expected Value of Perfect Information

学习如何问好的问题：通过完全信息下的期待值为追问问题排序

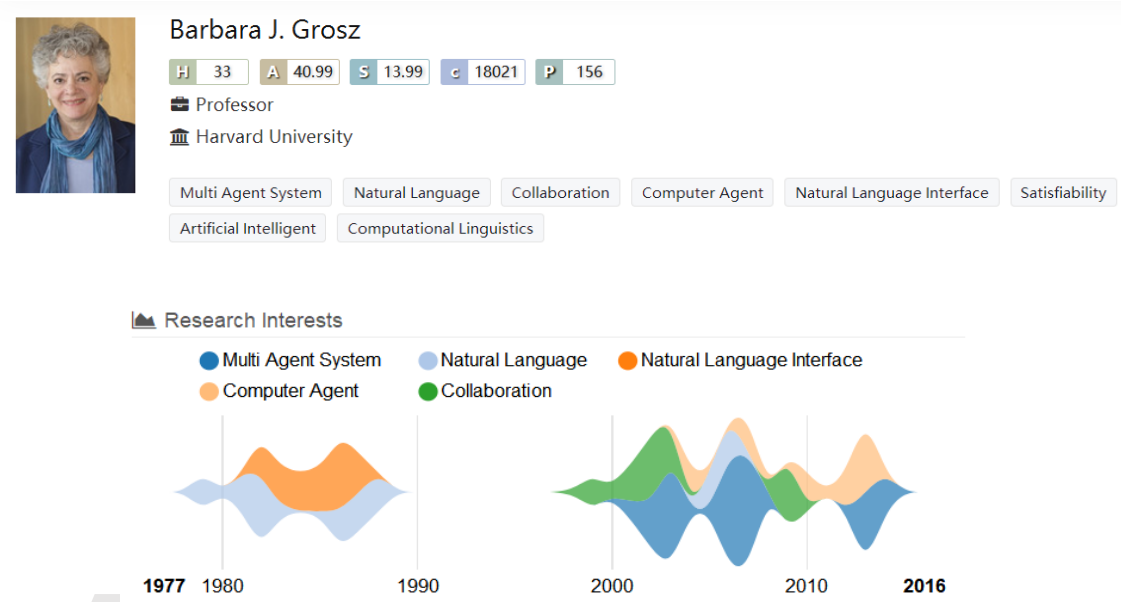
论文作者：Sudha Rao，Hal Daumé III

论文摘要：提问是沟通中的一大基本要素，如果机器不知道如何问问题，那它们也就无法高效地与人类合作。在这项研究中，作者们构建了一个神经网络用于给追问的问题做排名。作者们模型设计的启发来源于完全信息情况下的期待值：一个可以期待获得有用答案的问题就是一个好问题。作者们根据 StackExchange 上抓取的数据研究了这个问题。StackExchange 是一个内容丰富的在线咨询平台，其中有用户发帖咨询以后，别的用户会在下面追问起到解释澄清作用的问题，以便更好地了解状况、帮助到发帖人。论文作者们创建了一个由这样的追问问题组成的数据集，其中包含了 StackExchange 上 askubuntu、unix、superuser 这三个领域的约 7.7 万组发帖、追问问题以及问题的回答。作者们在其中的 500 组样本上评估了自己的模型，相比其他基准模型有显著的提高，不仅如此，他们也与人类专家的判断进行了对比。

论文地址：https://arxiv.org/abs/1805.04655

- 2017 年终身成就奖

2017 年 ACL 终身成就奖由 Barbara J. Grosz 获得。



Barbara J. Grosz，现任哈佛大学希金斯自然科学教授，哈佛大学拉德克利夫高等研究院院长，圣菲研究所科学委员会研究员。美国国家工程院院士、美国哲学学会会员、爱丁堡皇家学会和美国艺术与科学学院成员、AAAI 会士、ACM 会士。

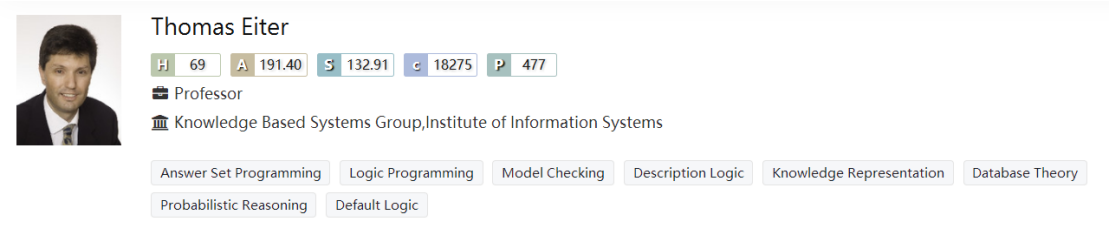
Barbara J. Grosz 的研究侧重于多代理系统、自然语言交互以及协作活动建模等方向。在与知识图谱领域相关的理解人类思维和智慧现象，进而让计算机系统具有智能行为并构建出能够思考和有智慧的系统等研究方向开展了深入工作。

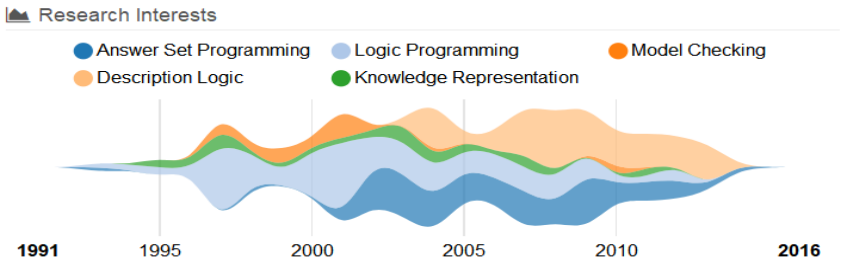
2.7.4. AAAI 奖项介绍

通过整理 AAAI 会议最近 3 年知识图谱领域获奖论文、学者名单，汇总 AAAI 知识图谱领域论文、奖项如下：

- 2017 年杰出高级计划委员会成员奖

2017 年 AAAI 杰出高级计划委员会成员奖由 Thomas Eiter 获得。



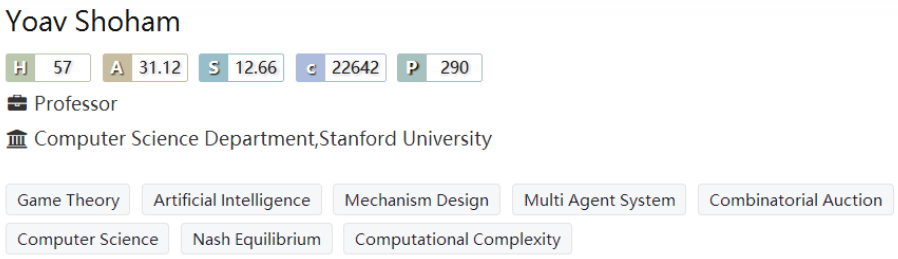


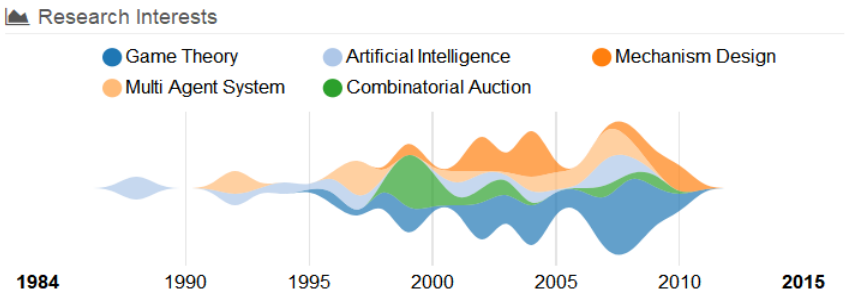
Thomas Eiter，现任奥地利维也纳技术大学计算机科学教授，信息系统研究所所长，基础知识体系集团主管，AIREV（人工智能评论）副主编。欧洲科学院院士，ACM 会士，IEEE 会士。

