科技大数据知识图谱

杜一

中国科学院计算机网络信息中心

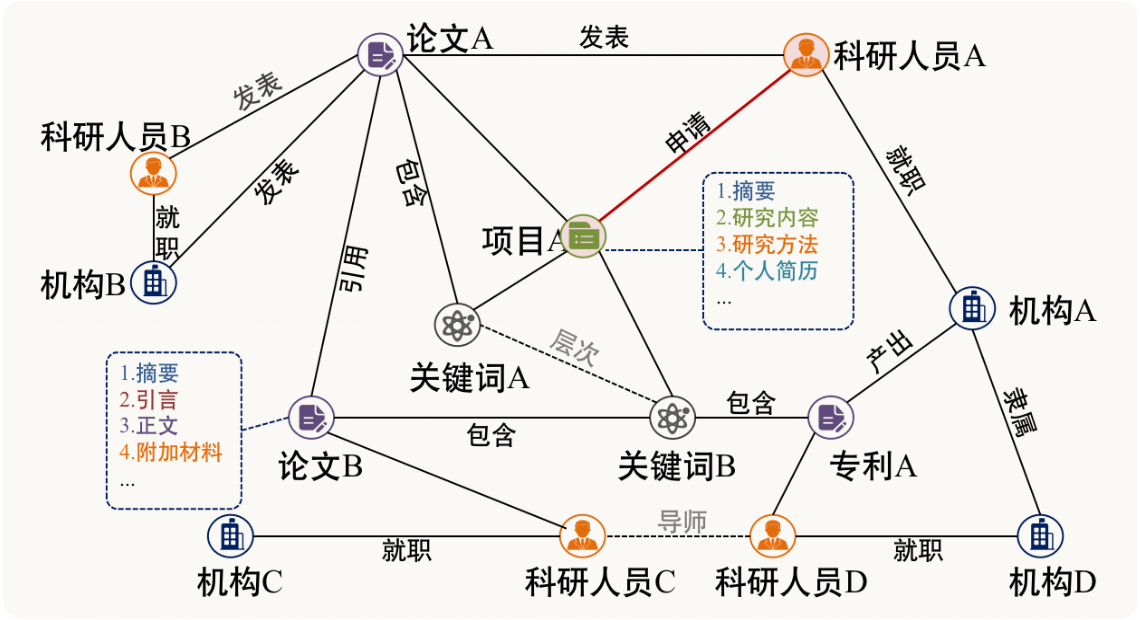
研究背景：科技管理(管理科学)需求与大数据(信息科学)的交叉背景

5亿+篇
论文报告

100+EB
科学数据

500万
项目申请

科技精细化决策面临挑战：如何从海量数据中，发现辅助决策的知识



科研实体 (亿级)与关系(百亿级)
构成的大规模语义网络

- 实体：学者、机构、项目、论文等
- 关系：合作、隶属、引用、师承等

科技大数据知识图谱可针对科技决策精细化需求，给出可解释性的推荐

国家自然科学基金大数据知识管理服务平台

基金委原有信息系统

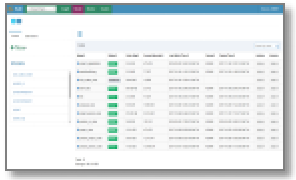
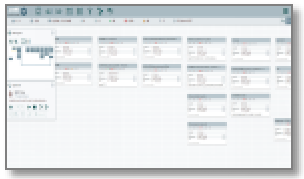
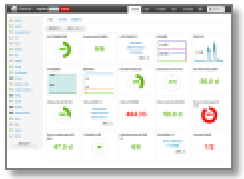
数据量增大，原有数据库无法满足存储、计算需求

数据孤岛，降低了数据之间连通并发挥更大作用的空间

原有系统不支持更加深入的分析与挖掘（知识图谱）

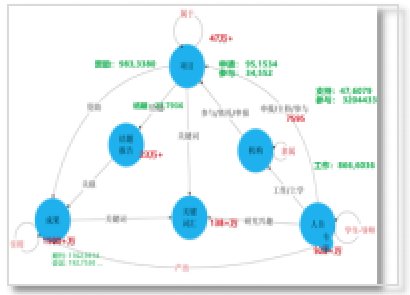
一套平台

构建了一套大数据**全栈技术平台**，为基金大数据的采集、存储、分析、检索与应用等服务提供了分布式、可扩展的存储与计算环境。

一张图谱

全面升级基金数据知识模型，建立“**机构-项目-人员-成果-关键词**”的知识图谱，打破数据孤岛格局，形成面向服务的数据湖。



一批应用

基于统一的服务接口，**建成**知识服务门户、多维统计、网络挖掘、交叉预测、专家画像等示范应用，形成了“**口同径、数同源、能洞察**”的大数据应用新能力。




接口累计服务3亿+次

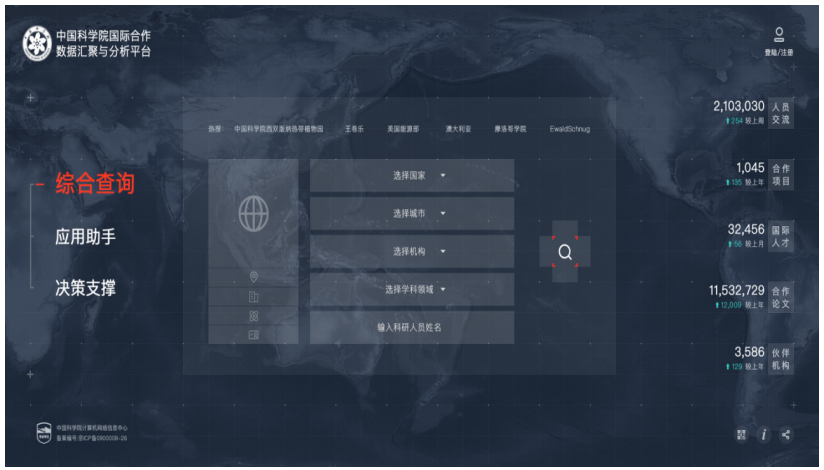
姚畅;王晓帆;杜一;张兆田;李建军;郝艳妮;;国家自然科学基金大数据知识管理服务平台总体方案及关键技术研究[J];中国科学基金;2019年01期

