大数据领域知识图谱——项目与技术

杜一

中国科学院计算机网络信息中心

主要内容

- 国家自然科学基金大数据知识管理服务平台
- 科技领域知识图谱的横向应用
- 科技领域知识图谱的纵向深挖
- 基于大数据知识工程的工具积累

国家自然科学基金大数据知识管理服务平台

项目背景 – 业务系统现状

网络信息系统

基础研究知识库

共享服务网

成果之友

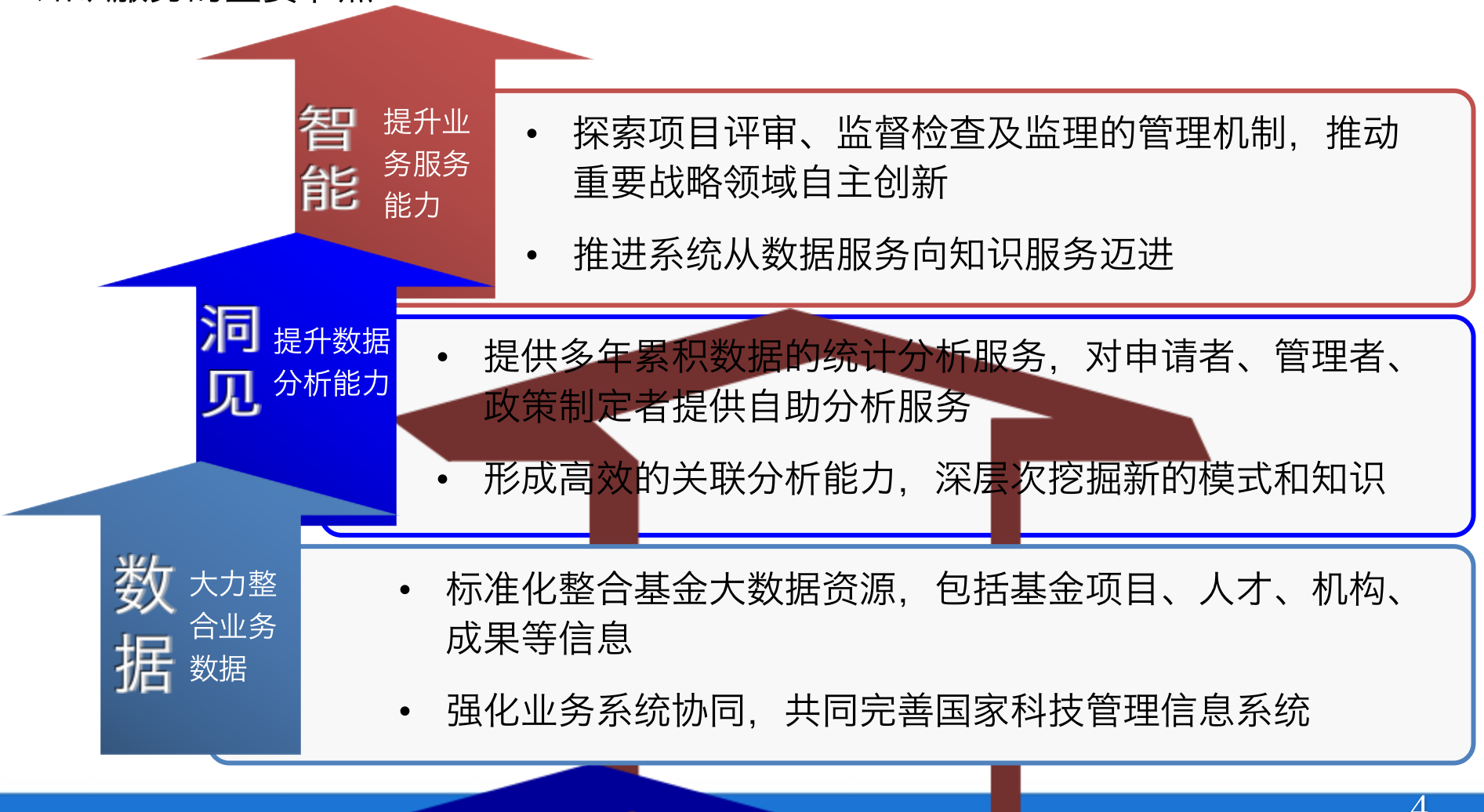
数据量增大，原有数据库无法满足存储、计算需求

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）

国家自然科学基金大数据知识管理服务平台

目标：遵循大数据、语义网、关联数据等国际标准，建设成为国内领先、国际知名的科研管理信息大数据中心（数据完备、分析高效、知识丰富），成为科研管理信息资源和知识服务的重要节点



国家自然科学基金大数据知识管理服务平台

应用系统

数据服务

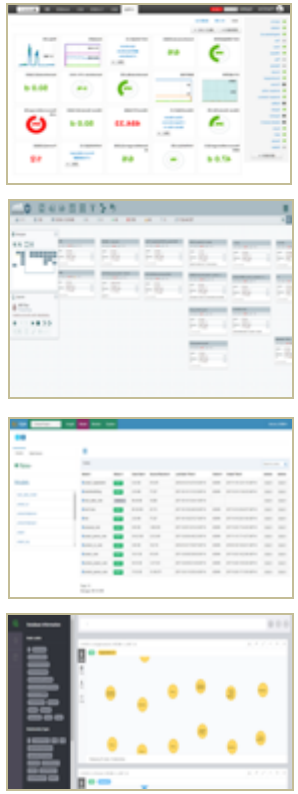
多维即时分析统计　网络查询与分析　全文检索

数据模型

人员　项目　成果　机构

数据源

ISIS业务系统数据　ISIS成果库　专家Profile　Or系统　DBLP　Scopus　CSCD　Springer

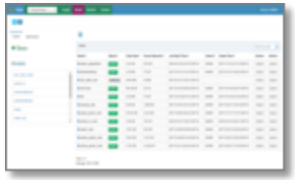


大数据管理

建设成效

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，实现流水线汇聚路径，形成异构融合、面向服务的数据湖。

一批应用

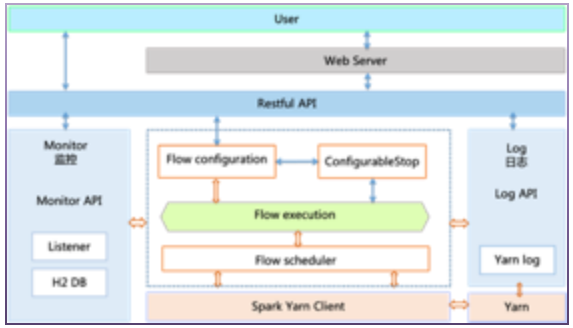
基于统一的服务接口，建成知识服务门户、多维统计、网络挖掘、交叉预测、专家画像等示范应用，形成了“口同径、数同源、能洞察”的大数据应用新能力。





