AI TIME
Artificial Intelligence Time

智谱·AI

开题？

发Paper?

寻找解题思路？

作报告？

毕业论文？

如何做好学术检索

分享人：仇瑜

清华大学知识工程实验室

2021年5月

AI TIME
Artificial Intelligence Time

智谱·AI

目 录

为什么要进行学术检索?

“如果说我比别人看得更远些，那是因为我站在了巨人的肩膀上。” —— 牛顿

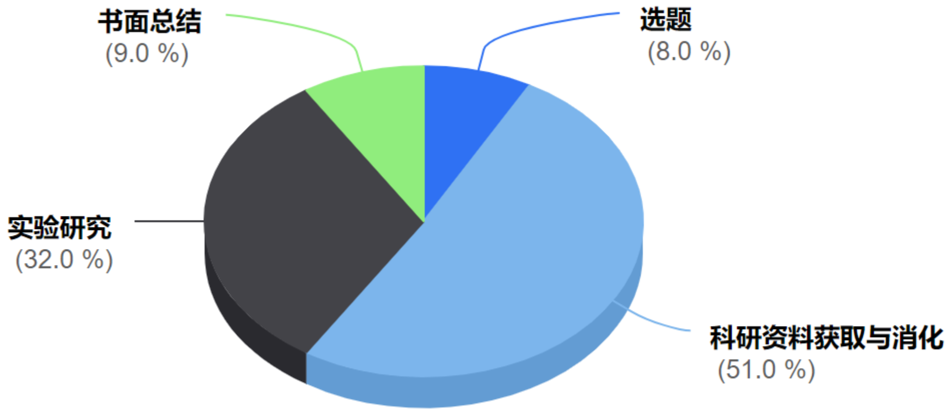
“为学之道，莫先于穷理；穷理之要，必先于读书。 ” —— 朱熹

A **literature search** is a well thought out, organized search and **evaluation** of literature available on a topic.

- 了解领域研究概况，科研选题
- 发现研究方法，收集实验资料
- 了解最新进展，扩展研究思路
- 科技查新，避免重复工作

A well-structured literature search is the most effective and efficient way to **locate sound evidence on the subject** you are researching.

科研人员时间分配

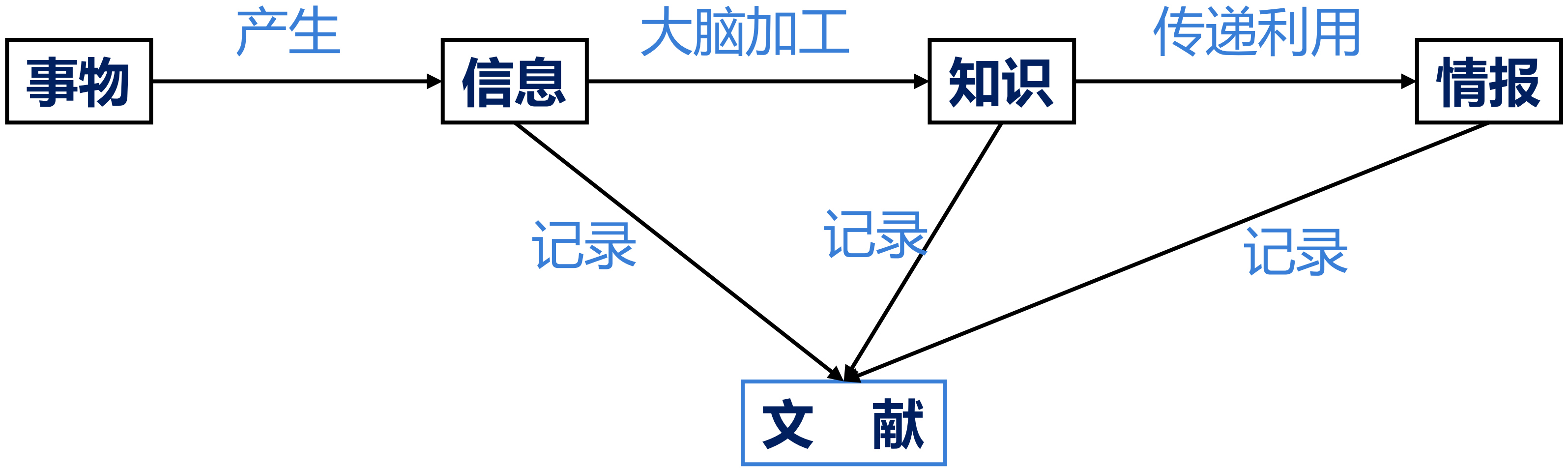


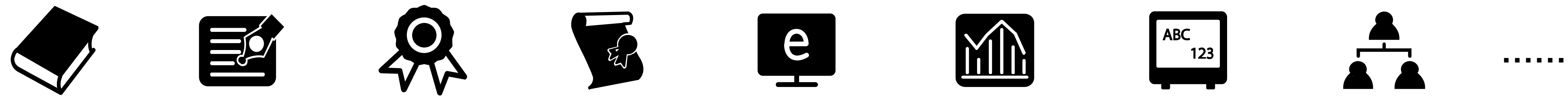
美国科学基金会（National Science Foundation）统计

学术检索是科学研究中一种学习的方法，一种解决问题的方法

学术资源产生

- **信息**广泛存在于自然界和人类社会。
- **知识**是人类在改造世界的过程中所获得的认识和经验的总和。
- **情报**是人们在一定时间内为一定目的而传递的有使用价值的知识或信息。
- **文献**是记录知识或信息的一切载体。