科技领域知识服务平台

中国科学院
计算机网络信息中心
Computer Network Information Center,
Chinese Academy of Sciences

空间科学领域知识图谱与决策支持系统

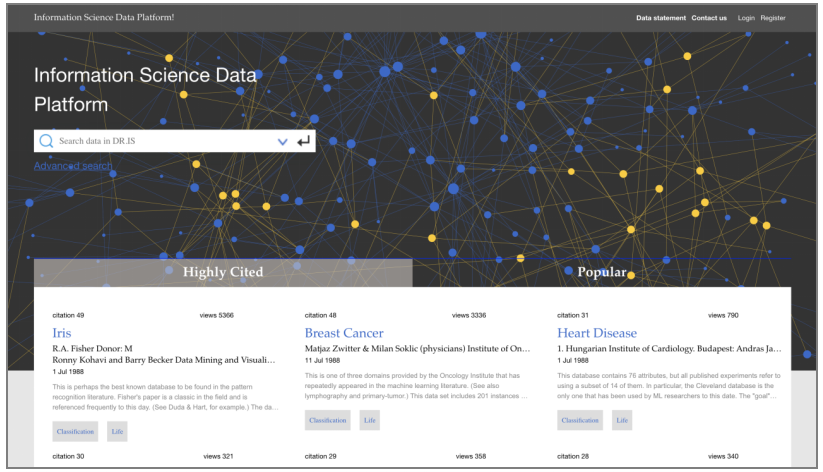
基于知识图谱的科技合作智能化管理服务平台



中科院国际合作数据汇聚与分析平台

中科院国际合作数据汇聚与分析平台

烟草科技知识图谱服务平台

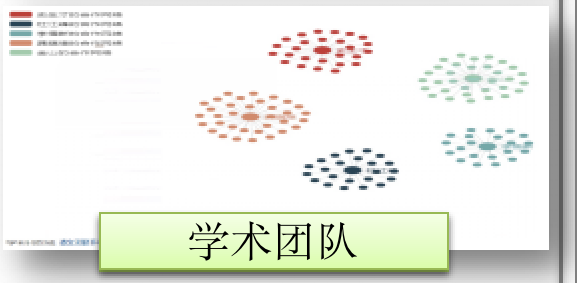


Information Science Data Platform

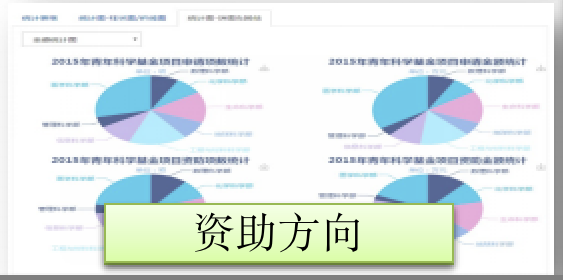
科技领域知识服务平台

学科趋势

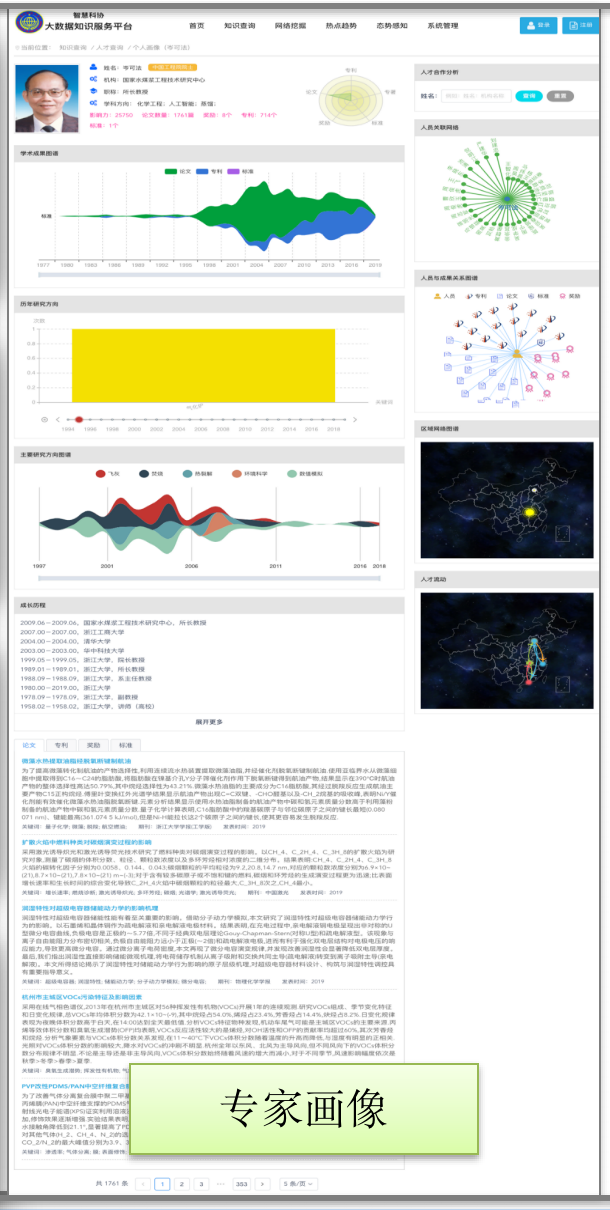
交叉学科



学术团队

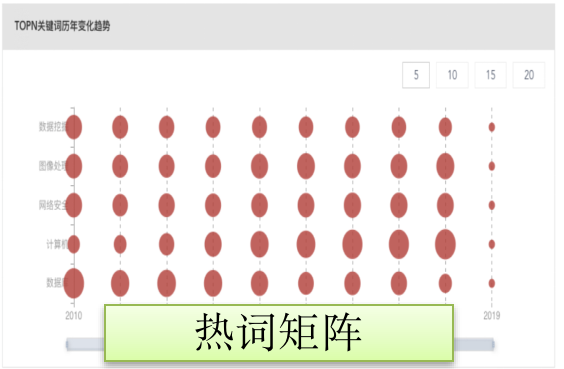


资助方向

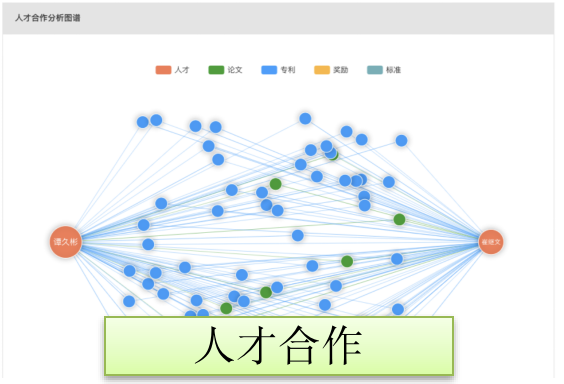


专家画像

学科热词

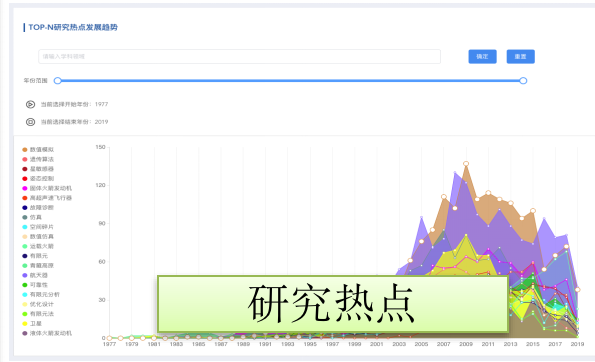


热词矩阵

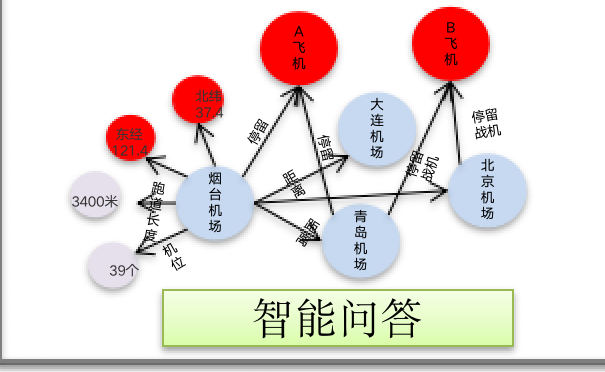


人才合作

关联路径



研究热点

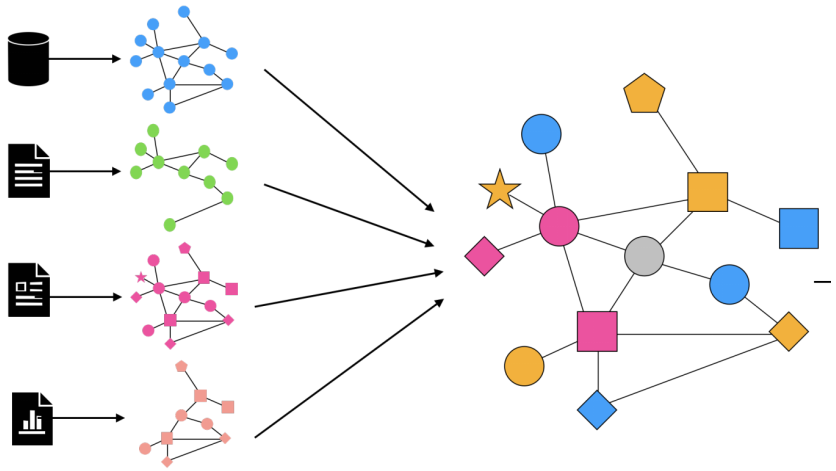


智能问答

问题与挑战

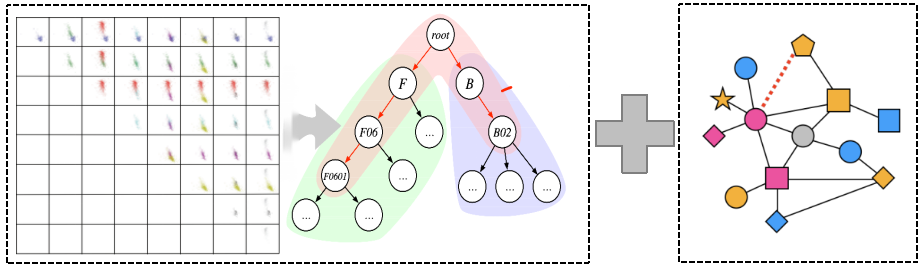
“构建-使用-拓展”的知识图谱技术，在科技大数据知识图谱辅助决策时面临特殊挑战

“建”图谱



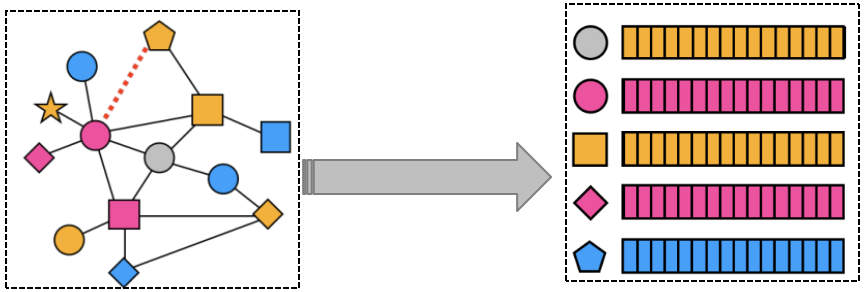
科技情报大数据存在大量人名、概念的歧义，使得知识图谱建不准

“用”图谱



科技管理复杂精细，使得知识图谱与科技决策需求结合难

“拓”图谱



知识图谱结构本身的非度量特性，使得不同领域的扩展难

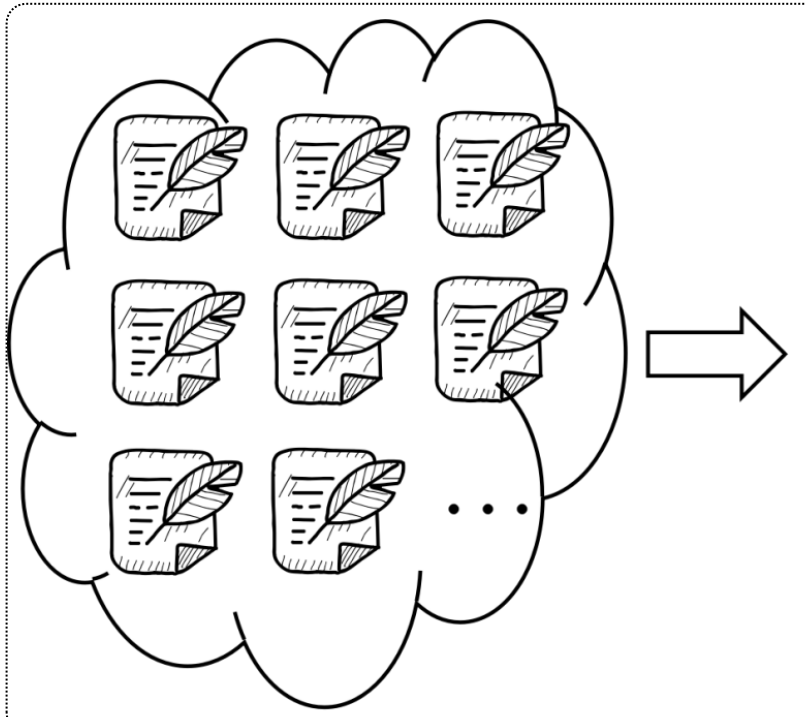
工作1.“建图谱”-知识图谱构建过程中的人名与概念消歧

难点：海量科技情报数据中人名、概念等实体的表达方式多样

2亿科研论文，涉及10亿人次姓名，上百亿次机构与关键词，准确消歧难度大

“建图谱” → “用图谱”