关键技术

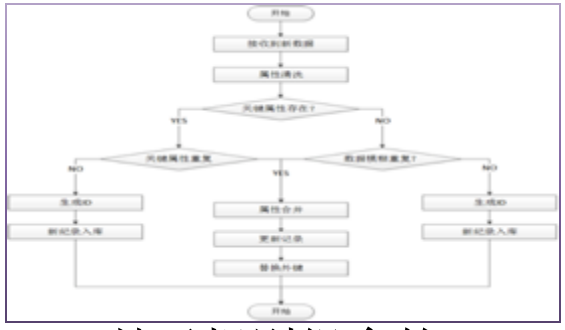
自动及半自动的数据采集、处理流水线机制



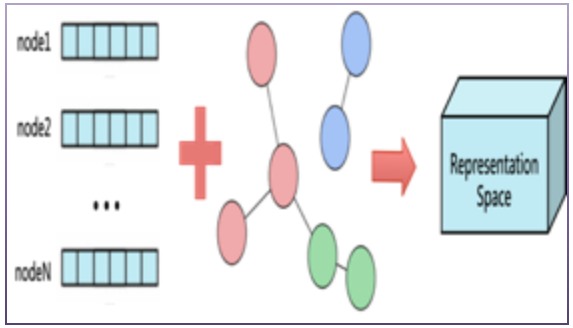
PiFlow 架构

流水线实例

基于统计规则及深度学习方法的实体融合方法

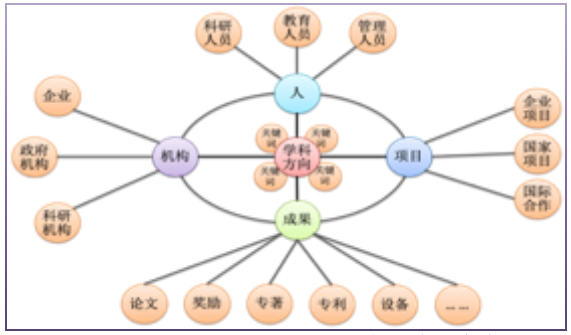


基于规则组合的实体融合

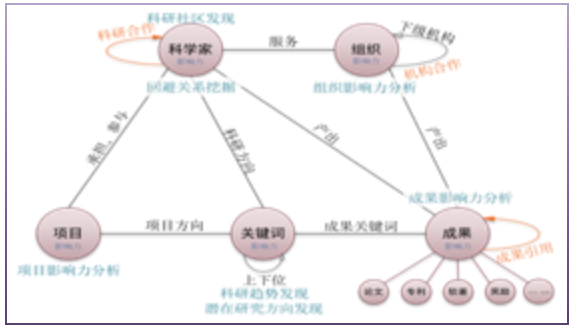


基于GraphEmbedding的实体融合

基于属性图的科技领域知识图谱构建方法



基于规则的科技领域知识图谱

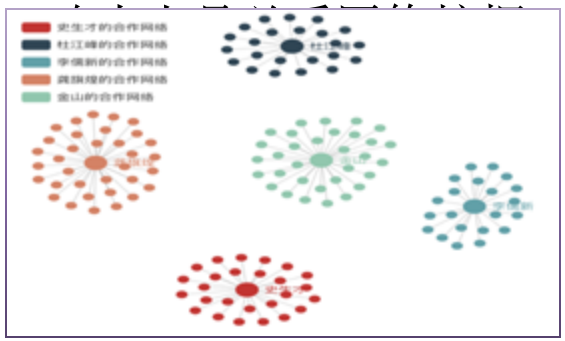


基于推理学习的科技领域知识图谱

关键技术

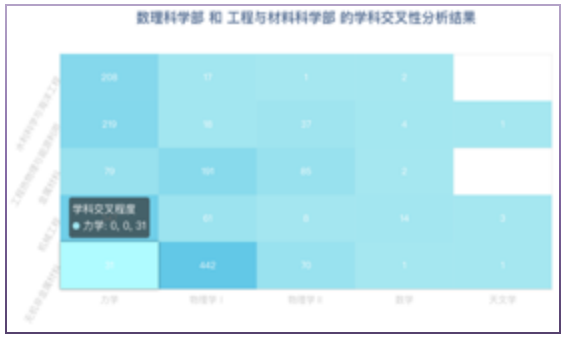
基于学科的科研社区发现算法

基于PageRank+SVD的高影