图书　论文　专利　学术报告　新闻资讯　数据集　源代码　专家　……

学术资源分类和用途

知识产生

记录

原始文献

知识重组

知识浓缩

词典、名录、指南、手册、年鉴等

综述、专著、教科书、论文丛集、百科全书等

期刊论文
学术论文
科技报告
学位论文
专利说明书

文献整理

书目、索引、文摘、检索该工具等

三次文献

选用大量有关的文献，经过综合、分析、研究而编写出来的文献。

二次文献

文献工作者对一次文献进行加工、提炼和压缩后的产物。

零次文献

口头信息、未公开发表的文字资料。

一次文献

以研究成果为基本素材创作或撰写，研究人员最终索取的文献。

可以由二次文献作为入口，满足整体的需求；然后，通过入口来获取有用的高质量的全文（一、三次文献）来满足纵深的研究需要。

学术检索策略

科学研究工作是一个不断往复、螺旋式展开的过程。

引自清华大学图书馆钱俊雯

导师给定方向后泛调研

- 收集领域综述文献、图书、博士学位论文；重点利用本领域经典或综述文集数据库；
- 重点阅读英文综或研究论文标题、摘要；了解前沿、难点、创新点、并收集关键词；
- 确定研究题目（大课题研究背景+当前研究热点+自身兴趣点）；

确定研究题目后精调研

- 有针对性的收集文献，重点在于确定内容，了解技术主要研究者和机构；
- 文献阅读（泛读和精读相结合）；
- 确定课题实施方案（技术和方法的创新）

学术检索发展历程

学术搜索发展历程与计算机和网络技术发展密切相关。

文献资源	纸质文献	电子文献	数字图书馆	开放多源集成数据库	开放互联网学术资源
文献检索	人工查找	人工分类目录	关键词搜索 自建数据库	开放学术搜索	智能语义搜索
搜索技术	无	主题标引+ 传统信息检索	目录门户 自动聚类 内容分析技术 数字图书馆技术	爬虫技术 特征提取 对象实体识别 文本结构化	知识图谱 可视化技术 智能推荐 技术趋势分析
文献计量	人工计量	简单计量 (按年份排序)	指数化和模型化计量 (影响因子、h指数)	可视化和智能化阶段 (显示领域论文量年度 分布、关键词共现网络)	

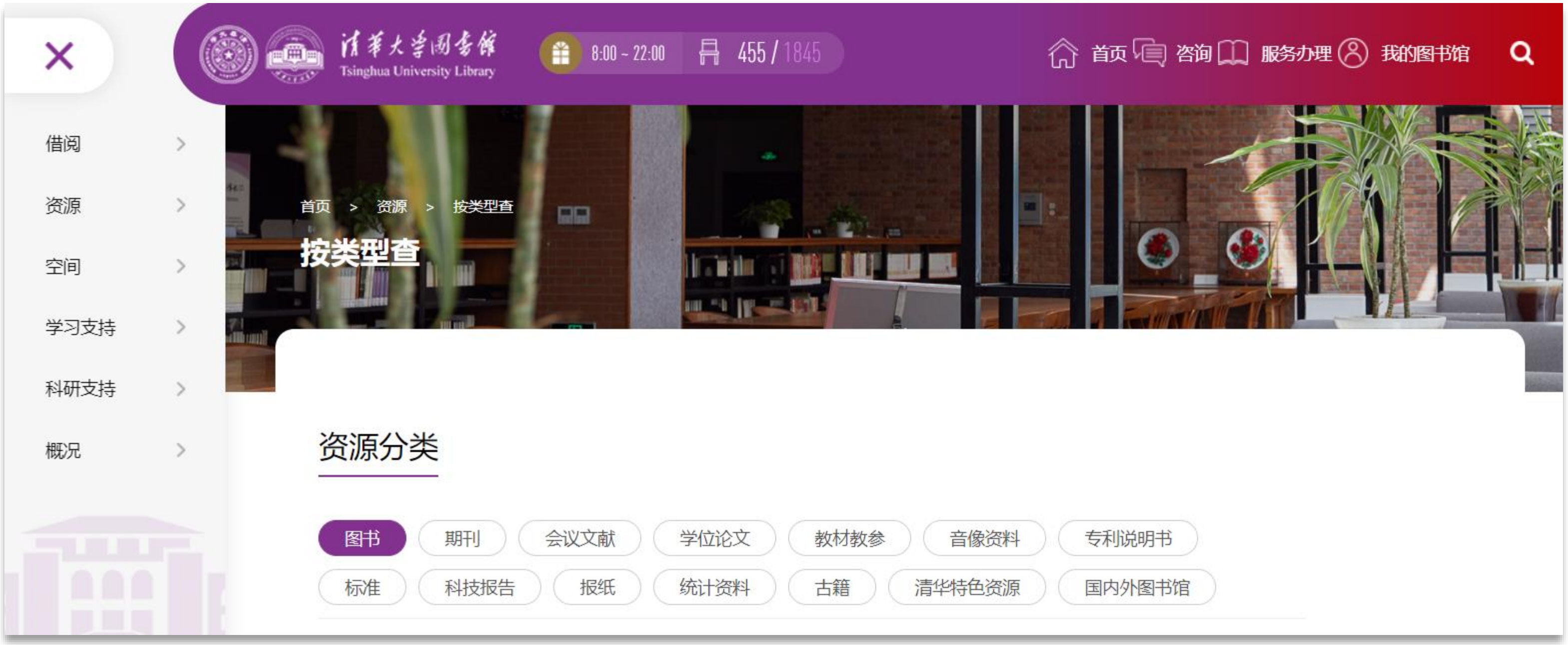
1980 PC Era 1990 Web 1.0 2000 Web 2.0 2010 Web 3.0 2020