“建图谱” → “拓图谱”



Synonyms
异名同义

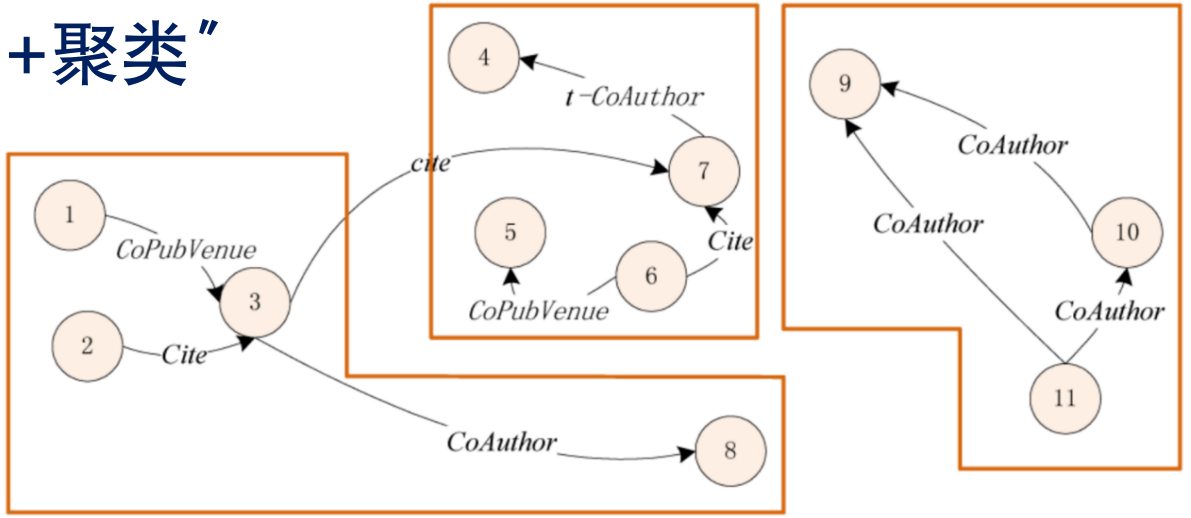
Homonyms
同名异义

举例：在谷歌学术搜索(Google Scholar)中，W.Wang相关成果409万项。

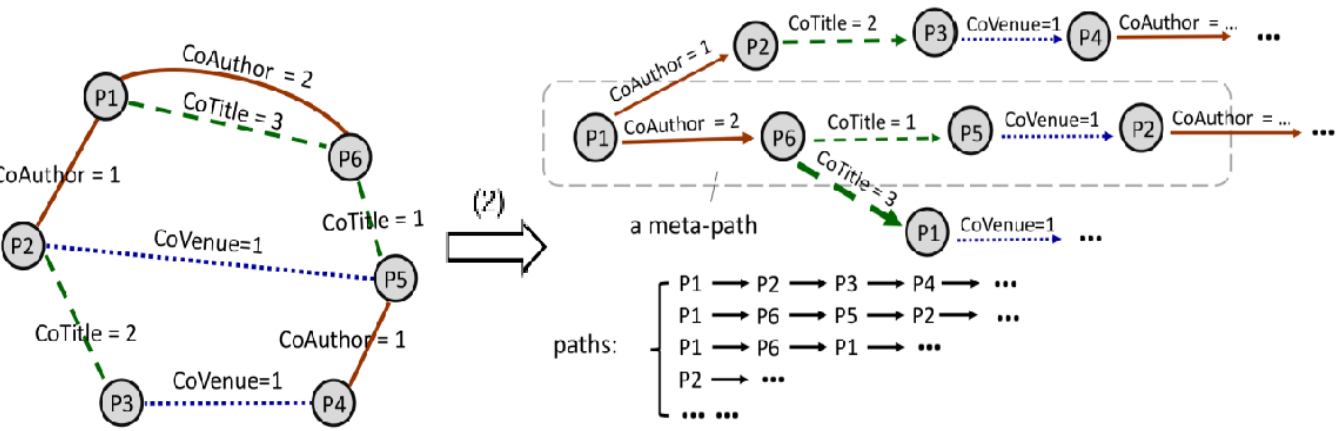
提出基于知识表示与聚类的消歧模型TGNN

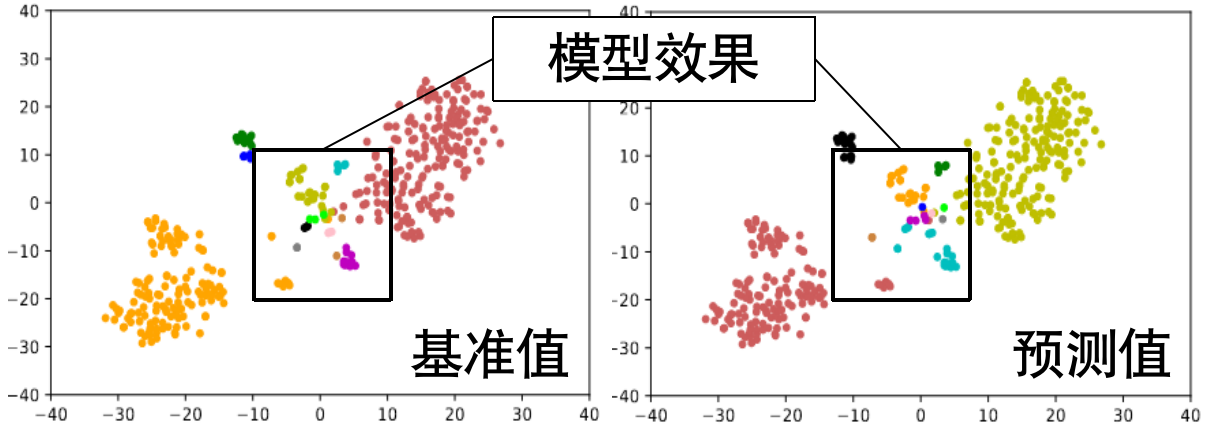
将科技大数据中的歧义问题定义为“表示+聚类”

- 对于M项待消歧的人名与概念
  - Input: $D_a = \{p_1^{\mathrm{a}}, p_2^{\mathrm{a}}, \ldots, p_m^{\mathrm{a}}\}$
  - Output: $C^a = \{C_1^a, C_2^a, \ldots, C_k^a\}$