基于网络度量指标+标签传播的科研社区发现

学科交叉性评估及学科趋势预测方法



基于统计指标的学科交叉性评价

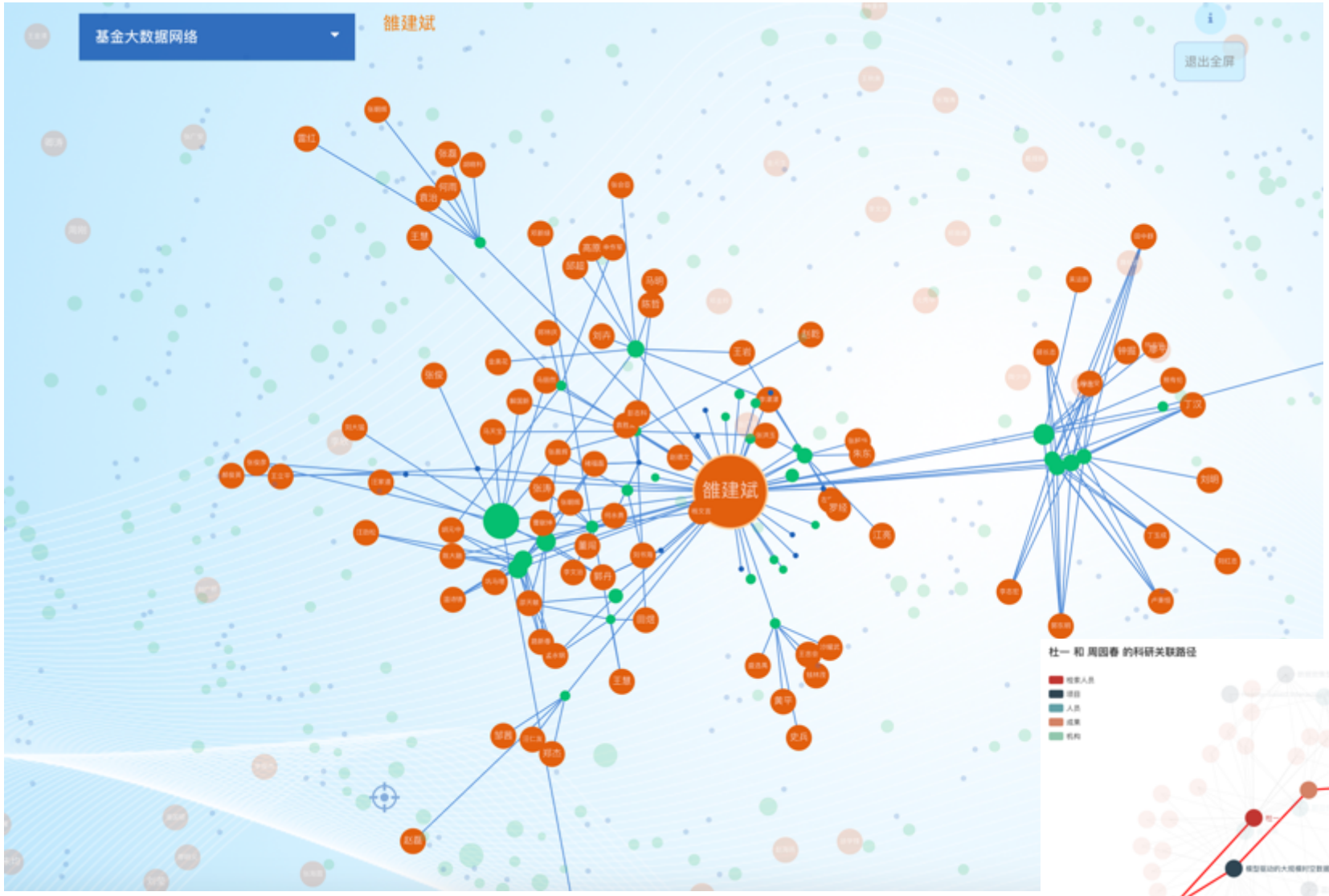
基于时序+图数据库的学科趋势评估

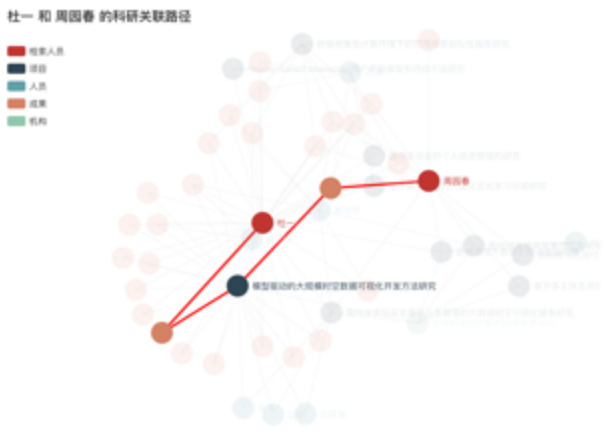
学术影响力评价方法

基于网络度量指标+计量学的科研人员影响力评价

基于网络度量指标+计量学的科研项目影响力评价

演示:大数据知识管理服务门户

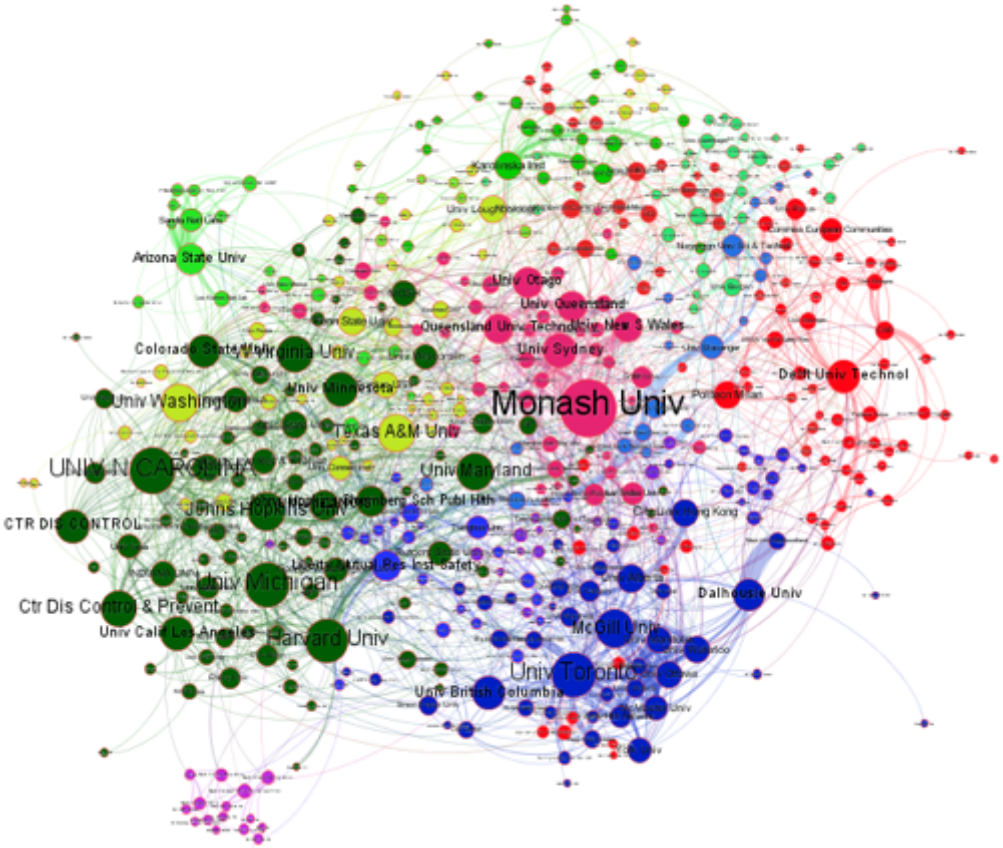




科技领域知识图谱的横向应用

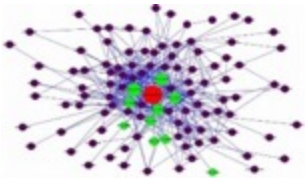
烟草科技领域知识图谱——总体需求

采用大数据知识加工、知识挖掘、精准画像等知识图谱技术，构建覆盖烟草“**科研机构、科技人员、科技项目、科技成果**”等核心实体的**领域科技资源图谱**，形成面向烟草科研活动的知识服务平台与能力。