引自AMiner 《人工智能之学术搜索》

学术资源获取途径

智谱·AI

- **图书馆**
  - 授权数据库
  - 馆藏图书/电子书
  - 多媒体资源
  - 特色领域资源
  - **科研数据**
- **网络学术资源**
  - 专业数据库
  - 搜索引擎
  - 门户、论坛
  - 电商
- **老师、师兄师姐**
  - 学术社交圈



学术检索引擎	社交网络	学会/协会网站
学术论坛	开放存取资源	电子商务网站
学科信息门户	学者主页博客	……

学术资源获取途径

□ 综合性论坛

- 科学网：http://www.sciencenet.cn/
- 小木虫：http://muchong.com/bbs/
- 科研速递：http://www.expaper.cn/

□ 专业性论坛

- CCF：https://www.ccf.org.cn/
- 丁香园：https://portal.dxy.cn/
- 经管之家：https://bbs.pinggu.org/

□ 其他

- 国家数字图书馆：http://mylib.nlc.cn
- 工程院知识中心：http://www.ckcest.cn/
- 学习强国：https://www.xuexi.cn/

中华人民共和国国民经济和社会发展第十四个五年规划和2035年远景目标纲要

AI TIME
Artificial Intelligence Time

智谱·AI

目 录

学术检索步骤

学术检索是情报获取的过程，查找、评估和有效利用需要的信息来解决实际问题或者做出决策。

- 分析课题、明确检索目的
- 选择检索工具，确定检索方法
- 查找文献线索
- 获取原始文献

Step 01

分析主题，提取关键词

人工智能，知识图谱

Artificial intelligence,

knowledge graph

Step 02

扩充同义词，上下位词

AI，知识工程，语义网

Knowledge engineering semantic web

Step 03

选择工具/数据库检索文献

逻辑运算符，检索式等

Step 04

检索结果利用

多种属性筛选

优化检索词

文献管理

文献检索工具

将所有收录文献按一定规则编制成的具有存储、检索和报道功能的工具。

通用搜索引擎

谷歌、百度、必应等

- 信息覆盖范围广，类型全面；
- 检索结果干扰项较多；
- 适用于课题调研阶段；
- 能反映领域的研究关注度；

文摘型数据库

AMiner、谷歌学术、SCI、EI等

- 本身无全文，但对文献有深层次加工；
- 适用于文献系统调研和分析，学术评价；
- 收录范围广、数据量大、有连续性；
- 能反映领域的学术进展；

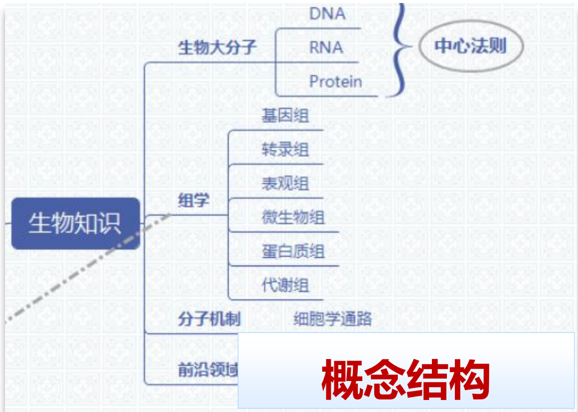
全文数据库