Thomas Eiter 的研究侧重与于知识表示与推理、计算逻辑、逻辑程序、模型校验、AI 算法及复杂性等方向。主持了语义 Web 的答案集编程、混合知识库中的推理、基于知识的高级信息访问代理等研究课题，在知识表示与推理、智能代理商、非单调逻辑程序设计和数据库等研究方向开展了深入工作。

- 2017 年 FEIGENBAUM 奖

2017 年 AAAI FEIGENBAUM 奖由 Yoav Shoham 获得。

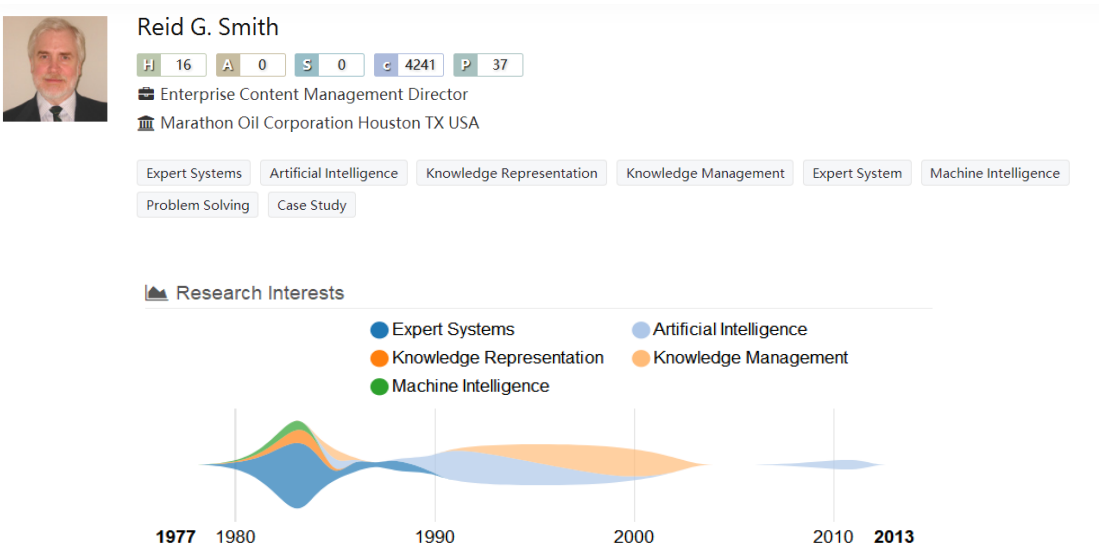




Yoav Shoham，现任斯坦福大学计算机科学教授。AAAI 会士，ACM 会士，GTS 会士 Yoav Shoham 的研究侧重于人工智能、多代理系统、博弈论、机制设计等方向。在知识表示、多智能体系统和计算博弈论等人工智能高影响力基础研究领域开展了深入工作。

- 2016 年恩格尔莫尔奖

2016 年 AAAI 恩格尔莫尔奖由 Reid G. Smith 获得。



Reid G. Smith，现任 i2k Connect 联合创始人兼首席执行官，曾任 Marathon 石油公司企业内容管理总监和 IT 上行业务经理。AAAI 会士、ACM 会士、AAAS 会士、IEEE 会士。

Reid G. Smith 的研究侧重于知识表示、知识管理、人工智能、机器智能、专家系统等方向。社会工作经历丰富，在实际企业管理工作在对知识管理领域做出了开创性贡献，提高了知识管理领域的影响力。

2.7.5. ISWC 奖项介绍

通过整理 ISWC 最近 3 年获奖论文、学者名单，汇总 ISWC 知识图谱领域论文、奖项如下：

ISWC 2017 的具体获奖情况如下：

最佳研究性论文

- 最佳论文奖

A Formal Framework for Comparing Linked Data Fragments

作者：Olaf Hartig, Ian Letter, Orge Pérez

论文摘要：提出链接数据片段（LDF）框架作为统一视图权衡由服务器提供不同接口的来访问关联数据，每个接口都有自己关于性能、宽带需求的特定属性。链接数据碎片机（LDFM）作为经典的图灵机工作，具有模拟服务器和客户端功能的额外功能。

论文地址：https://www.aminer.cn/archive/a-formal-framework-for-comparing-linked-data-

fragments/5a260c2817c44a4ba8a23dac

- 最佳学生论文奖

Improving Visual Relationship Detection Using Semantic Modeling of Scene Descriptions

作者：Stephan Baier, Volker Tresp， Yunpu Ma

论文摘要：图像的结构化场景描述对于大型图像数据库的自动处理和查询是比较有用的，展示了统计语义模型和视觉模型的组合如何改进将图像映射到其相关场景描述的任务，将场景描述表示为三元组，其中每个三元组由一对视觉对象组成，它们出现在图像中，并且展现出它们之间的关系。

论文地址：https://www.aminer.cn/archive/improving-visual-relationship-detection-using-semantic-modeling-of-scene-descriptions/5a260c2e17c44a4ba8a23dda

最佳工业界与政府论文

- 最佳论文奖

Semantic Rule-Based Equipment Diagnostic

作者：Gulnar Mehdi, Evgeny Kharlamov, Ognjen Savkovic, Guohui Xiao, Elem Guzel Kalayci, Sebastian Brandt, Ian Horrocks, Mikhail Roshchin， Thomas Runkler

论文摘要：基于工业规则的诊断系统通常依赖于数据，因为它们依赖于各个设备的特定特性。这种依赖性给规则编写、重用和维护方面带来了重大挑战。因此我们提出了一种语义规则语言 sigRL，同时通过实验来检验了它的可用性和效率。

论文地址：
http://xueshu.baidu.com/usercenter/paper/show?paperid=8a3907ab83a14e4c3a7314575de7d35a&site=xueshu_se

ISWC 2016 的具体获奖情况如下：

最佳研究性论文

- 最佳论文奖

The multiset semantics of SPARQL patterns

作者：Renzo Angles，Claudio Gutierrez

论文摘要：本文确定了 SPARQL 核心模式的多集语义的代数和逻辑结构。

论文地址：https://www.aminer.cn/archive/the-multiset-semantics-of-sparql-patterns/58437725ac44360f1082f7a9