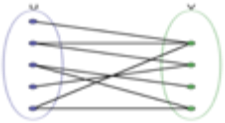


最短路径

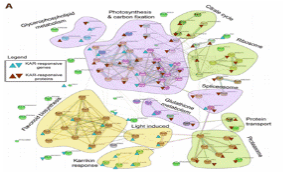
连通分析



关键节点



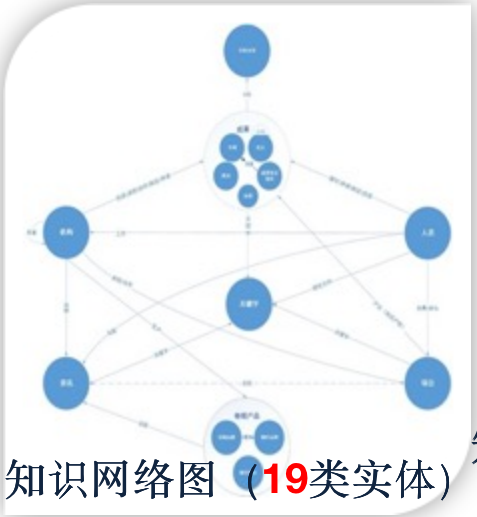
二分图



群团聚类

烟草科技领域知识图谱——建设成效

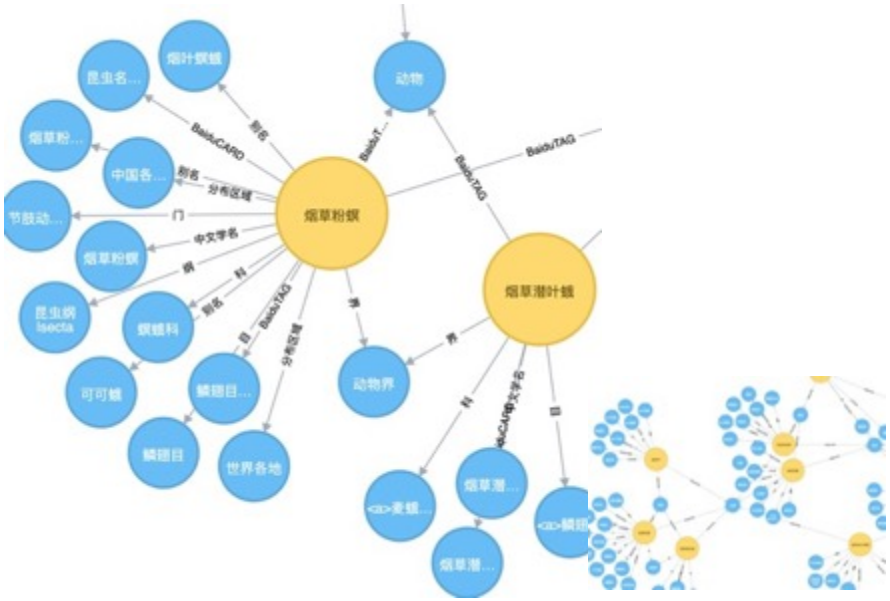
构建“一张”知识图谱



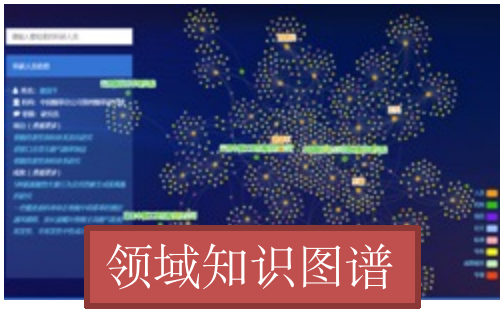
知识网络图（19类实体）

知识实体关系（33类关系）

突破领域内、外数据关联



建立“一批”知识示范功能



领域知识图谱

影响力分析

专家画像

社区与热点

中国科协计算机与人工智能大数据知识管理服务平台

- 面向计算机与人工智能领域，构建中国科协知识管理平台，为管理决策提供服务

- 自动及半自动汇聚5亿+海量数据
- 汇聚科学家、项目、论文、成果实体
- 实现人员画像、学科趋势分析等智能服务

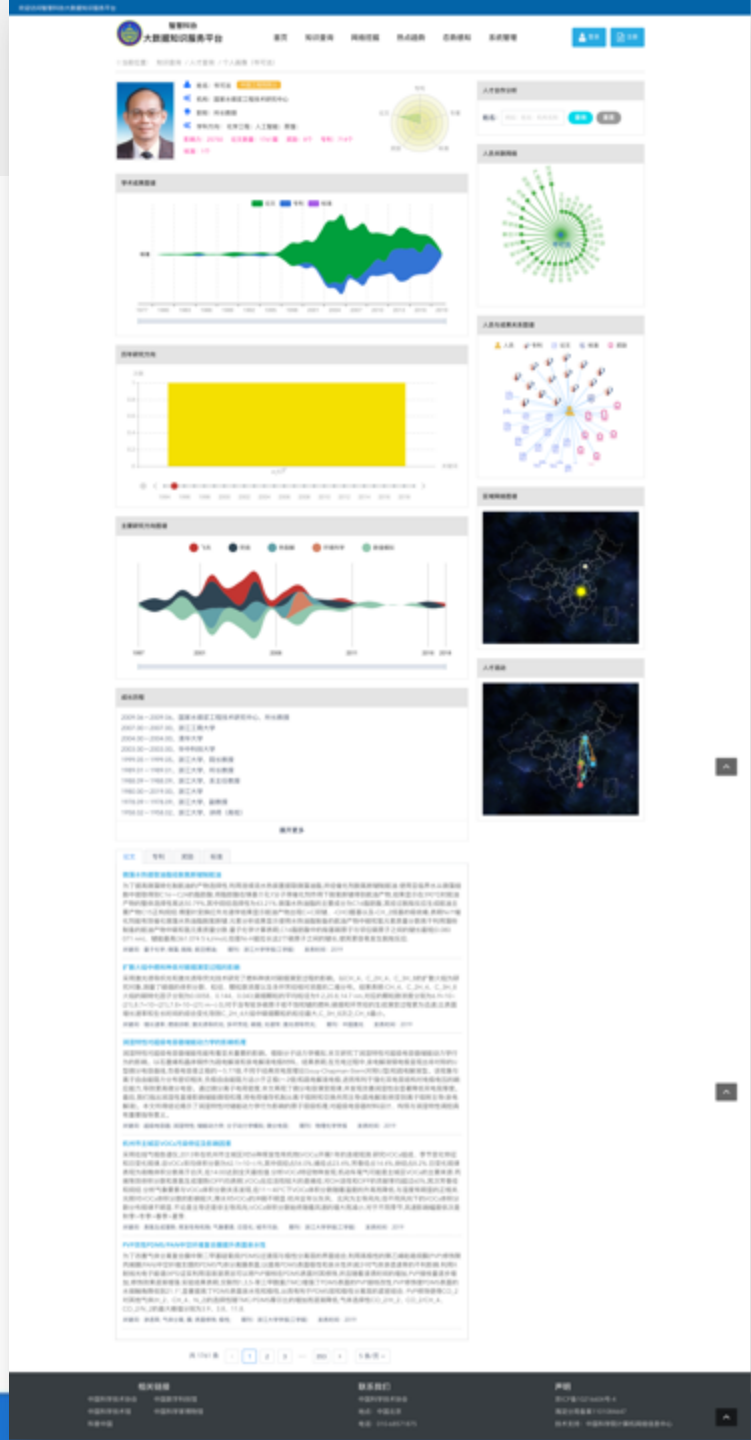
中国科学技术协会
China Association for Science and Technology

中国科协计算机与人工智能大数据知识管理服务平台

态势感知系统（八大实体）

中国科协计算机与人工智能大数据知识管理服务平台

中国科协计算机与人工智能大数据知识管理服务平台



空间科学领域知识图谱与决策支持

其它领域扩展

- 医疗（致病菌、病例）
- 教育（知识点图谱）
- 工业互联网（工业设备）
- 军事（武器装备）

科技领域知识图谱的纵向深挖

面向领域大数据的知识图谱构建

- 国家自然科学基金重点项目

实证研究

4： 面向科技、POI领域大数据知识图谱的实证

理论研究

高效计算

3： 分布式知识图谱数据管理系统关键技术

知识推理

2： 面向领域大数据的知识图谱推理与认知技术

知识表征

1： 面向领域大数据的知识图谱表示与构建技术

面向领域大数据的知识图谱表示与构建技术

- 基于科技论文的异质网络表示学习及应用：架构

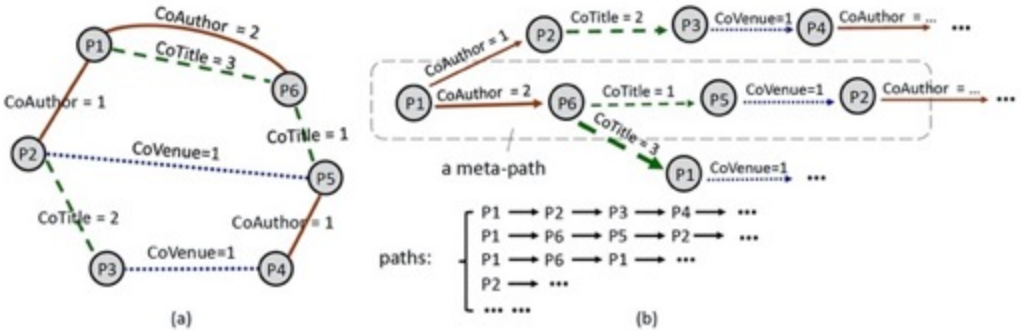


Fig. 1. An illustration of (a) a publication heterogeneous network with one node type, three relation types, and (b) paths sampled by meta-path ($r = CoAuthor \circ CoTitle \circ CoVenue$) and relation weight guided random walks.