在metapath2vec基础上，引入领域知识定义游走概率，进行图表示学习后聚类





科技领域异质网络表示

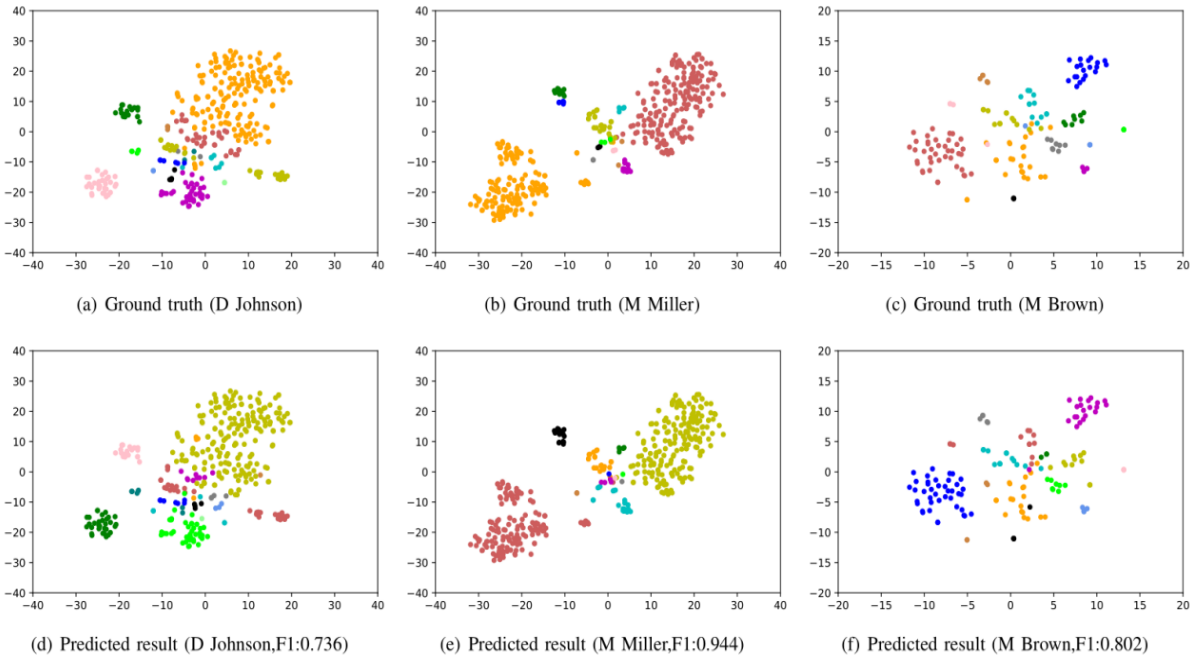
模型在多个学术知识图谱中进行验证，并取得了较好的效果。

TABLE I
THE PERFORMANCE OF DIFFERENT METHODS ON AMINER DISAMBIGUATION DATASET

Name	Our method	Component	Zhang et al. 2018 [1]	Xu et al. 2018	Zhang et al. 2017 [5]	DeepWalk	LINE	Metapth2Vec	Hin2Vec	GraphSAGE
Ajay Gupta	**0.750**	0.329	0.568	0.552	0.618	0.370	0.578	0.298	0.684	0.654
Alok Gupta	**1**	0.690	0.689	0.892	0.590	0.582	0.835	0.663	0.734	0.651
Bin Yu	**0.696**	0.292	0.431	0.585	0.614	0.490	0.475	0.354	0.490	0.441
David Cooper	0.900	0.327	0.737	0.884	**0.931**	0.737	0.833	0.833	**0.931**	0.862
David Nelson	**0.944**	0.219	0.750	0.735	0.556	0.353	0.523	0.788	0.635	0.710
Fei Su	**1**	0.648	0.933	0.630	0.941	0.684	0.721	0.930	0.917	0.948
Hao Wang	0.604	0.086	0.403	0.557	0.543	0.382	0.400	0.420	**0.624**	0.192
Jie Tang	**0.982**	0.883	0.657	0.522	0.910	0.738	0.432	0.902	0.825	0.741
Thomas Wolf	**0.860**	0.502	0.703	0.522	0.352	0.320	0.357	0.390	0.516	0.710
Yang Wang	0.548	0.118	0.273	**0.574**	0.409	0.171	0.211	0.310	0.443	0.204
Avg.	**0.786**	0.507	0.715	0.681	0.680	0.563	0.606	0.643	0.629	0.678

TABLE II
THE PERFORMANCE OF DIFFERENT METHODS ON CITESEEX DISAMBIGUATION DATASET

Name	Our method	Component	Zhang et al. 2018 [1]	Xu et al. 2018	Zhang et al. 2017 [5]	DeepWalk	LINE	Metapth2Vec	Hin2Vec	GraphSAGE
A Kumar	**0.648**	0.392	0.412	0.443	0.307	0.367	0.389	0.478	0.498	0.369
C Chen	**0.442**	0.091	0.299	0.437	0.384	0.155	0.239	0.274	0.431	0.248
D Johnson	0.736	0.454	**0.745**	0.696	0.667	0.487	0.613	0.595	0.590	0.729
J Martin	**0.731**	0.512	0.649	0.495	0.481	0.483	0.529	0.567	0.665	0.596
J Robinson	**0.695**	0.360	0.384	0.626	0.369	0.498	0.450	0.507	0.540	0.554
J Smith	**0.889**	0.201	0.613	0.824	0.753	0.296	0.671	0.796	0.717	0.657
M Brown	**0.802**	0.368	0.710	0.590	0.498	0.526	0.617	0.729	0.560	0.752
M Miller	**0.944**	0.578	0.730	0.913	0.885	0.566	0.621	0.621	0.639	0.895
S Lee	**0.602**	0.078	0.401	0.573	0.553	0.121	0.417	0.417	0.560	0.411
Y Chen	**0.777**	0.124	0.441	0.762	0.770	0.446	0.664	0.665	0.402	0.436
Avg.	**0.698**	0.288	0.538	0.646	0.561	0.429	0.507	0.547	0.563	0.523



(a) Ground truth (D Johnson) (b) Ground truth (M Miller) (c) Ground truth (M Brown)

(d) Predicted result (D Johnson,F1:0.736) (e) Predicted result (M Miller,F1:0.944) (f) Predicted result (M Brown,F1:0.802)

Fig. 3. t-SNE Visualization of embedding spaces on publications of three different names (D Johnson, M Miller, M Brown). Each color in (a), (b), (c) denotes an ground truth cluster which contains publications of a distinct author entity, while each color in (d), (e), (f) denotes a predicted cluster generated by our graph enhanced hierarchical agglomerative clustering.

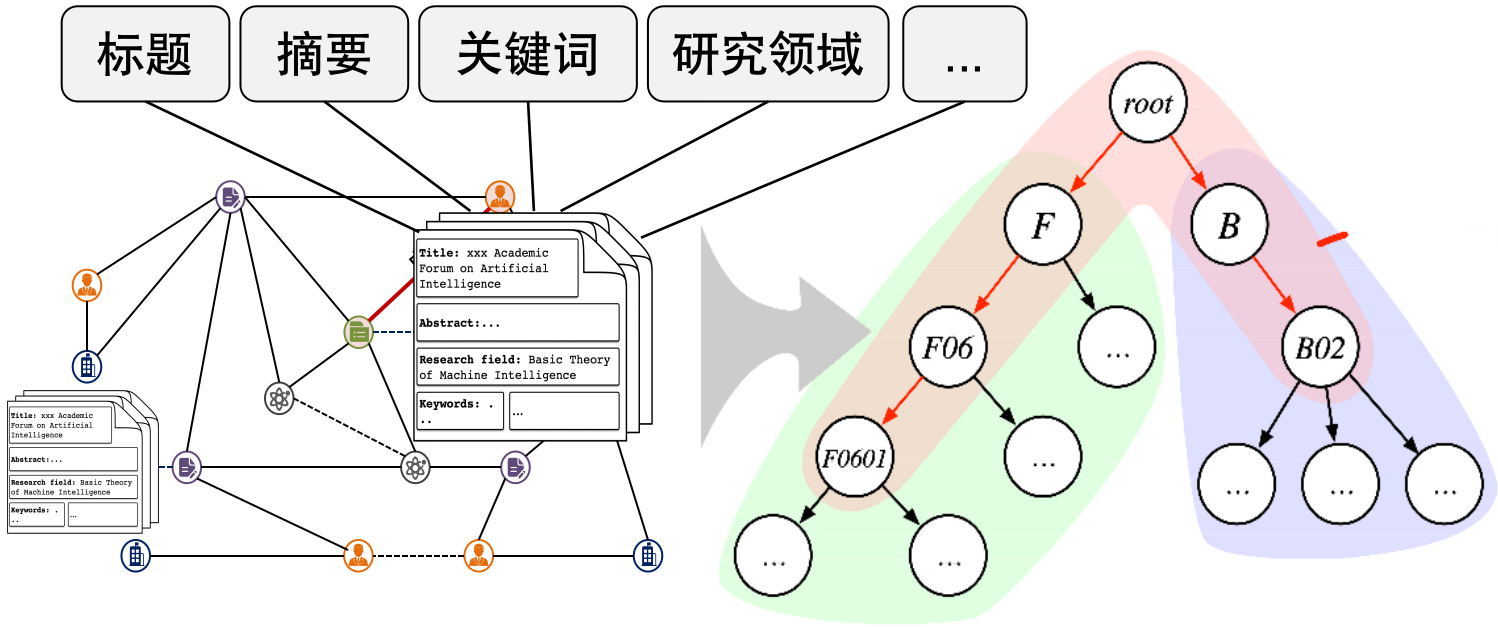
Ziyue Qiao, Yi Du*, Yanjie Fu, Pengfei Wang, Yuanchun Zhou. (2019) Unsupervised Author Disambiguation using Heterogeneous Graph Convolutional Network Embedding, IEEE International Conference on Big Data,2019.

工作2.“用图谱”-科技文本交叉性研判

难点：科技文本的复杂性与研究领域的多样性

如何将复杂科技文本信息与科技大数据知识图谱融合，准确分类到上百万交叉组合中

“建图谱” → “用图谱” → “拓图谱”