Unsupervised Entity Resolution on Multi-type Graphs

作者：Linhong Zhu, Majid Ghasemi-Gol, Pedro Szekely, Aram Galstyan，Craig A. Knoblock

论文摘要：实体解析是识别在知识库或跨多个知识库中代表同一现实世界实体的所有提及的任务。我们解决了在包含多种类型节点的 RDF 图上执行实体解析的问题，使用不同类型的实例之间的链接来提高准确性。

论文地址：https://www.aminer.cn/archive/unsupervised-entity-resolution-on-multi-type-graphs/58437789ac44360f108435f4

- 最佳学生论文奖

A Probabilistic Model for Time-Aware Entity Recommendation

作者：Lei Zhang, Achim Rettinger, Ji Zhang

论文摘要：近年来，开发用于相关实体推荐的技术的努力越来越多，其中就有给定关键词查询的情况下检索相关实体的排序列表。本文提出第一个概率模型，通过利用从 Web 上公开的不同数据源提取的实体的异构知识，从而将时间意识考虑在实体推荐中。

论文地址：http://xueshu.baidu.com/usercenter/paper/show?paperid=2b236e3f814471dec6ada55d514e6f7a&site=xueshu_se

最佳工业界与政府论文

- 最佳论文奖

Semantic Technologies for Data Analysis in Health Care

作者：Robert Piro, Ian Horrocks, Peter Hendler, Yavor Nenov, Boris Motik, Michael Rossman，Scott Kimberly

论文摘要：美国的 HMO 必须每年向美国当局提供有关其护理质量的测量结果。这些测量集之中有在 HEDIS 的规范中定义，我们应用语义技术来计算 HEDIS 测量中最困难部分的项目，从而得到一个干净、结构清晰、易懂的 HEDIS 编码。

论文地址：http://xueshu.baidu.com/usercenter/paper/show?paperid=a531519a4e8ad22edaa2b383e7401e01&site=xueshu_se&hitarticle=1

ISWC 2015 的具体获奖情况如下：

最佳研究性论文

- **最佳论文奖**

LOD Lab: Experiments at LOD Scale

作者：L. Rietveld, W. Beek, S. Schlobach

论文地址：https://www.aminer.cn/archive/lod-lab-experiments-at-lod-scale/5736984b6e3b12023e713efc

- **最佳学生论文奖**

A Flexible Framework for Understanding the Dynamics of Evolving RDF Datasets

作者：Y. Roussakis, I. Chrysakis, K. Stefanidis, G. Flouris, Y. Starakas

论文摘要：在文中，我们提出了一个框架，可以识别、分析和理解由于 Web 数据的动态演化引起的问题。

论文地址：https://www.aminer.cn/archive/a-flexible-framework-for-understanding-the-dyna-mics-of-evolving-rdf-datasets/5736984b6e3b12023e713e7b

最佳工业界与政府论文

最佳论文奖

Building and Using a Knowledge Graph to Combat Human Trafficking

作者：Pedro Szekely, Craig Knoblock, Jason Slepicka, Chengye Yin, Andrew Philpot, Amandeep Singh, Dipsy Kapoor, Prem Natarajan, Daniel Marcu, Kevin Knight, David Stallard, Steve Minton, Brian Amanatullah, Todd Hughes, Mike Tamayo, David Flynt, Rachel Artiss, Shih-Fu Chang, Tao Chen， Subessware S. Karunamoorthy

论文摘要：论文提出了一种构建知识图谱的方法，它利用语义技术来协调从不同源不断爬行的数据，扩展到从爬行内容中提取的数十亿三元组，并支持对数据的交互式查询。论文将方法应用于打击人口贩运问题，并将其部署到六个执法机构和几个非政府组织，以帮助他们找到贩运者并帮助受害者。

论文地址：https://www.aminer.cn/archive/building-and-using-a-knowledge-graph-to-combat-human-trafficking/5736984b6e3b12023e71402c

3. 应用篇

知识图谱分为通用知识图谱与领域知识图谱两类，两类图谱本质相同，其区别主要体现在覆盖范围与使用方式上。通用知识图谱可以形象地看成一个面向通用领域的结构化的百科知识库，其中包含了大量的现实世界中的常识性知识，覆盖面广。领域知识图谱又叫行业知识图谱或垂直知识图谱，通常面向某一特定领域，可看成是一个基于语义技术的行业知识库，因其基于行业数据构建，有着严格而丰富的数据模式，所以对该领域知识的深度、知识准确性有着更高的要求。

3.1. 通用知识图谱应用

通用知识图谱可以形象地看成一个面向通用领域的“结构化的百科知识库”，其中包含了大量的现实世界中的常识性知识，覆盖面极广。由于现实世界的知识丰富多样且极其庞杂，通用知识图谱主要强调知识的广度，通常运用百科数据进行自底向上（Top-Down）的方法进行构建。表 3 展示的即是常识知识库型知识图谱。

表 3 常识知识库型指示图谱

<table>
<tr><td>类别</td><td colspan="3">名称</td></tr>
<tr><td>通用型</td><td colspan="3">DBpedia，Yago，Freebase，BabelNet，ResearchCyc，WordNet，ConceptNet，KnowItAll，Microsoft Probase，Microsoft ConceptGraph，复旦 GDM（已更名为 CN-DBpedia），XLORE，zhishi. me，知识魔方（SSCO）</td></tr>
<tr><td rowspan="4">垂直型</td><td>金融</td><td colspan="2">海致智能金融知识图谱，财经知识图谱，宜信反欺诈知识图谱</td></tr>
<tr><td>医疗</td><td colspan="2">中西医知识库系统，护理知识图谱，VoxelCloud AI 医学影像知识图谱</td></tr>
<tr><td rowspan="2">企业</td><td>金融</td><td>海致企业知识图谱（涉及用户画像、智能运维等），百度云企业图谱（财经，辅助企业风险监控、营销规划），阿里云企业图谱（展示企业间及个人关系链路），知因智慧知识图谱（提供信贷风险解决方案）</td></tr>
<tr><td>其他</td><td>SCOPA 知识图谱（公安，支持并发、快速查询等），探迹（提供企业咨询报告）</td></tr>
<tr><td>其他</td><td colspan="2">舟山海洋数字图书馆知识图谱，百度教育知识图谱，阿里商品知识图谱</td></tr>
</table>

国外的 DBpedia 使用固定的模式从维基百科中抽取信息实体，当前拥有 127 种语言的超过两千八百万实体以及数亿 RDF 三元组；YAGO 则整合维基百科与 WordNet 的大规模本体，拥有 10 种语言约 459 万个实体，2400 万个事实；Babelnet 则采用将 WordNet 词典与 Wikipedia 百科集成的方法，构建了一个目前最大规模的多语言词典知识库，包含 271 种语言 1400 万同义词组、36.4 万词语关系和 3.8 亿链接关系。

国内的 Zhishi.me 从开放的百科数据中抽取结构化数据，当前已融合了包括百度百科、互动百科、中文维基三大百科的数据，拥有 1000 万个实体数据、一亿两千万个 RDF 三元