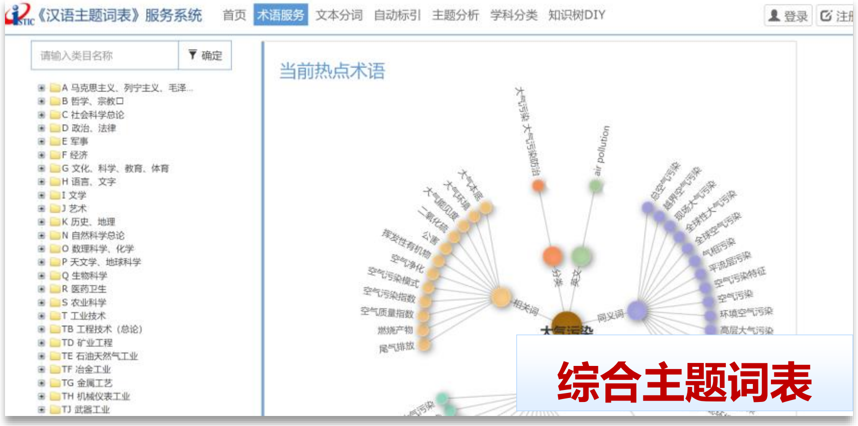
Springer、IEEE、Nature等

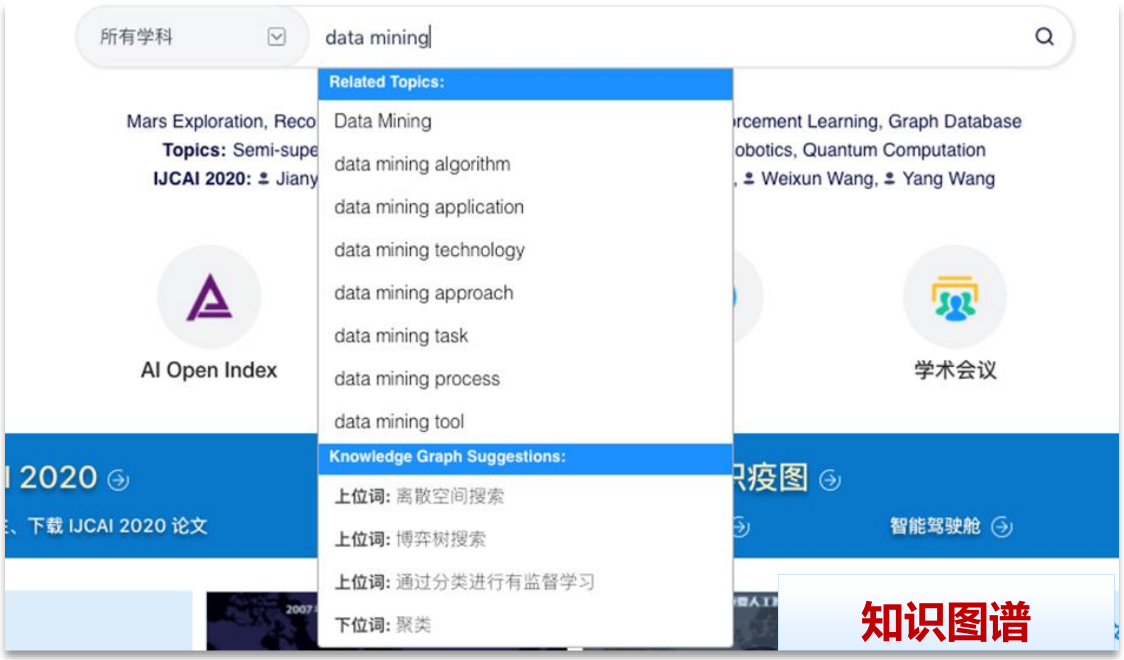
- 提供原始原文信息；
- 文献更新速度快，时效性强；
- 收录时间范围一般小于文摘库；
- 不具备较强的文献分析功能；

主题关键词确定

根据主题词表/知识图谱/领域词典等，优化扩展搜索关键词。







检索技巧

不断试错的过程， 在检索结果的基础上逐渐优化，最终得到相对理想。

逻辑运算符	精确匹配	截词检索
AND OR NOT 改变运算优先级（）	“” 表示精确匹配，或固定短语，过滤相关度不高的文献	截词符*可代替检索词中的一个或多个字母：词根、单复数、变形等

- **site:限定特定域名站点中检索**
  artificial intelligence site:nature.com
- **filetype:限定检索结果为某一类型文件**
  artificial intelligence filetype:ppt
- **inurl:返回网址中包含关键字的页面**
  artificial intelligence inurl:ai
- **Intitle:返回标题包含指定关键词的结果**
  artificial intelligence intitle:国家

- **link:搜索连接到某URL的网页**
  link:arxiv.org/abs/1810.04805
- **related:列出于目标URL相似的网页**
  related:arxiv.org/abs/1810.04805
- **intext:寻找特定网页里的关键字**
  intext:人工智能
- **define:查询关键词的词义**
  define: artificial intelligence

建立领域知识库

对某一科学领域进行全面、客观的梳理，建立自己的知识体系。

□ 综述

提供这一学科系统的、周期性的最新研究状况，不仅总结，还找出错误来引起讨论，形成新的研究。

- 领域最重要的文献，这比任何搜索引擎更有效。
- 从综述文献开始，是科研的最高起点……

□ 图书、硕博士论文

目录梳理了领域技术分类及发展现状。

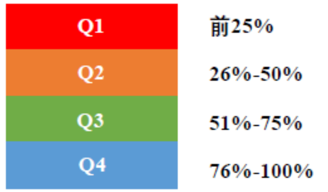
□ 梳理领域重要关注对象

- 文献发表可能存在“二八定律”，关注**学科带头人**及相关课题组；
- 除了关键词检索外，也可根据已找到或导师推荐的文献**作者或机构**检索，发现研究问题的延续；
- 关注高质量**期刊/会议**。

JCR分区：

✓ Clarivate发布

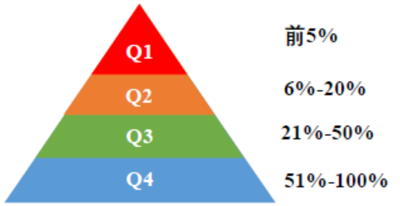
✓ 按照期刊当年的影响因子，将SCI&SSCI期刊分为4个分区：Q1> Q2> Q3> Q4