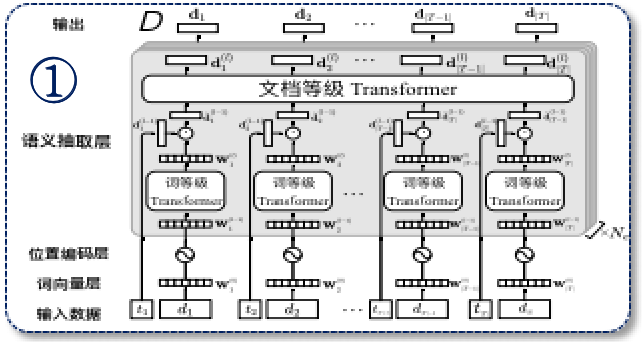


申请代码数	理论组合数
1	2494
2	3,108,771
3	~25亿

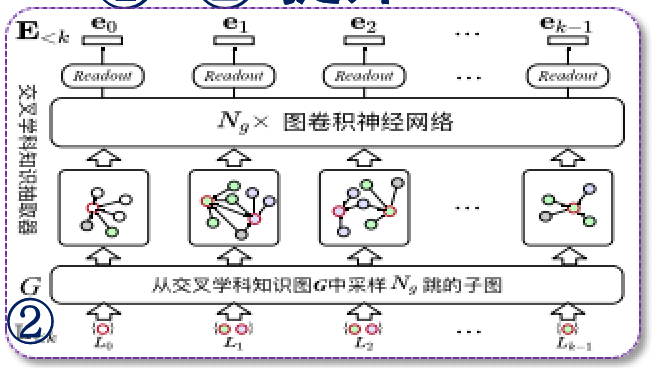
创新点:基于科技知识图谱的多标签、多粒度交叉学科分类模型

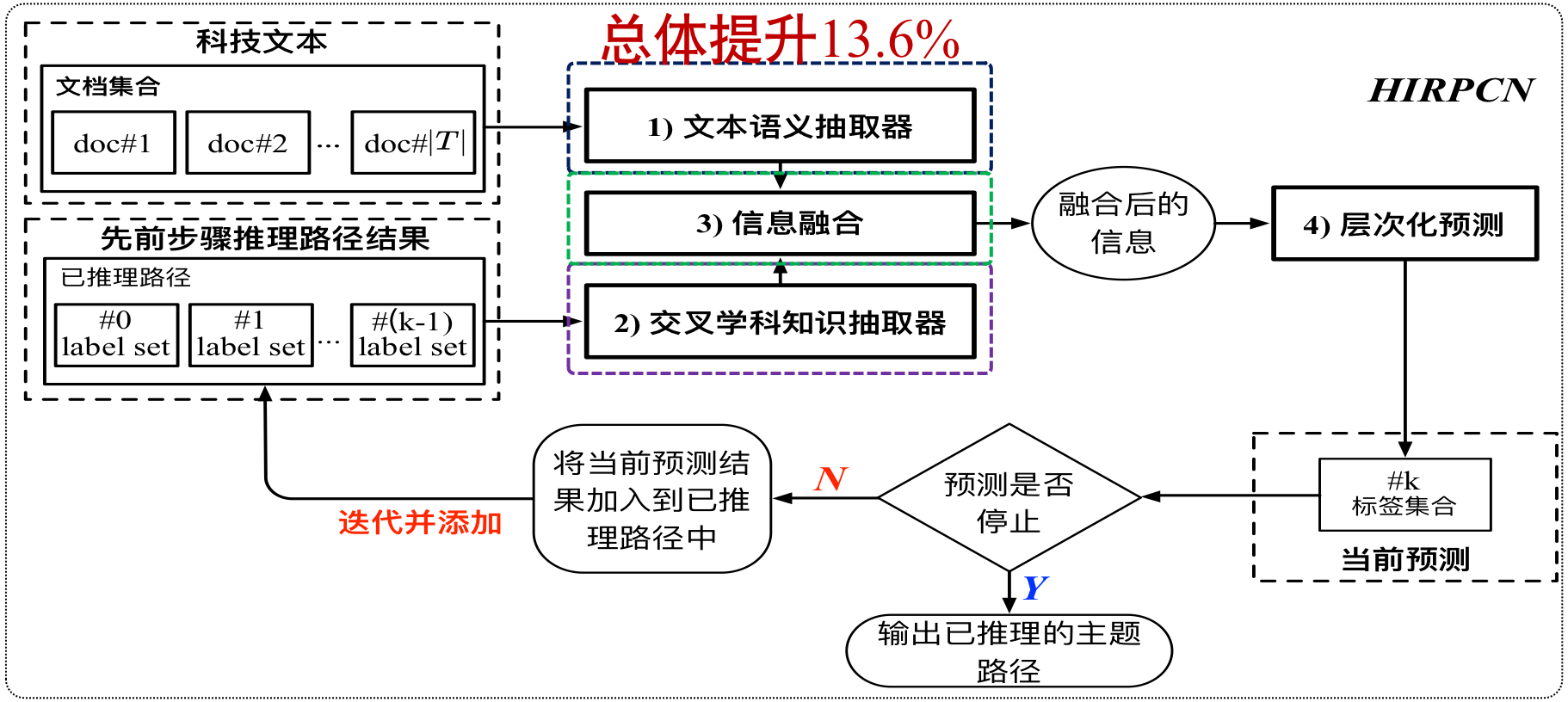
$$\Omega(D, G, \gamma, \mathbb{L}_{<k}; \Theta) \rightarrow L_k, \qquad P(\mathbb{L}|D, G, \gamma; \Theta) = \prod_{k=1}^{H_A} P(L_k|D, G, \gamma, \mathbb{L}_{<k}; \Theta)$$

科技大数据知识图谱



①+②:提升5.9%





模型效果

预测出的交叉代码

Discipline #1 : C060703 (Biological Data Integration and Biological Big Data)

Discipline #2 : F0305 (Biological and Medical Information Systems and Technology)

Title: Genetic Variation Driven 遗传变异驱动的肿瘤耐药相关非编码RNA深度挖掘算法及在肺癌中的功能研究 (Drug Resistance; Encode; Deep Mining Algorithm; Lung Cancer)

Keywords: 人非编码RNA肿瘤化疗耐药深度挖掘算法生物大数据 (Encode; Deep Mining Algorithm; Bio Big Data)

Abstract: 肿瘤细胞对化疗药物产生耐药性是肿瘤治疗失败的重要原因长链非编码RNA lncRNA可以参与肿瘤耐药调控网络为挖掘关键治疗靶点和解析肿瘤耐药机制提供了新的机遇但利用多组学肿瘤大数据系统预测肿瘤耐药相关lncRNA遗传变异并将其用于肿瘤病人的化疗用药指导仍是挑战本项目在前期工作的基础上将大数据分析算法开发和实验验证相结合系统整合肿瘤测序数据基因组注释数据肿瘤组学特征和临床用药数据构建完整的肿瘤化疗耐药 (Mining; Algorithm; Genetic Variation)

Research of Field: 生物数据与信息挖掘与共享 (Bio data; Information Mining)

交叉代码预测

主题代码	真实主题：H - 医学科学	额外补充：B - 化学科学
标题	环境应激与化学致癌	
关键词	环境毒理	化学致癌 … NRF1/2
研究领域	卫生毒理	
摘要	…围绕砷中毒，将实验室研究与流行病学调查相结合，率先揭示慢性砷暴露人群氧化-还原稳态失衡是砷代谢影响砷致癌的介导事件；发现成体干细胞的"旁观者"效应及NRF2作为应激网络枢纽分子在砷致癌过程中的重要作用…探究NRF1在化学物致肺癌中的作用与机制，构建基于NRF1/2-ARE通路的化学致癌预防策略和方法。	

主题代码	真实主题：A - 数理科学	额外补充: E - 工程与材料科学部
标题	基于多级结构微纳米纤维…非织造布高效空气过滤材料…	
关键词	空气过滤	熔喷 … 力学机理
研究领域	工业流体力学	
摘要	微纳米纤维非织造布作为高效空气过滤材料…将仿生结构引入纤维，设计、制备由多级结构微纳米纤维形成的非织造布正在成为提升高效空气过滤材料性能…。本项目通过设计…形貌以优化微纳米…的目标，通过模拟…；同时通过研究…运动来分析…过滤性能。最后，以获…主要目标，优化…。	

隐含交叉特征

隐含交叉特征预测

工作3.“拓图谱”-大数据知识图谱的高效表示

中国科学院 计算机网络信息中心 Computer Network Information Center, Chinese Academy of Sciences

难点：知识图谱中的实体之间难以度量，制约了其扩展性

异质网络表示学习方法，可以将知识图谱映射到度量空间，但无法嵌入层次、上下文等图谱结构信息

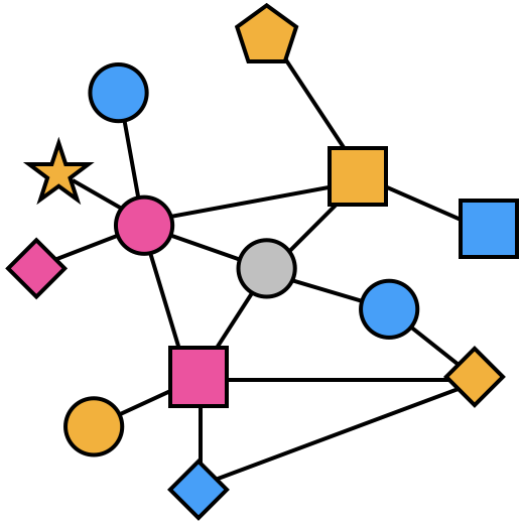
“建图谱” → “用图谱”

“建图谱” → “拓图谱”