组；以通用百科为主线，结合垂直领域的 CN-DBPedia，则从百科类网站的纯文本页面中提取信息，经过滤、融合、推断等操作后形成高质量的结构化数据；XLore 则是基于中文维基百科、英文维基百科、百度百科、互动百科构建的大规模中英文知识平衡知识图谱。

3.2. 领域知识图谱应用

如图 31 所示，领域知识图谱常常用来辅助各种复杂的分析应用或决策支持，在多个领域均有应用，不同领域的构建方案与应用形式则有所不同，本文将以电子商务、图书情报、企业商业、船业投资、生物医疗五个领域为例，从图谱构建与知识应用两个方面介绍领域知识图谱的技术构建应用与研究现状。

图 31 行业知识图谱应用

3.2.1. 电子商务

当下，电商的交易规模巨大，对我们每个人的生活都有影响。因而电商知识图谱这个垂直图谱变得非常重要。

如图 32 所示，电商知识图谱以商品为核心，以人、货、场为主要框架。目前共涉及 9 大类一级本体和 27 大类二级本体。一级本体分别为：人、货、场、百科知识、行业竞对、品质、类目、资质和舆情。人、货、场构成了商品信息流通的闭环，其他本体主要给予商品更丰富的信息描述。下图描述了商品知识图谱的数据模型，数据来源包含国内-国外数据，商业-国家数据，线上-线下等多源数据。目前有百亿级的节点和百亿级的关系边。

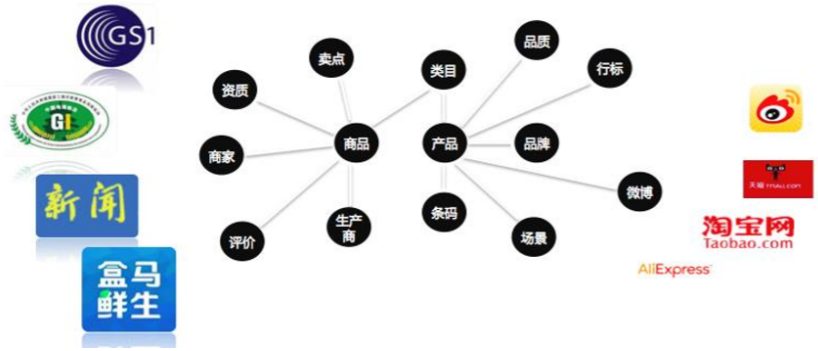


图 32 电商图谱 Schema

电商知识图谱，这个商品‘大脑’的一个应用场景就是导购。而所谓导购，就是让消费者更容易找到他想要的东西，比如说买家输入“我需要一件漂亮的真丝丝巾”，“商品大脑”会通过语法词法分析来提取语义要点“一”、“漂亮”、“真丝”、“丝巾”这些关键词，从而帮买家搜索到合适的商品。在导购中为让发现更简单，“商品大脑”还学习了大量的行业规范与国家标准，比如说全棉、低糖、低嘌呤等。此外，它还有与时俱进的优点。“商品大脑”可以从公共媒体、专业社区的信息中识别出近期热词，跟踪热点词的变化，由运营确认是否成为热点词，这也是为什么买家在输入斩男色、禁忌之吻、流苏风等热词后，出现了自己想要的商品。最后，智能的“商品大脑”还能通过实时学习构建出场景。比如输入“海边玩买什么”，结果就会出现泳衣、游泳圈、防晒霜、沙滩裙等商品。再者，电商平台管控从过去的“巡检”模式升级为发布端实时逐一检查。在海量的商品发布量的挑战下，最大可能地借助大数据、人工智能阻止坏人、问题商品进入电商生态。为了最大限度地保护知识产权，保护消费者权益，电商知识图谱推理引擎技术满足了智能化、自学习、毫秒级响应、可解释等更高地技术要求。实现了良好的社会效益。例如：上下位和等价推理，检索父类时，通过上下位推理把子类的对象召回，同时利用等价推理（实体的同义词、变异词、同款模型等），扩大召回。以为保护消费者我们需要拦截“产地为某核污染区域的食品”为例，推理引擎翻译为“找到产地为该区域，且属性项与‘产地’同义，属性值是该区域下位实体的食品，以及与命中的食品是同款的食品”。

3.2.2. 图书情报

图情知识图谱是指聚焦某一特定细分行业，以整合行业内资源为目标的知识图谱。提供知识搜索、知识标引、决策支持等形态的知识应用，服务于行业内的从业人员，科研机构及行业决策者。

图情领域与知识图谱的结合由来已久。以图 33 中大英博物院语义搜索为例，英国的大英博物馆通过结合语义技术对馆藏品各类数据资源进行语义组织，通过语义细化、多媒体资源标注等方式提供多样化的知识服务形式；英国广播公司 BBC 在其音乐、体育野生动物等板块定义了知识本体，将新闻转化为机器可读的信息源（RDF/XML，JSON 和 XML）进行

内容管理与报道自动生成。国内图情领域也越来越重视对知识图谱技术的利用。上海图书馆借鉴美国国会书目框架 BIBFRAME 对家谱、名人、手稿等资源构建知识体系，打造家谱服务平台为研究者们提供古籍循证服务；中国农科院则聚焦于水稻细分领域，整合论文、专利、新闻等行业资源，构建水稻知识图谱，为科研工作者提供了行业专业知识服务平台。

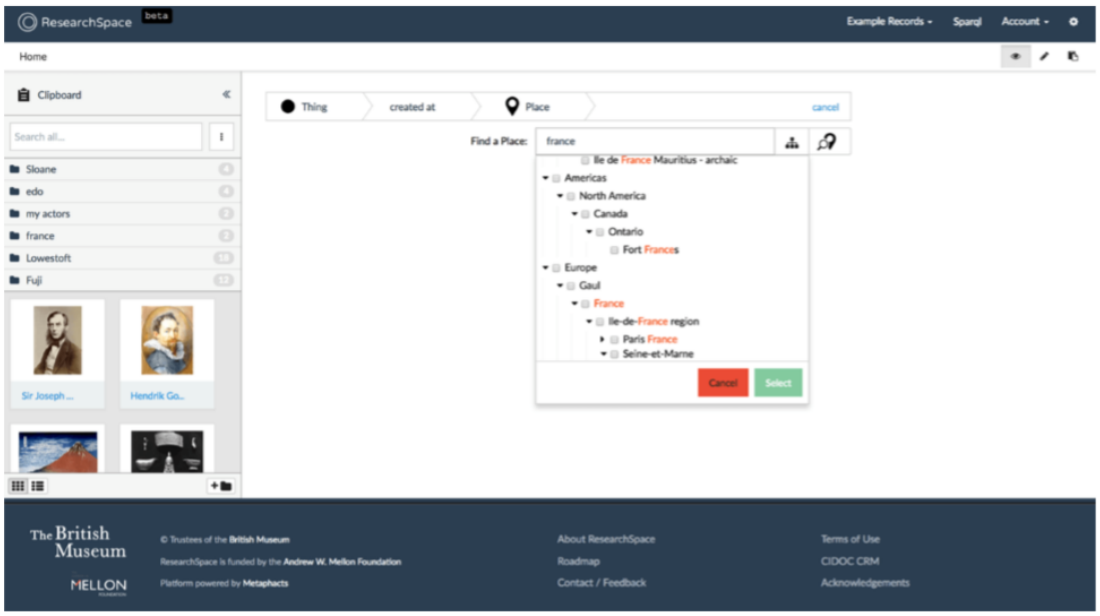


图 33 大英博物院语义搜索

3.2.3. 企业商业

全国企业知识图谱通过异常关联挖掘、企业风险评估、关联探索、最终控制人和战略发展等方式为行业客户提供智能服务和风险管理。

异常关联挖掘是通过路径分析、关联探索等操作，挖掘目标企业谱系中的异常关联。基于企业商业知识图谱从多维度构建数据模型可以进行全方位的企业风险评估，有效规避潜在的经营风险与资金风险，如图 34 所示：