中科院分区：

✓ 中科院发布

✓ 按照期刊3年影响因子的平均值，将SCI&SSCI期刊分为4个分区：Q1> Q2> Q3> Q4



SCI | **EI** | **中文核心**

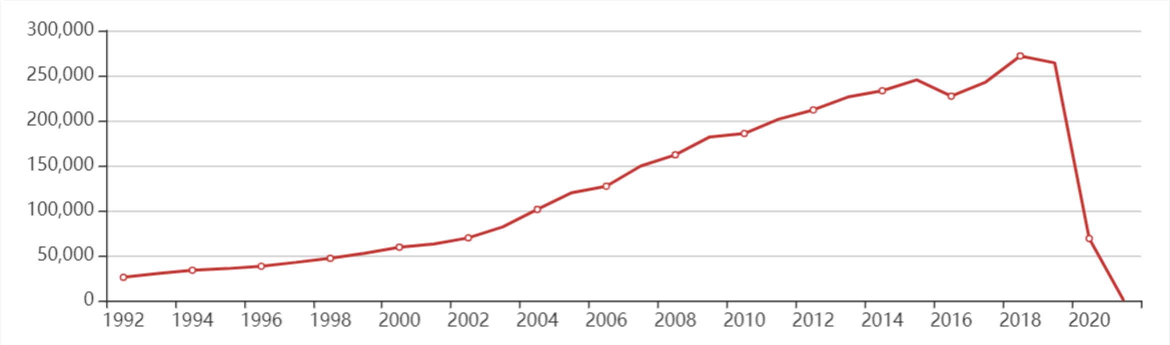
AI TIME Artificial Intelligence Time

选题调研阶段

智谱·AI

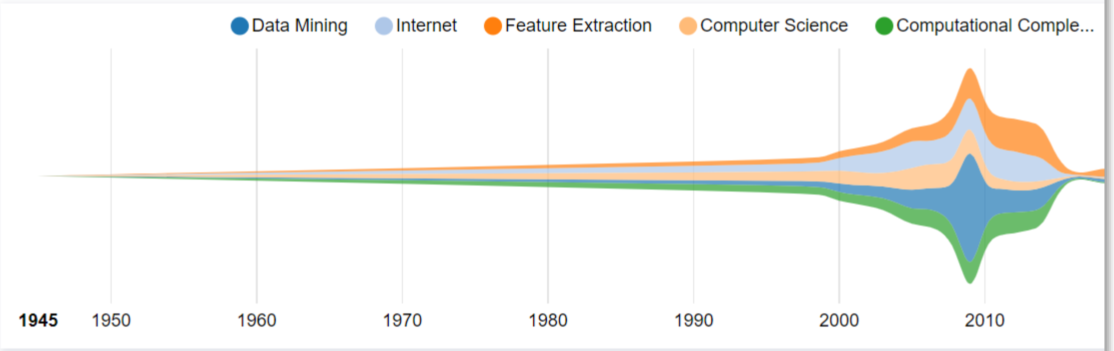
领域发展态势分析，发现热点方向、重要学者/机构、顶级期刊/会议。

发展态势



研究趋势

搜索该学科下的论文



热点主题

1. Machine Learning (3999)
2. Deep Learning (3417)
3. Feature Extraction (3229)
4. Task Analysis (2562)
5. Computational Modeling (2382)
6. Mathematical Model (2318)

概览 杰出作者 顶尖论文 热门期刊/会议

关键情报

所有时间 近 5 年

杰出作者 → 更多

1. *Jiawei Han*
2. *Anil K. Jain*
3. *Philip S. Yu*

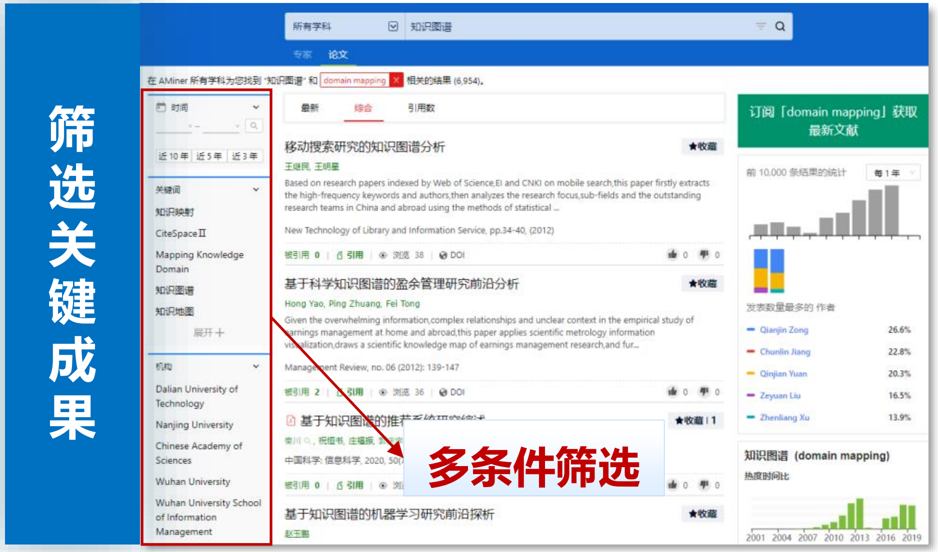
热门期刊/会议 → 更多

1. *CVPR Workshops*
2. *IEEE Transactions on Pattern Analysis and Machine Intelligence*
3. *Computer Applications in the Biosciences*

AI TIME 研究方案设计阶段

智谱·AI

快速定位某个领域的关键文献，追踪前沿技术。



研究方案设计阶段

AI TIME Artificial Intelligence Time

智谱·AI

文献精读，技术点梳理，总结技术方案。

论文精读

简介
- Unsupervised pre-trained Language Representation (LR) models like BERT (Devlin et al 2018) have achieved promising results in multiple NLP tasks.
- Publicly-provided models, like BERT, GPT (Radford et al 2018), and XLNet (Yang et al 2019), who were pre-trained over open-domain corpora, act just like an ordinary people.
- Even though they can refresh the state-of-the-art of GLUE (Wang et al 2018) benchmark by learning from open-d… specific tasks, due to little knowledge connection between specific and open domain.
- Pre-training is time-consuming and computationally expensive, making it unacceptable to most users