科技情报决策

军事指挥决策

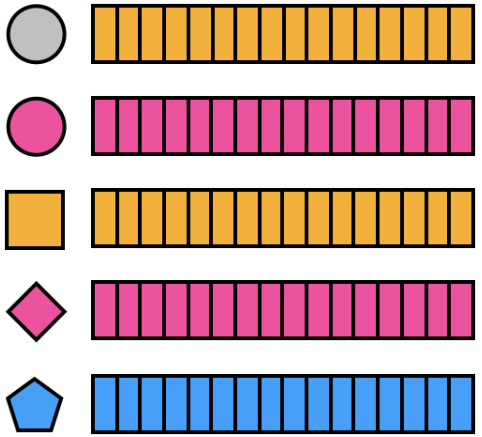
其它决策场景



非度量空间

知识图谱表示学习

将知识图谱中节点表示为低维、稠密的向量



度量空间

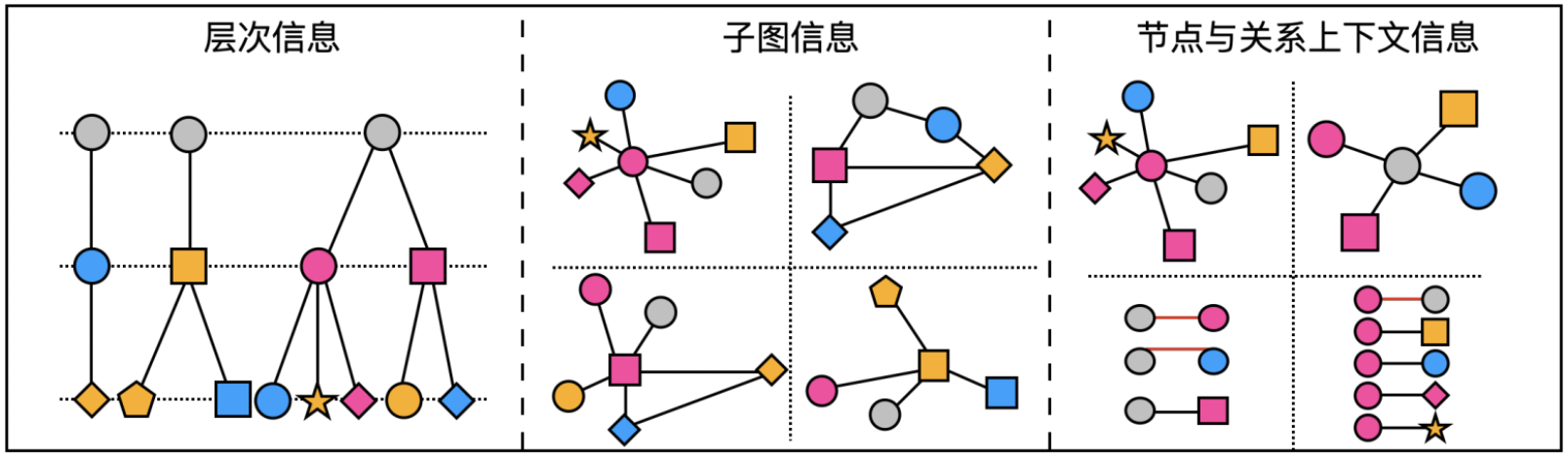
提出融合邻域信息的知识图谱表示学习方法

Input:

$$G = (N, E, \mathcal{T}, \mathcal{R})$$

Output:

$$\{f_t : N_t \rightarrow \mathbb{R}_t^{d'}\}, t \in \mathcal{T}$$



DBLP 科技情报数据	IMDB 电影娱乐数据	YELP 食品点评数据

- 对比模型：DeepWalk、Metapath2vec、RHIN、GraphSAGE、GAT、HAN、HetGNN等

聚类任务(+5.36%)　　分类任务(+4.77%)　　链接预测任务(+6.67%)

在3个基准数据的8个不同任务上，相比14个主流模型的最好效果均明显提升

层次结构邻域：融合层次结构的知识图谱表示学习方法

中国科学院 计算机网络信息中心 Computer Network Information Center, Chinese Academy of Sciences

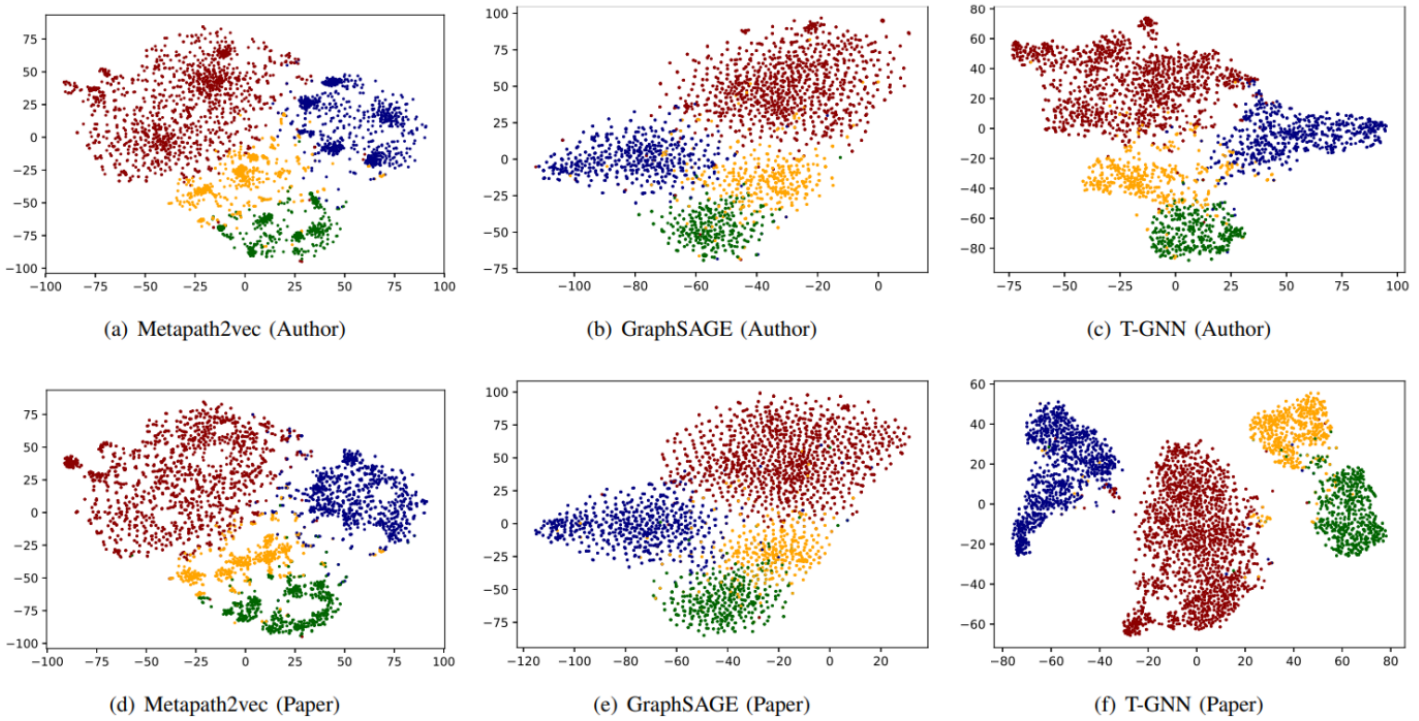
模型在学术知识图谱、电影知识图谱以及餐饮点评知识图谱中进行验证，并取得了较好的效果。

TABLE II: Results of Clustering

Dataset	DBLP		IMDB		YELP	
Metric	NMI	ARI	NMI	ARI	NMI	ARI
DeepWalk	0.735	0.616	0.023	0.015	0.260	0.279
Metapath2vec	0.864	0.899	0.096	0.091	0.261	0.282
RHINE	0.866	0.902	0.055	0.036	0.342	0.351
GraphSAGE	0.865	0.912	0.128	0.135	0.396	0.433
GAT	0.855	0.897	0.114	0.113	0.413	0.457
HAN	0.900	0.933	0.136	0.144	0.413	0.458
HetGNN	0.891	0.939	0.131	0.139	0.403	0.440
T-GNN	**0.916**	**0.955**	**0.145**	**0.152**	**0.420**	**0.484**

TABLE III: Results of Multi-class Classification

Dataset	DBLP		IMDB		YELP	
Metric(F1)	Micro	Macro	Micro	Macro	Micro	Macro
DeepWalk	0.905	0.896	0.478	0.473	0.703	0.660
Metapath2vec	0.922	0.918	0.505	0.509	0.705	0.660
RHINE	0.927	0.920	0.449	0.448	0.726	0.676
GraphSAGE	0.962	0.943	0.583	0.584	0.739	0.748
GAT	0.967	0.958	0.543	0.542	0.744	0.758
HAN	0.979	0.971	0.596	0.598	0.746	0.759
HetGNN	0.983	0.979	0.594	0.593	0.730	0.753
T-GNN	**0.997**	**0.996**	**0.608**	**0.609**	**0.760**	**0.772**



(a) Metapath2vec (Author) (b) GraphSAGE (Author) (c) T-GNN (Author)

(d) Metapath2vec (Paper) (e) GraphSAGE (Paper) (f) T-GNN (Paper)

Fig. 7: t-SNE Visualization of representation distribution of authors and papers learned by three different graph representation learning methods: Metapath2vec, GraphSAGE and T-GNN. (a), (b), (c) denotes the author representations learned by these three methods and (d), (e), (f) denotes the paper representations. The four different research areas which authors and papers belong to are colored differently.

Ziyue Qiao, Pengyang Wang, Yanjie Fu, Yi Du*, Pengfei Wang, Yuanchun Zhou. Tree Structure-Aware Graph Representation Learning via Integrated Hierarchical Aggregation and Relational Metric Learning, The 20th IEEE International Conference on Data Mining (ICDM),2020.

子图结构邻域：融合邻域子图结构的知识图谱表示学习方法

中国科学院 计算机网络信息中心 Computer Network Information Center, Chinese Academy of Sciences

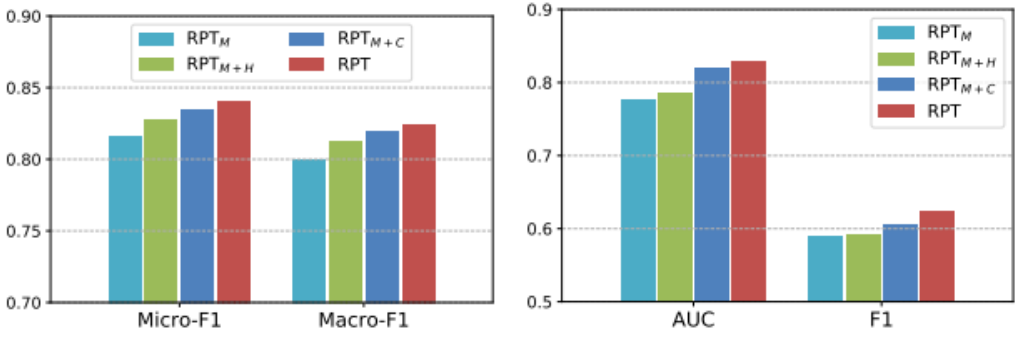
应用于大量无监督的研究人员数据上，并运用到下游多个研究人员数据挖掘任务中，例如研究人员分类，合作者推荐，相关性检索等。

TABLE III
RESULTS OF RESEARCHER CLASSIFICATION.