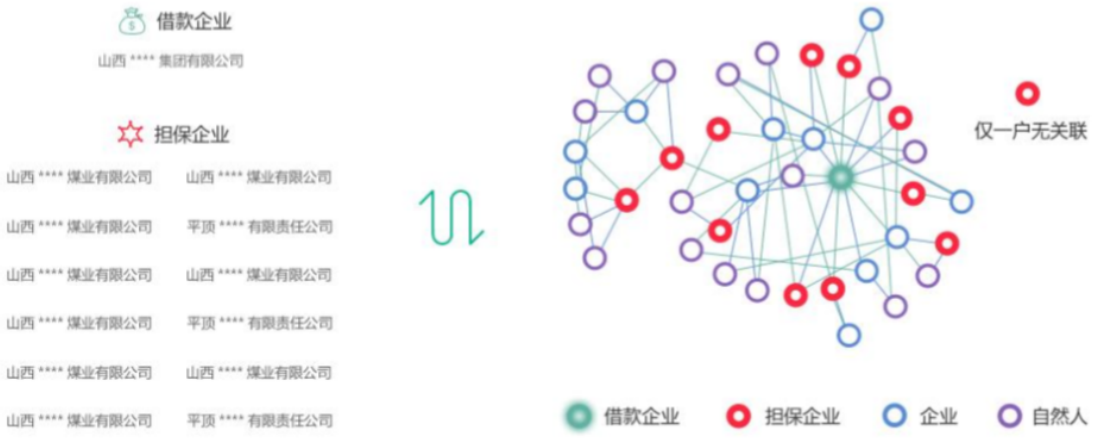


图 34 异常关联挖掘

最终控制人是基于股权投资关系寻找持股比例最大的股东，最终追溯至自然人或国有资产管理部门，如图 35 所示：

图 35 最终控制人分析

战略发展则以“信任圈”的形式将目标企业的对外投资企业从股权上加以区分，探寻其全资、控股、合营、参股的股权结构及发展战略，从而理解竞争对手和行业企业的真实战略，发现投资行业结构、区域结构、风险结构、年龄结构等，如图 36 所示：

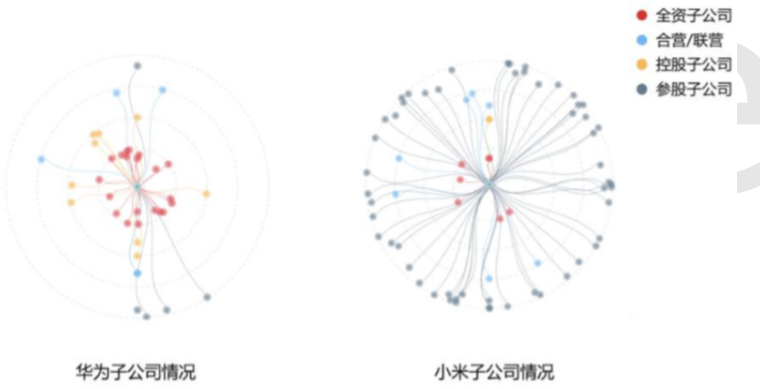


图 36 企业社交图谱

3.2.4. 创业投资

创业投资知识图谱聚焦于工商知识图谱的一部分数据内，创投领域知识图谱主要应用形态包括知识检索以及可视化决策支持。

知识检索依托创投知识图谱可以在原有知识全文搜索的基础上实现语义搜索与智能问答的应用形态。其中，语义搜索提供自然语言式的搜索方式，由机器完成用户搜索意图识别。而作为知识搜索的终极形态，智能问答允许用户通过对话的方式对领域内知识进行问答交互，同时通过配置问题模板实现复杂业务问题的回答。如图 37 所示：

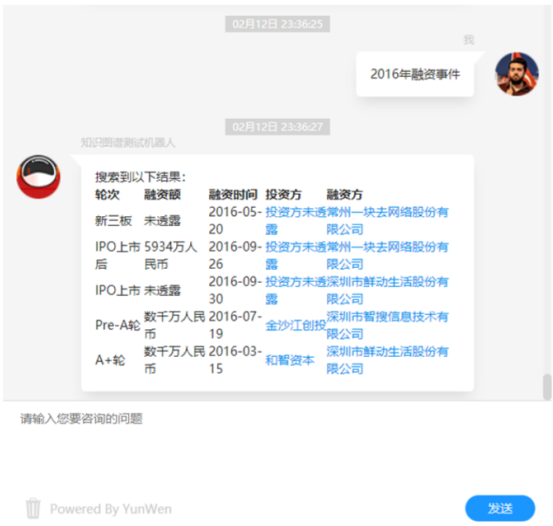


图 37 智能问答

3.2.5. 生物医疗

随着技术的不断进步，采用理论研究与实证分析、应用研究相结合的方法，在收集大量资料与数据、阅读文献的基础上梳理和总结经典的医学管理与决策理论以及大数据管理与分析方法的医疗知识图谱已经实现。整体技术路线如图 38 所示，在此基础上开展应用研究，研发系统对理论成果进行验证，根据评测标准对应用效果进行测评。总体技术路线为建立知识表示模型、构建医学知识图谱、提供医学知识服务、研发知识服务系统，具有较强的可行性和创新性。

首先基于资源描述框架网络本体语言建立医学知识表示模型，包括医学体分类体系以及建模实体不确定性关联；然后从电子病历、临床指南和医学主题词表等多源异构医学大数据中抽取医学信息，采用条件随机场模型实体、朴素贝叶斯模型抽取实体关系，关联规则挖掘方法抽取实体属性。提出实体链接方法和基于图的重启随机游走方法进行知识融合，进一步提高知识质量，构建医学知识图谱。