简介

重点内容
- Unsupervised pre-trained Language Representation (LR) models like BERT (Devlin et al 2018) have achieved promising results in multiple NLP tasks
- This paper proposes a knowledge-enabled Language Representation model, namely knowledge-enabled language representation model, which is compatible with BERT and can incorporate domain knowledge without Heterogeneous Embedding Space and knowledge noise issues;
- We propose knowledge-enabled language representation model to enable language representation with knowledge graphs, achieving the capability of commonsense or domain knowledge
- Soft-position and visible matrix are adapted to control the scope of knowledge, preventing it from deviating from its original meaning
- Empirical results demonstrate that knowledge graphs is especially helpful for knowledge-driven specific-domain t… require domain experts
- knowledge-enabled language representation model is compatible with the model parameters of BERT, which mea… pre-trained BERT parameters (e.g., Google BERT, Baidu-ERNIE, etc.) on knowledge-enabled language representatio…

研究内容

方法
- The authors detail the implementation of K-BERT and its overall framework is presented in Figure 1.
- The authors denote a sentence s = {w0, w1, w2, ..., wn} as a sequence of tokens, where n is the length of this sentence.
- Each token wi is included in the vocabulary V, wi ∈ V.
- KG, denoted as K, is a collection of triples ε =, where wi and wk are the name of entities, and rj ∈ V is the relation
- All the triples are in KG, i.e., ε ∈ K

方法步骤

结论
- The authors propose K-BERT to enable language representation with knowledge graphs, achieving the capability of commonsense or domain knowledge.
- Soft-position and visible matrix are adapted to control the scope of knowledge, preventing it from deviating from its original meaning.
- Despite the challenges of HES and KN, the investigation reveals promising results on twelve open-/specific- domain NLP tasks.
- Empirical results demonstrate that KG is especially helpful for knowledge-driven specific-domain tasks and can be… experts.
- K-BERT is compatible with the model parameters of BERT, which means that users can directly adopt the available… BERT, Baidu-ERNIE, etc.) on K-BERT without pre-training by themselves

实验结论

总结
- Introduction: Unsupervised pre-trained Language Representation (LR) models like BERT (Devlin et al 2018) have achieved promising results in multiple NLP tasks.
- Publicly-provided models, like BERT, GPT (Radford et al 2018), and XLNet (Yang et al 2019), who were pre-trained over open-domain corpora, act just like an ordinary people.
- Even though they can refresh the state-of-the-art of GLUE (Wang et al 2018) benchmark by learning from open-d… specific tasks, due to little knowledge connection between specific and open domain.
- Pre-training is time-consuming and computationally expensive, making it unacceptable to most users

总结

快速了解论文

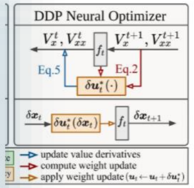
AMiner CONF | 秒读论文

一分钟了解顶会论文

Calliope Powered by Calliope

多媒体形式的论文摘要

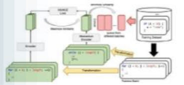
论文解读



ICLR2021论文推荐 | DDPNOpt: 最优控制理论训练方法

小脉带你读论文!

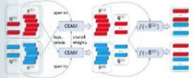
AMiner科技



ICLR2021 | 近期必读表示学习精选论文

论文可下载~

AMiner科技



ICLR 2021 | 近期必读少样本学习精选论文

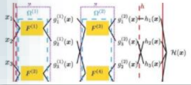
论文可直接下载!

AMiner科技

ICLR2021 | 近期必读图神经网络精选论文

论文可直接下载!