Proportions(%)	10/10/80		20/10/70		30/10/60	
Metric(F1)	Micro	Macro	Micro	Macro	Micro	Macro
Doc2vec	0.781	0.759	0.799	0.780	0.804	0.787
Metapath2vec	0.760	0.733	0.769	0.743	0.784	0.763
AHNE	0.790	0.770	0.811	0.786	0.810	0.797
BERT	0.815	0.792	0.824	0.805	0.838	0.825
GraphSAGE	0.800	0.777	0.812	0.795	0.820	0.808
RGCN	0.835	0.818	0.838	0.823	0.841	0.825
RPT (fb)	0.827	0.805	0.848	0.833	0.852	0.837
RPT (e2e)	**0.840**	**0.824**	**0.850**	**0.835**	**0.855**	**0.840**

TABLE IV
RESULTS OF COLLABORATING PREDICTION. 1K = 1000.

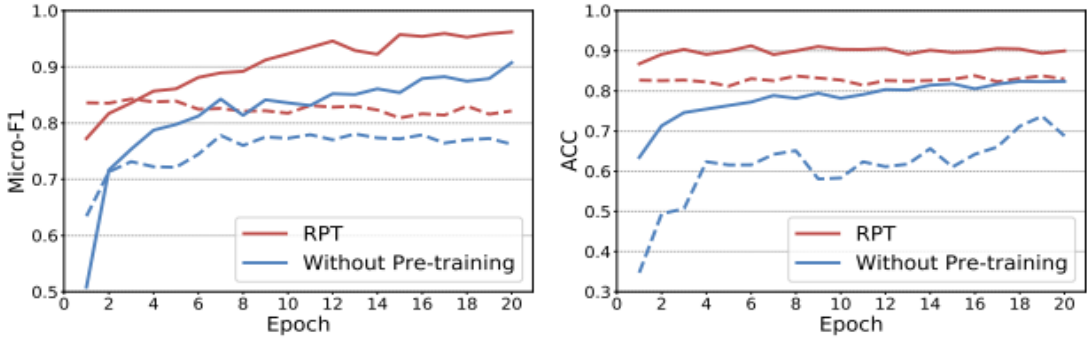
Number of links	1k/1k/10k		3k/1k/10k		5k/1k/10k	
Metric(F1)	ACC	F1	ACC	F1	ACC	F1
Doc2vec	0.764	0.240	0.767	0.310	0.778	0.363
Metapath2vec	0.778	0.347	0.794	0.391	0.796	0.416
AHNE	0.796	0.483	0.809	0.506	0.816	0.508
BERT	0.785	0.378	0.799	0.520	0.805	0.543
GraphSAGE	0.777	0.308	0.781	0.477	0.795	0.482
RGCN	0.811	0.502	0.833	0.608	0.841	0.640
RPT (fb)	0.805	**0.624**	0.827	0.622	0.828	0.633
RPT (e2e)	**0.829**	0.623	**0.840**	**0.641**	**0.860**	**0.674**



(a) Researcher Classification (b) Collaborating Prediction

TABLE V
RESULTS OF TOP-K RESEARCHER RETRIEVAL

K	1		5		10		15		20	
Metric	Pre@K	Rec@K	Pre@K	Rec@K	Pre@K	Rec@K	Pre@K	Rec@K	Pre@K	Rec@K
Doc2vec	0.278	0.041	0.152	0.099	0.107	0.131	0.088	0.154	0.076	0.171
Metapath2vec	0.506	0.076	0.259	0.177	0.189	0.235	0.153	0.272	0.131	0.299
ASNE	0.514	0.077	0.288	0.192	0.198	0.244	0.159	0.279	0.134	0.304
BERT	0.594	0.089	0.299	0.205	0.195	0.246	0.149	0.272	0.124	0.291
GraphSAGE	0.722	0.104	0.389	0.253	0.252	0.304	0.191	0.333	0.158	0.355
RGCN	0.757	**0.106**	0.480	0.288	0.322	0.352	0.246	0.386	0.201	0.409
RPT (fb)	0.597	0.090	0.337	0.217	0.231	0.273	0.182	0.309	0.154	0.337
RPT (e2e)	**0.787**	**0.106**	**0.525**	**0.309**	**0.357**	**0.381**	**0.275**	**0.421**	**0.226**	**0.447**



(a) Researcher Classification (b) Collaborating Prediction

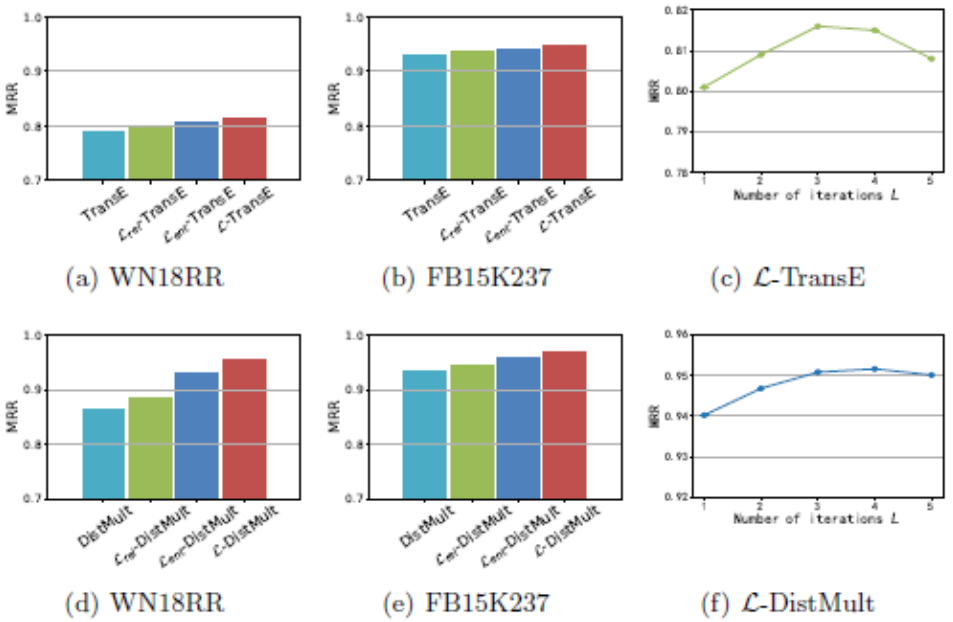
Ziyue Qiao, Pengyang Wang, Yanjie Fu, Yi Du*, Pengfei Wang, Yuanchun Zhou. Tree Structure-Aware Graph Representation Learning via Integrated Hierarchical Aggregation and Relational Metric Learning. The 38th IEEE International Conference on Data Engineering (ICDE 2022). Under review.

上下文：融合实体上下文和关系上下文的表示学习方法

中国科学院 计算机网络信息中心 Computer Network Information Center, Chinese Academy of Sciences

模型在多个通用知识图谱数据集中进行验证，并在多个评价指标下取得了较好的效果。消融实验验证了两种上下文的有效性。

Method	WN18RR			FB15K-237			NELL995			DDB14		
	MRR	MR↓	Hit@3	MRR	MR↓	Hit@3	MRR	MR↓	Hit@3	MRR	MR↓	Hit@3
ComplEx	0.840	2.053	0.880	0.924	1.494	0.970	0.703	23.040	0.765	0.953	1.287	0.968
SimplE	0.730	3.259	0.755	**0.971**	1.407	0.987	0.716	26.120	0.748	0.924	1.540	0.948
RotatE	0.799	2.284	0.823	0.970	1.315	0.980	0.729	23.894	0.756	0.953	1.281	0.964
RGCN	0.823	2.144	0.854	0.954	1.498	0.973	0.731	22.917	0.749	0.951	1.278	0.965
TransE	0.789	1.755	0.918	0.932	1.979	0.952	0.719	16.654	0.766	0.936	1.487	0.957
$\mathcal{L}$-TransE	0.813	1.648	0.933	0.943	2.281	0.962	0.793	9.325	0.831	0.964	1.184	0.969
DistMult	0.865	1.743	0.922	0.935	1.920	0.979	0.712	22.340	0.744	0.937	1.334	0.958
$\mathcal{L}$-DistMult	**0.955**	**1.134**	**0.988**	0.967	**1.174**	**0.988**	**0.852**	**2.271**	**0.914**	**0.972**	**1.097**	**0.991**



(a) WN18RR (b) FB15K237 (c) $\mathcal{L}$-TransE

(d) WN18RR (e) FB15K237 (f) $\mathcal{L}$-DistMult

Fig. 3. The performance of model variants for (a)$\mathcal{L}$-TransE and (d)$\mathcal{L}$-DistMult on WN18RR dataset. The performance of model variants for (b)$\mathcal{L}$-TransE and (e)$\mathcal{L}$-DistMult on FB15K237 dataset. The performance of various L for (c)$\mathcal{L}$-TransE and (f)$\mathcal{L}$-DistMult on WN18RR dataset.