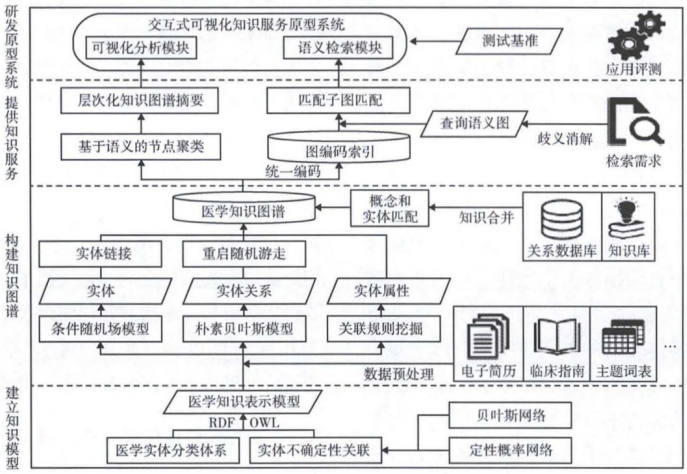


图 38 生物医疗

4. 趋势篇

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更多事实性知识。然而，人类知识类型丰富多样，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了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，并已经成功抽取出大量有用的知识。但在抽取知识的准确率、准确率和效率方面，都仍不尽如人意，有极大的提升空间。

- **知识融合**

从不同来源数据中抽取的知识可能存在大量噪声和冗余，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。

整体而言，知识图谱领域的发展将会呈现以下趋势：

- 特色化

构建大规模知识图谱多基于 Web 信息、知识库:国外以 Web 开放信息为主、结构化知识库为辅快速构建大规模、跨领域知识图谱，如 Google 基于 Web 开放资源、知识库（维基百科、Freebase 等）采集信息并构建知识图谱;国内早期采用该类方法构建并通过增加中文特性扩充语义范畴、满足用户需求。但限于中英文信息处理差异性，当前中文知识图谱构建多基于中文知识百科整合 Web 开放信息构建特色垂直型中文百科知识图谱及其应用。

- 开放化

大规模知识图谱多依赖开放域数据（（半）结构化数据）抽取知识（如 Freebase，CN-DBPedia）并基于 Web 传播但当前开放度较低（尤其是商用知识图谱），不利于知识图谱构建、垂直应用落地，与其开放、互联初衷相悖。新近出现的开放知识图谱社区（Open KG）制定协议规范（遵循商业规则、知识产权、数据开放许可协议等），通过开源软件方式在保障各方权益前提下开放知识图谱以实现整体利益最大化:基于关联数据技术实现多知识图谱关联，基于知识图谱链接封闭域数据与开放域数据（有效弥补封闭域数据知识不完全缺陷），提供 API 方便用户访问，以发现、共享知识并增加其价值。

- 智能化

为更好发挥现有知识图谱知识表达、知识资源优势，需与其他技术（信息推荐、事理图谱、机器学习、深度学习等）融合以提升应用智能性:中文知识图谱个性化推荐系统利用大规模知识图谱中概念、实体间超链关系度量任意词条间语义关联并结合显式语义分析模型实现用户与项目（用两组标签分别描述）间精准推荐;表示事件逻辑关系的事理图谱辅助知识图谱定位、拓展事态进程并可用于智能推荐、常识推理等。工业界基于大数据、知识图谱、人工智能、机器学习等技术构建机器智脑，通过知识规则或深度学习模型积累知识、经验以模拟、抽象人类智慧，提升商业应用可行性及机器智能性。

基于分析，AMiner 数据平台绘制了知识图谱领域近期与全局热点词汇，分别如图 39、图 40 所示：

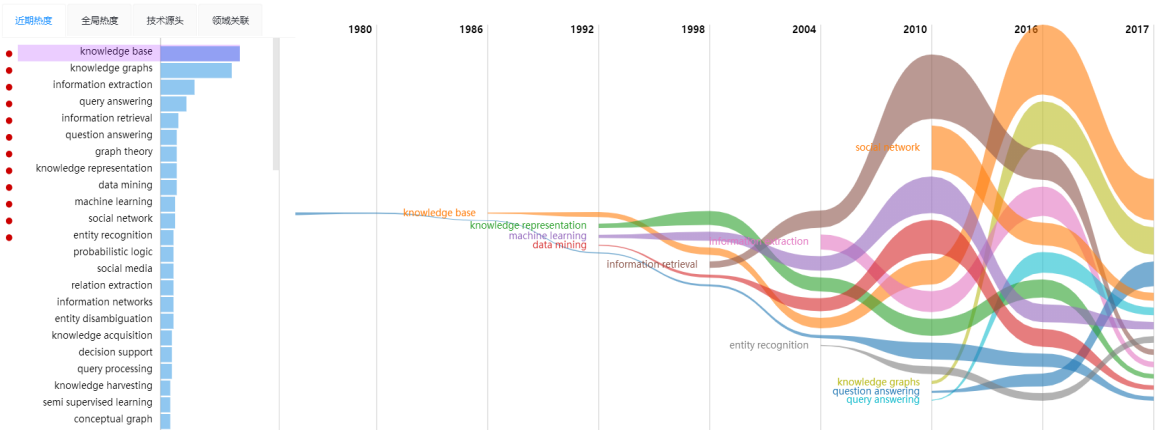


图 39 知识图谱领域近期热度

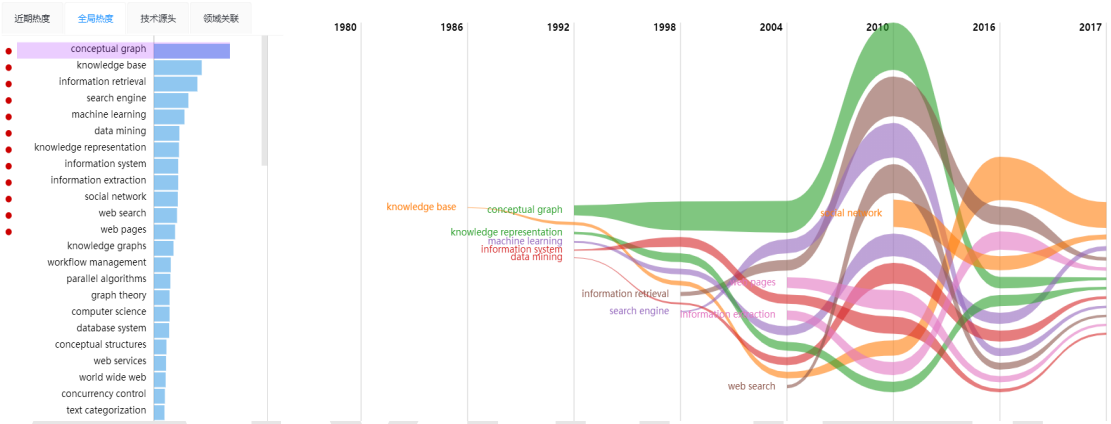


图 40 知识图谱领域全局热度

由以上两图可知，知识库、信息检索、数据挖掘、知识表示、社会网络等方向在知识图谱领域的热度长盛不衰。除此之外，信息提取、查询应答、问题回答、机器学习、概率逻辑、实体消歧、实体识别、查询处理、决策支持等方向的研究热度在近年来逐渐上升，概念图、搜索引擎、信息系统等方向的热度逐渐消退。

知识图谱作为人工智能技术中的知识容器和孵化器，会对未来 AI 领域的发展起到关键性的作用。无论是通用知识图谱还是领域知识图谱，其构建技术的发展和对应用场景的探索仍然会不断的持续下去。知识图谱技术不单指某一项具体的技术，而是从知识表示、抽取、存储、计算、应用等一系列技术的集合。随着这些相关技术的发展，我们有理由相信，知识图谱构建技术会朝着越来越自动化方向前进，同时知识图谱也会在越来越多的领域找到能够真正落地的应用场景，在各行各业中解放生产力，助力业务转型。

参考文献

[1] 知识图谱发展报告[R].北京：中国中文信息学会语言与知识计算专委会,2018.08.