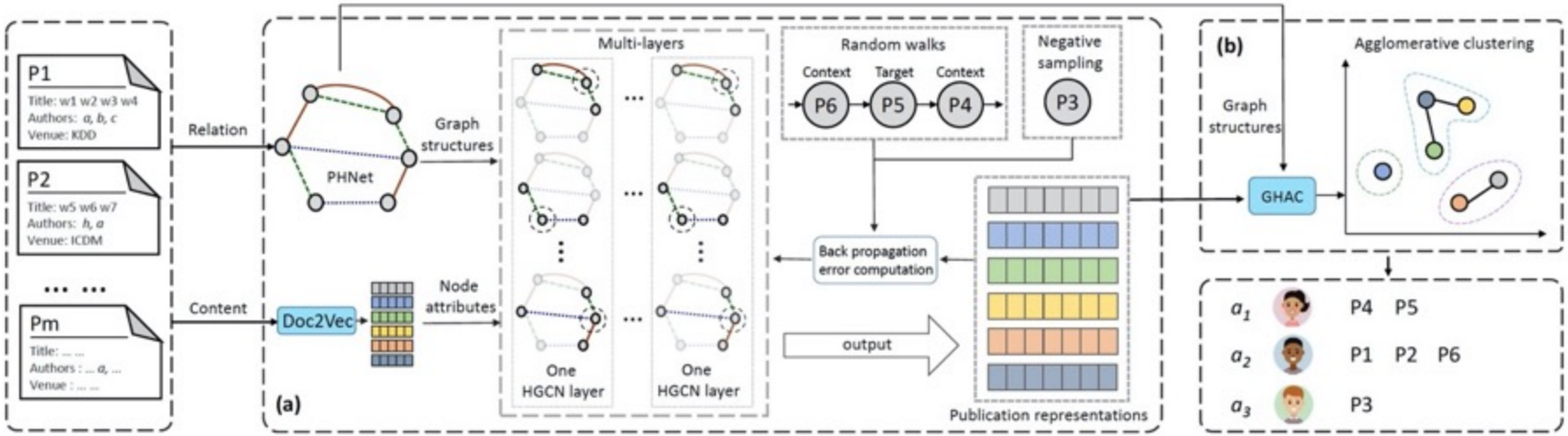


Fig. 2. An illustration of disambiguating the author name a. This framework consists of (a) the publication heterogeneous network embedding method to learn publication representations and (b) the graph-enhanced hierarchical agglomerative clustering method to partition the publications.

面向领域大数据的知识图谱表示与构建技术

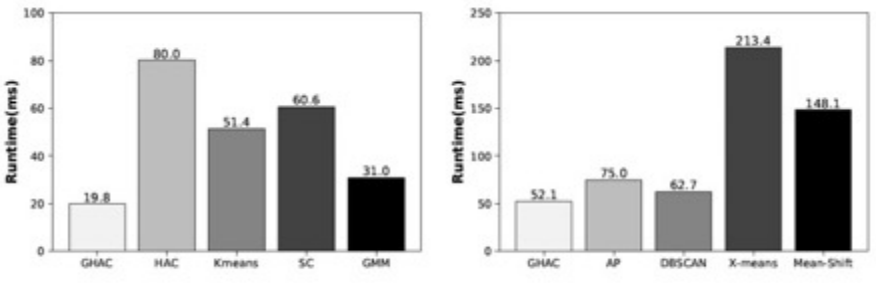
- 基于科技论文的异质网络表示学习及应用：性能

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

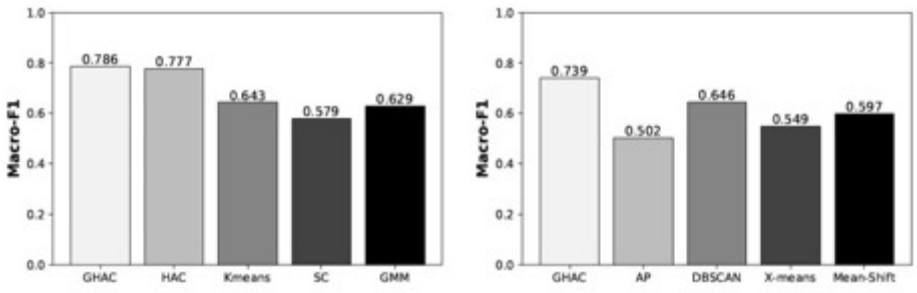
TABLE II
THE PERFORMANCE OF DIFFERENT METHODS ON CITESEEX DISAMBIGUATION DATASET

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) K is known (b) K is unknown

Fig. 5. The average running time of different clustering methods on Aminer dataset.