Ning, Zhiyuan, Ziyue Qiao, Hao Dong, Yi Du, Yuanchun Zhou. “LightCAKE: A Lightweight Framework for Context-Aware Knowledge Graph Embedding.” Pacific-Asia Conference on Knowledge Discovery and Data Mining(PAKDD), 2021,.

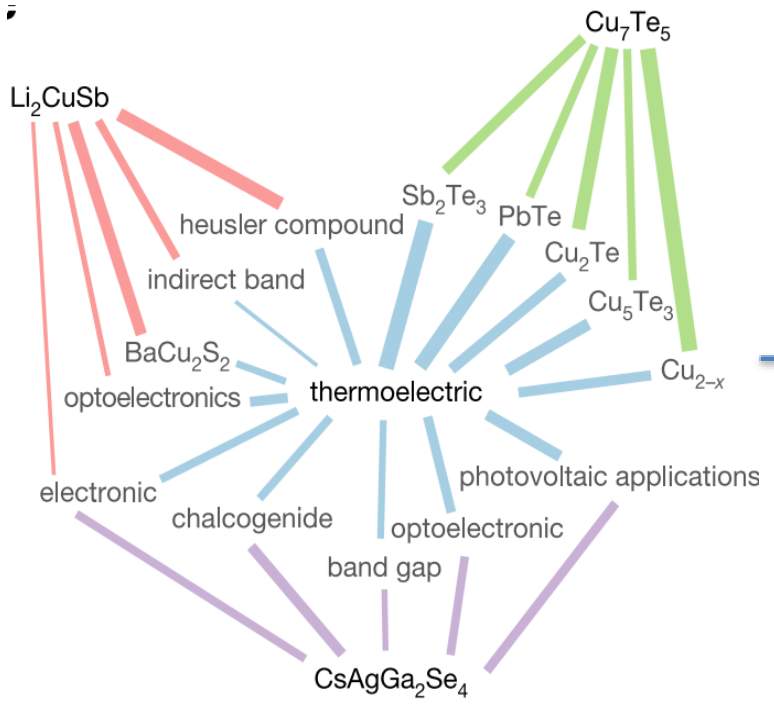
Knowledge Graph for Science

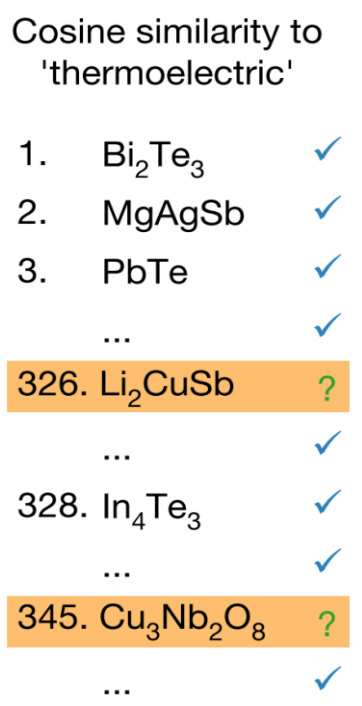
中国科学院
计算机网络信息中心
Computer Network Information Center,
Chinese Academy of Sciences

知识图谱在辅助科学发现中展现出潜力

建图谱 → 用图谱 → 拓图谱

知识图谱在解决以科技管理、军事决策为代表的管理决策问题上，发挥了重要作用



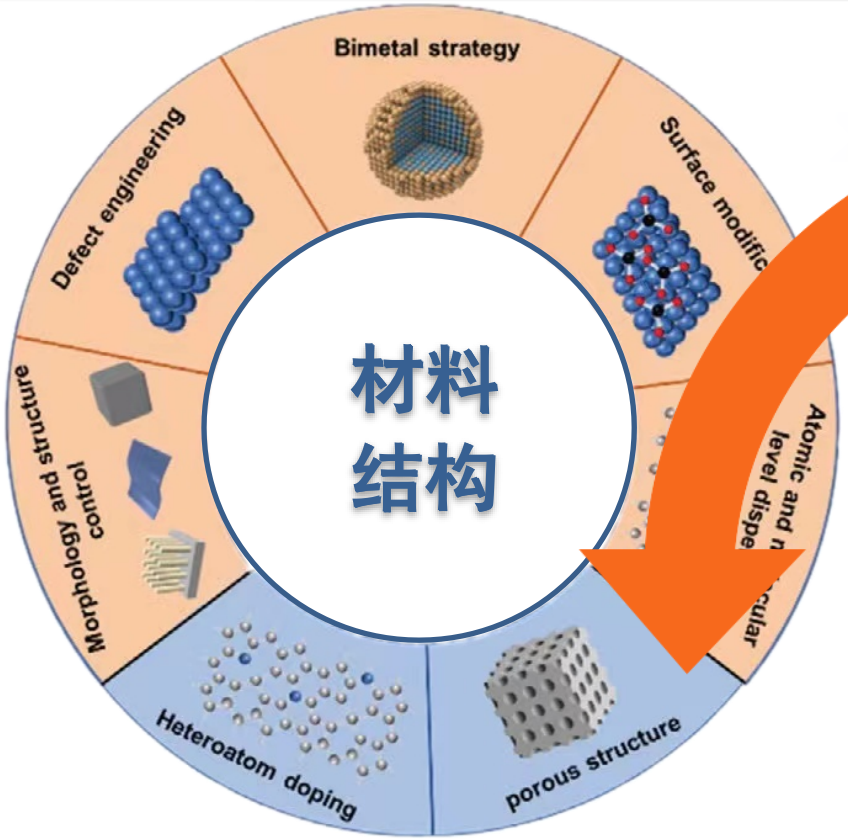


从材料领域文献/专利数据中抽取材料及性质，构建材料-性质关联知识图谱，实现新热电性质材料的预测

Nature 2019,571;Science 2020,370； Nature Comm. 2021,2573

契机:基于领域知识图谱的光电催化材料挖掘软件

中国科学院
计算机网络信息中心
Computer Network Information Center,
Chinese Academy of Sciences



计算/测量

科技文献数据
模拟计算数据

材料性质
价带结构、载流子浓度 … …

测量

材料功能
选择性、量子效率 … …

发现

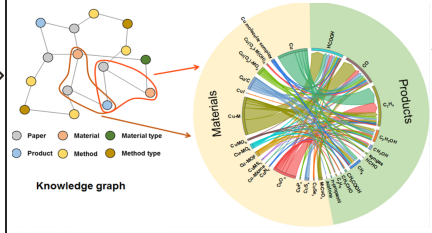
- 数据库：ICSD(无机晶体结构/德国)、MAGNDATA(磁结构/西班牙)、AFLOW(金属材料/美国)、Springer Material(通用/德国)、Material Project(电池材料/美国)
- 软件：VASP(商业/奥地利)、SM Search(商业/德国)、Materials Studio(商业/美国)、CP2K(德国)

根据材料功能直接预测材料的结构

工作特色

1.大模型与领域知识图谱融合研究

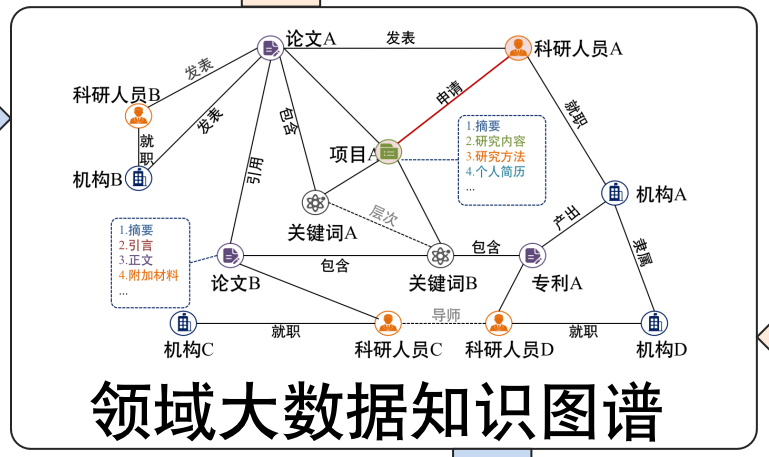
辅助多模态信息抽取
(知识体系指导)



文本、图像
表格、3D
序列、光谱

多模态大数据

知识图谱增强
(多模态表示与融合)



领域大数据知识图谱

辅助复杂知识抽取
(专家在回路+Prompt)

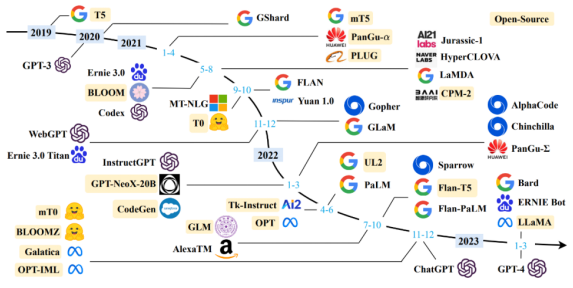


Fig. 1. A timeline of existing large language models (having a size larger than 10B) in recent years. We mark the open-source LLMs in yellow color.

大语言模型(LLM)

高质量知识支撑
(Fine-tuning+Prompt)

2.利用多模态数据构建增强型知识图谱

请批评指正，谢谢！

中国科学院
计算机网络信息中心
Computer Network Information Center,
Chinese Academy of Sciences