[2] Bordes A, Weston J, Collobert R, et al. Learning structured embeddings of knowledge bases[C], in Proceedings of AAAI 2011. 301-306.

[3] Bordes A, Glorot X, Weston J, et al. Joint learning of words and meaning representations for open-text semantic parsing[C], in Proceedings of AISTATS 2012, 127-135.

[4] Socher R, Chen D, Manning C D, et al. Reasoning with neural tensor networks for knowledge base completion[C], in Proceedings of NIPS 2013: 926-934.

[5] Xiao H, Huang M, Hao Y, et al. TransG: A Generative Mixture Model for Knowledge Graph Embedding[J]. arXiv preprint arXiv:1509.05488. 2015.

[6] He S, Liu K, Ji G, et al. Learning to Represent Knowledge Graphs with Gaussian Embedding[C], in Proceedings of CIKM 2015, 623-632.

[7] Lafferty, J., McCallum, A. and Pereira, F.C., 2001. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In Proceedings of ICML 2001.

[8] Sundheim, B.M., 1996, May. Overview of results of the MUC-6 evaluation. In Proceedings of a workshop on held at Vienna, Virginia: May 6-8, 1996 (pp. 423-442). Association for Computational Linguistics.

[9] 陈维.电子商务语义库[R]．苏州：第一届全国中文知识图谱研讨会,2013．

[10] 胡国平.从应用角度来看知识图谱的价值和挑战[R]．武汉：第二届全国中文知识图谱研讨会，2014．

[11] 阮彤．垂直知识图谱构造工具与行业应用[R]．武汉：第二届全国中文知识图谱研讨会，2014.

[12] Ganea, O.E. and Hofmann, T., 2017. Deep Joint Entity Disambiguation with Local Neural Attention. In Proceedings of EMNLP 2017.

[13] Gupta, N., Singh, S. and Roth, D., 2017. Entity linking via joint encoding of types, descriptions, and context. In Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing (pp. 26812690).

[14] Sil, A., Kundu, G., Florian, R. and Hamza, W., 2018. Neural CrossLingual Entity Linking. In Proceedings of AAAI 2018.

[15] Suchanek, F. M. and Kasneci, G., et al. 2008. YAGO: A large ontology from Wikipedia and Wordnet. In: Web Semantics: Science, Services and Agents on the World Wide Web 6(3): 203-217.

[16] Jinxiu Chen, Donghong Ji, Chew Lim Tan, and Zhengyu Niu. Unsupervised feature selection for relation extraction. In Proceedings of the Second International Joint Conference on Natural Language Processing, IJCNLP ’05, pages 262–267, Berlin, Heidelberg, 2005. Springer-Verlag.

[17] Gabriel Pui Cheong Fung, Jeffrey Xu Yu, and Hongjun Lu. Discriminative category matching: Efficient text classification for huge document collections. In Data Mining, 2002. ICDM 2003.

Proceedings. 2002 IEEE International Conference on, pages 187–194. IEEE, 2002.

[18] Xiaotian Jiang; Quan Wang; Peng Li; Bin Wang. Relation Extraction with Multi-Instance Multi-Label Convolutional Neural Networks. COLING2016.

[19] Chai J Y, Biermann. AW. Learning and generalization in the creation of information extraction systems. [J]. Citeseer, 1998.

[20] Piskorski J, Tanev H, Atkinson M, et al. Online news event extraction for global crisis surveillance [J]. In Transactions on computational collective intelligence V, 2001: 182–212.

[21] Tanev H, Piskorski J, Atkinson M. Real-time news event extraction for global crisis monitoring. [J]. In Proceedings of the International Conference on Application of Natural Language to Information Systems, 2008: 207–218.

[22] 王伟, 赵东岩. 中文新闻事件本体建模与自动扩充[J]. 计 算机工程与科学, 2012, 34(4): 171-176.

[23] Zhengxiang Pan, Jeff Heflin. DLDB: Extending Relational Databases to Support Semantic Web Queries. In Proceedings of PSSS'2003.

[24] Kevin Wilkinson, Craig Sayers, Harumi A. Kuno, Dave Reynolds. Efficient RDF Storage and Retrieval in Jena2. SWDB 2003: 131-150.

[25] Kevin Wilkinson. Jena Property Table Implementation. in SSWS, Athens, Georgia, USA (2006), pp. 35–46.

[26] Marcin Wylot, Jigé Pont, Mariusz Wisniewski, Philippe CudréMauroux. dipLODocus[RDF] - Short and Long-Tail RDF Analytics for Massive Webs of Data. International Semantic Web Conference (1) 2011: 778-793.

[27] Jens Lehmann, Robert Isele, Max Jakob, Anja Jentzsch, Dimitris Kontokostas, Pablo N. Mendes, Sebastian Hellmann, Mohamed Morsey, Patrick van Kleef, Sören Auer, Christian Bizer. DBpedia - A large-scale, multilingual knowledge base extracted from Wikipedia. Semantic Web 6 (2), 167– 195 (2015).

[28] Ron J. Brachman. A Structural Paradigm for Representing Knowledge, Ph.D. thesis. Harvard University, May 1977 Also. BBN Report No.3605, Bolt Beranek and Newman Inc., May 1978.

[29] Demeester, T., Rocktäschel, T., and Riedel, S. (2016). Lifted rule injection for relation embeddings. In Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing (EMNLP), pages 1389-1399.

[30] Dettmers, T., Pasquale, M., Pontus, S., and Riedel, S. (2018). Convolutional 2D knowledge graph embeddings. In Proceedings of the 32nd AAAI Conference on Artificial Intelligence (AAAI).

[31] Yevgeny Kazakov, Pavel Klinov: Advancing ELK: Not Only Performance Matters. Description Logics, 233-248, 2015.

[32] Markus Krötzsch, Maximilian Marx, Ana Ozaki, Veronika Thost: Attributed Description Logics: Ontologies for Knowledge Graphs. International Semantic Web Conference (1) 2017: 418-435.