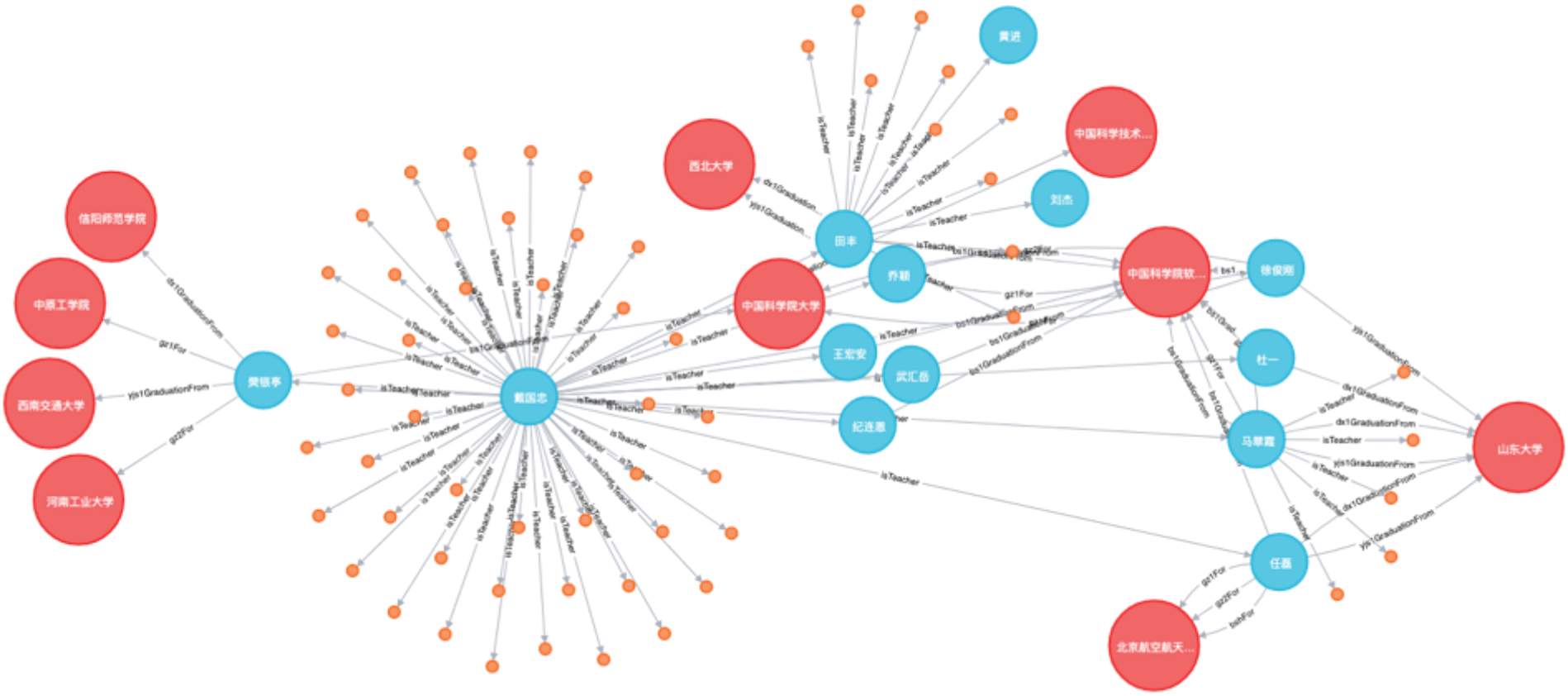


(a) K is known (b) K is unknown

Fig. 6. The performance of different clustering methods on Aminer dataset.