AMiner科技



ICLR 2021 | 自解释神经网络，直接写入了特征的重要值：Shapley Explanation Networks

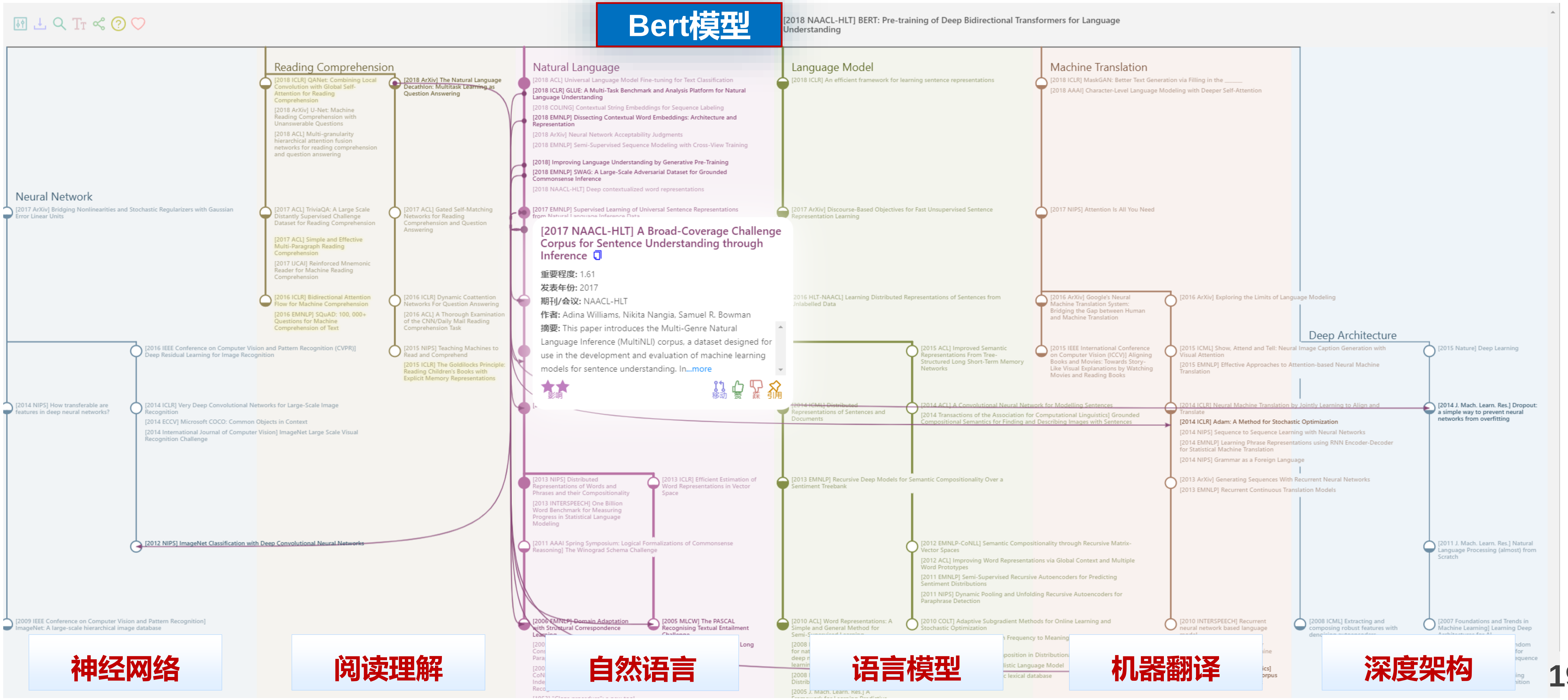
本文介绍了作者最近被ICLR2021录用的一篇关于自解释神经网络的文章。

极市平台

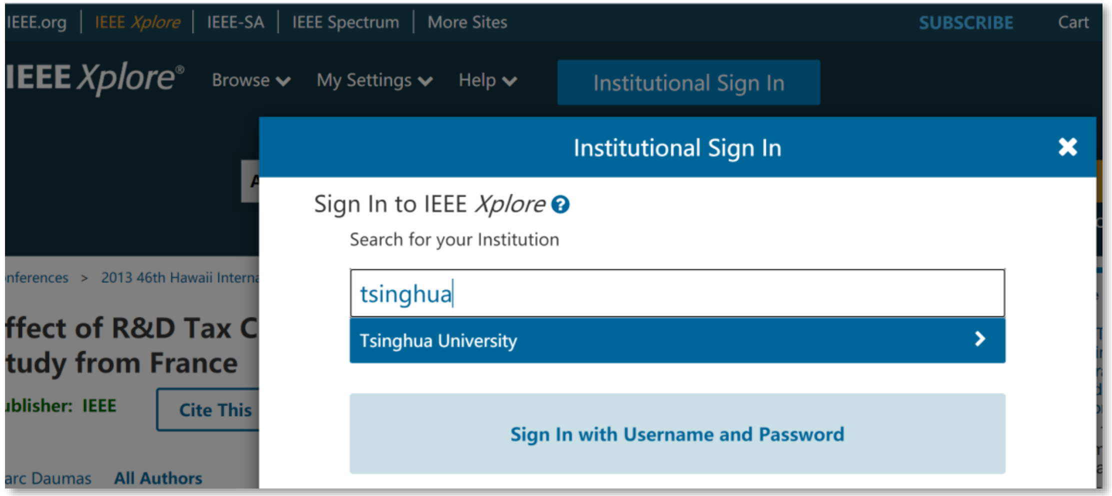
热点前沿论文详细解读

研究方案设计阶段

利用引文网络了解技术发展脉络，分析技术发展趋势。



学校购买数据库



文献求助