附录

知识图谱/知识工程的知识树共包括 10 个二级分类和 212 个三级分类。

图中带“<>”的节点表示关系，没有标”<>”的标明的节点关系是上下位关系

一级分类	二级分类	三级分类
A Knowledge Engineering	<is_kind _of>	knowledge technology 知识技术
		semantic (web) technology 语义技术
		web science 万维科学
		information science 情报科学
	<multidiscipline_of>	cognitive science 认知科学
		semantic web 语义网
		artificial intelligence 人工智能
		computer science 计算机科学
		natural language processing 自然语言处理
		information processing 信息处理
		social machine 社交机器
	<using_techniques>	unified modeling language 统一建模语言
		pattern recognition 模式识别
		information processing 信息处理
		clustering 聚类
		clistering algorithms 聚类算法
		data visualization 数据可视化
		data mining 数据挖掘
		quality management 质量管理
		design methodology 设计方法论
		feature extraction 特征提取
		feature space 特征空间
		feature selection 特征选择
		human centered computing 人机交互技术
		support vector machine 支持向量机
		statistical model 统计模型
		service oriented architecture 面向服务的体系结构
		markov chains 马尔可夫链
		social network analysis 社会网络分析
		decision models 决策模型
		data management 数据管理
		human interaction 人机交互
		decission tree 决策树
		genetic algorithm 遗传算法
		machine learning 机器学习

		information retrieval 信息检索
		semantic similarity 语义相似度
		semantic relatedness 语义相关性
		semantic computing 语义计算
		semantic analysis 语义分析
		graph theroy 图论
		explanation based learning 解释学习
		data integrity 数据完整性
		text anaylsis 文本分析
		text mining 文本挖掘
		bootstrapping method 拔靴法
		reinforcement learning 强化学习
		human computer interaction 人机交互
		transfer learning 迁移学习
		domain experts 领域专家
		situation aware 情境感知
		graphical user interface 图形用户界面
		predictive model 预测模型
	<aims_at>	better understanding 内涵理解
		computational intelligence 智能计算
		knowledge based system 知识系统
		knowledge base 知识库
		RDF repository 资源描述框架存储库
		knowledge management 知识管理
		knowledge management systems 知识管理系统
		decision support system 决策支持系统
		decision models 决策模型
		decision maker 决策者
		adaptive systems 自适应系统
		recommender systems 推荐系统
		multiagent systems 多智能体系统
		multi agnet systems 多智能体系统
		automomous systems 自动系统
		automomous agent 自动代理
		semantic search 语义检索
		question answering system 问答系统
		human robot interaction 人机交互
		intelligent assistant 智能辅助
		knowledege reuse 知识再利用
		knowledge sharing 知识共享
		expert system 专家系统

A		intelligent systems 智能系统
	<using_information_sources>	social networks 社会网络
		web resource 网络资源
		world wide web 万维网
		distributed databases 分布式数据库
		big data 大数据
		information sources 信息源
		xml database 可扩展标志语言数据库
		heterogeneous database 异构数据库
		heterogeneous data source 异构数据源
		multimedia 多媒体
		wireless networks 无线网络
		relational database 关系数据库
	<appliations_in>	electric commerce 电子商务
		disaster management 灾害管理
		computational biology 计算生物学
		biomedical domain 生物医学领域
		health care 卫生保健
		scientific domain 科学域
		education 教育
		open government data 政府公开数据
		life science 生命科学
		gene expression data 基因表达数据
	knowledge representation	data model 数据模型
		concept modelling 概念模型
		concept model 概念模型
		conceptual model 概念模型
		semantic model 语义模型
		knowledge model 知识模型
		structured data 结构化数据
		formal specification 形式描述
		formal meaning prepresentation 形式意义表示
		formal semantics 形式语义
		commonsense knowledge 常识
		world knowledge 世界知识
		web of data 数据网
		background knowledge 背景知识
		domain knowledge 领域知识
		semantic network 语义网络
		ontology 本体论
		rough set 粗糙集
		rough set theroy 粗糙集理论

		concept map 概念图
		fuzzy sets 模糊集合
		rule based 基于规则
		rule based system 基于规则系统
		heuristic rule 启发式规则
		object oriented 面向对象
		semantic workflow 语义工作流
		first order logic 一阶逻辑
		logic programming 逻辑编程
		frame based system 框为本的系统
		fuzzy logic 模糊逻辑
		fuzzy systems 模糊系统
		formal logic 形式逻辑
		decision rule 决策规则
		temporal logic 时态逻辑
		dynamic logic 动态逻辑
		domain specific language 领域专用语言
		resource description framework 资源描述框架
		ontology language 本体语言
		web ontology lanuage 网络本体语言
		semantic web rule language 语义网规则语言
		owl 2
		collabrative ontology engineering 联合本体工程
		ontology engineering 本体工程
		ontology development 本体开发
		collabrative ontology development 联合本体开发
		ontology extraction 本体抽取
		ontology evolution 本体演化
		ontology versioning 本体版本
	knowledge acquisition	knowledge extraction 知识提取
		knowledge capture 知识获取
		knowledge construction 知识建构
		knowledge building 知识建构
		information extraction 信息提取
		entity resolution 实体解析
		entity recognition 实体识别
		entity disambiguation 实体消歧
		semantic annotation 语义标注
		taxonmy induction 感应规范
		concept clustering 概念聚类

A		concept formation 概念形成
		concept learning 概念学习
		attribute value taxonomy 属性分类规范
		event detection 事件检测
		event identificaton 事件识别
		event extraction 事件抽取
		relation extraction 关系抽取
		semantic relation learning 语义关系学习
		relational learning 关系学习
		inference rule 推理规则
		rule learning 规则学习
	knowledge reasoning	case based reasoning 实例推理
		logical implication 逻辑蕴涵
		inference mechanisms 推理机制
		knowledge verification 知识验证
		semantic interpretation 语义解释
		uncertainty reasoning 不精确推理
		causual models 因果模型
		nonmonotonic reasoning 非单调推理
		spatial reasoning 空间推理
		temporal reasoning 时序推理
		abductive reasoning 溯因推理
		defult reasoning 默认推理
	knowledge integration	knowledge fusion 知识融合
		semantic integration 语义集成
		data fusion 数据融合
		inconsistent ontology 本体不一致
		heterogenous ontology 异构本体
		ontology interoperablity 互用性本体
		ontology mapping 本体映射
		ontology alignment 本体映射
		ontology matching 本体匹配
		schema mapping 模式映射
		schema matching 模式匹配
		matching function 匹配函数
		instance matching 实例匹配
		date linking 日期链接
		date interlinking 日期互联
		record linkage 记录链接
		thesaurus alignment 同义对齐
	knowledge storage	triple store 三元组存储
		RDF database 资源描述框架数据库

		RDF storge 资源描述框架存储
		graph database 图数据库
		exhaustive indexing 完整索引
		query language 查询语言
		conjunctive quries 合取查询
		RDF query 资源描述框架查询
		graph query 图查询
		query rewrite 查询重写
		distributed query 分布式查询
		subgraph matching 子图匹配
		graph partitioning 图划分
		data partitioning 数据划分

版权声明

AMiner 研究报告版权为 AMiner 团队独家所有，拥有唯一著作权。AMiner 咨询产品是 AMiner 团队的研究与统计成果，其性质是供用户内部参考的资料。

AMiner 研究报告提供给订阅用户使用，仅限于用户内部使用。未获得 AMiner 团队授权，任何人和单位不得以任何方式在任何媒体上（包括互联网）公开发布、复制，且不得以任何方式将研究报告的内容提供给其他单位或个人使用。如引用、刊发，需注明出处为“AMiner.org”，且不得对本报告进行有悖原意的删节与修改。

AMiner 研究报告是基于 AMiner 团队及其研究员认可的研究资料，所有资料源自 AMiner 后台程序对大数据的自动分析得到，本研究报告仅作为参考，AMiner 团队不保证所分析得到的准确性和完整性，也不承担任何投资者因使用本产品与服务而产生的任何责任。