面向领域大数据的知识图谱推理与认知技术

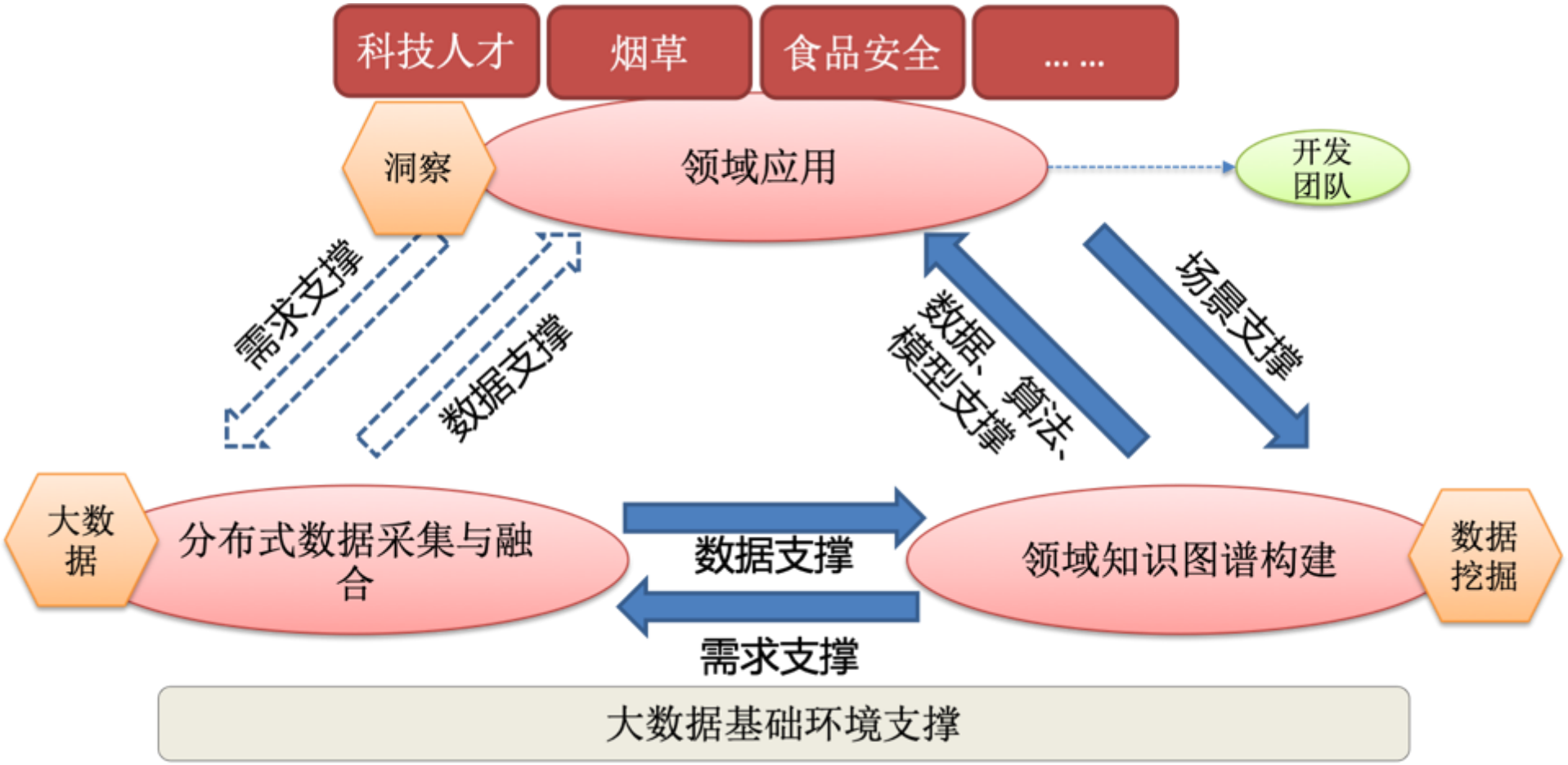
基于网络拓扑方法及表示学习方法的人员关系推理



基于大数据知识工程的工具积累

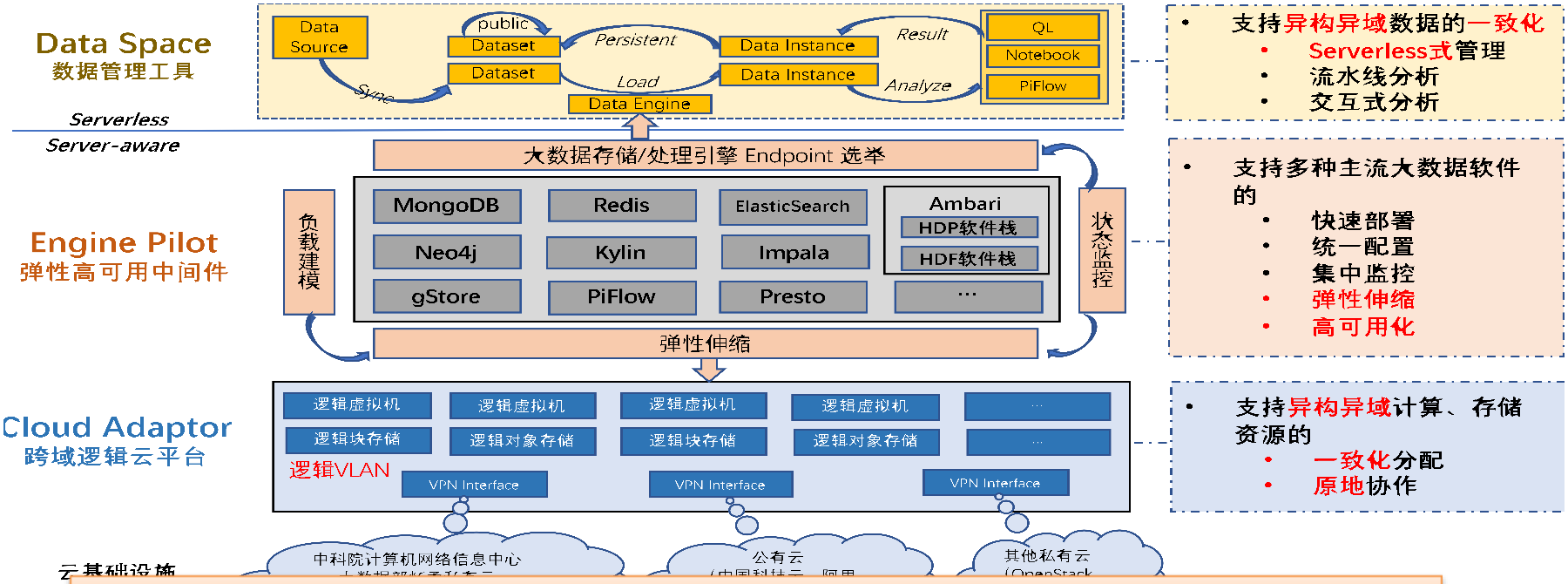
大数据工具体系架构

- 从基础环境、存储，到数据采集、分析与可视化的一站式工具



PackOne：主流大数据软件在云端的快速弹性部署

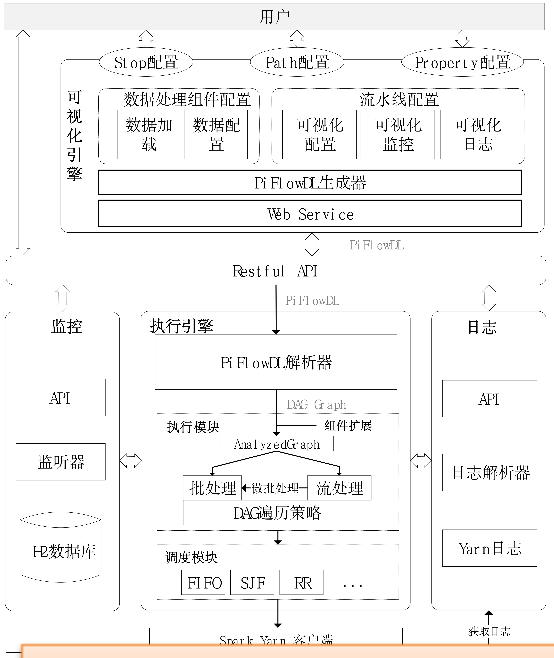
- Hadoop、Spark、NiFi、PiFlow、Kylin、MangoDB、Neo4J等流行的大数据管理/处理软件在云端的一键部署和一键伸缩

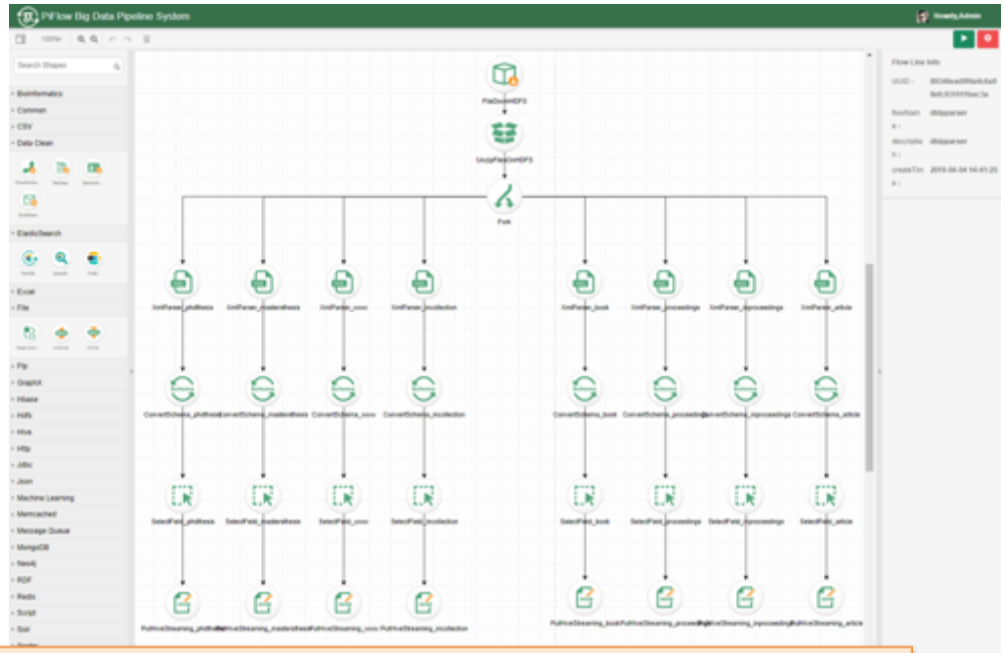


开源平台：STAR50+ FORK10+，支持科学大数据、基金委等相关项目

PiFlow：大数据采集与处理流水线平台

- 模型驱动的理论支持
- 基于Spark分布式存储与并行计算技术
- 内嵌100+数据处理组件
- 灵活的可扩展性





开源平台：STAR300+ FORK100+，在基金委、空间中心等项目中应用

请批评指正！