Open Access源

其他

AI TIME 网络资源下载

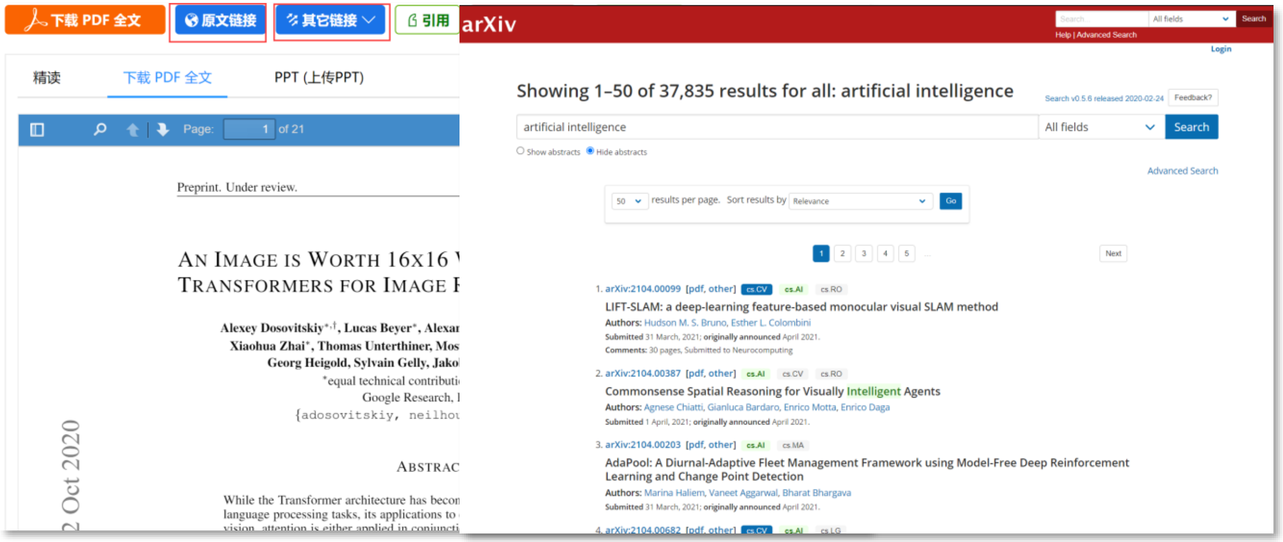
Artificial Intelligence Time

智谱·AI

学校购买数据库

文献求助

Open Access源



其他

AI TIME 关注学术动态

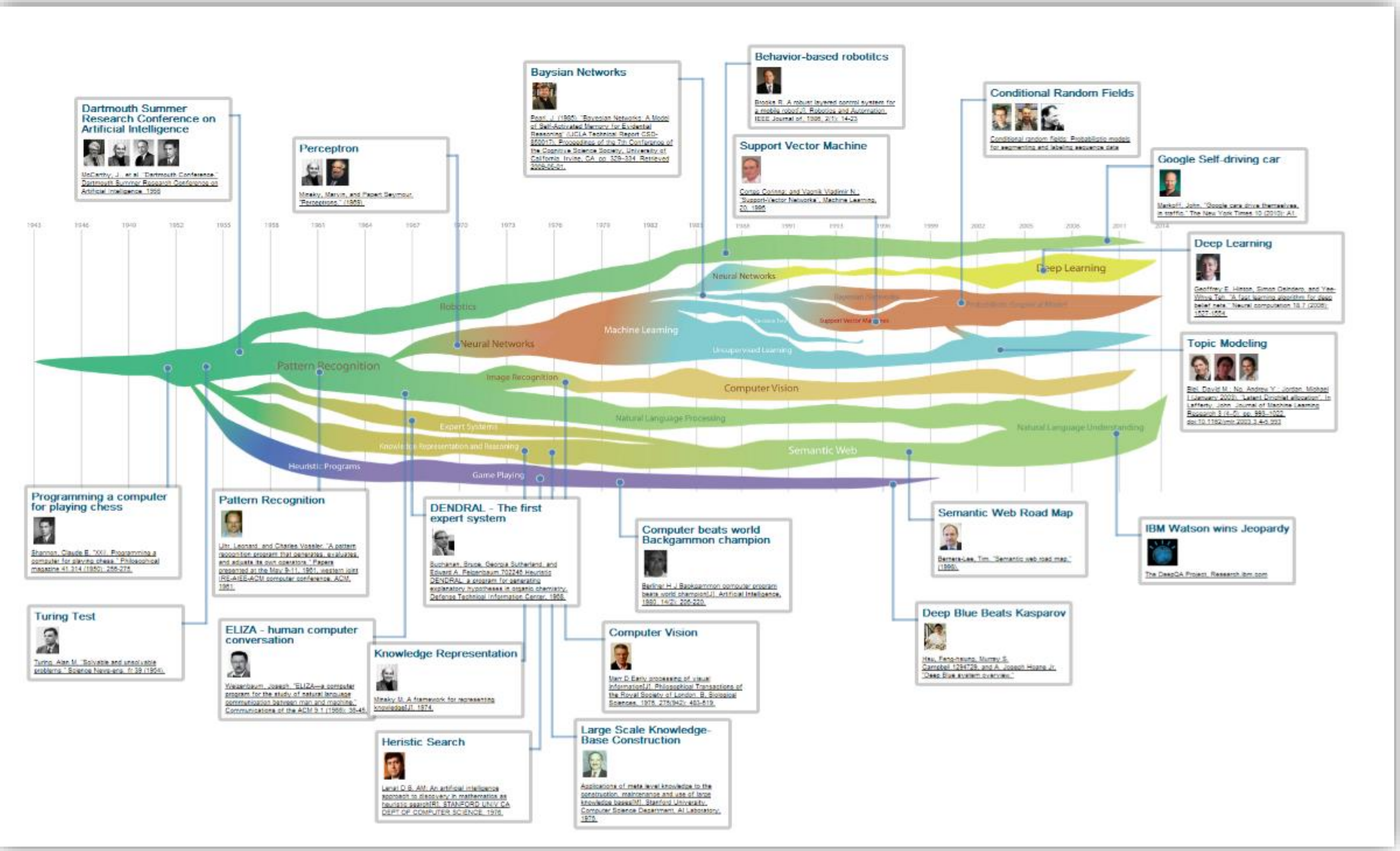
Artificial Intelligence Time

智谱·AI

- 领域科技新闻
- 领域顶会/顶刊动态
- 领域大牛最新研究进展

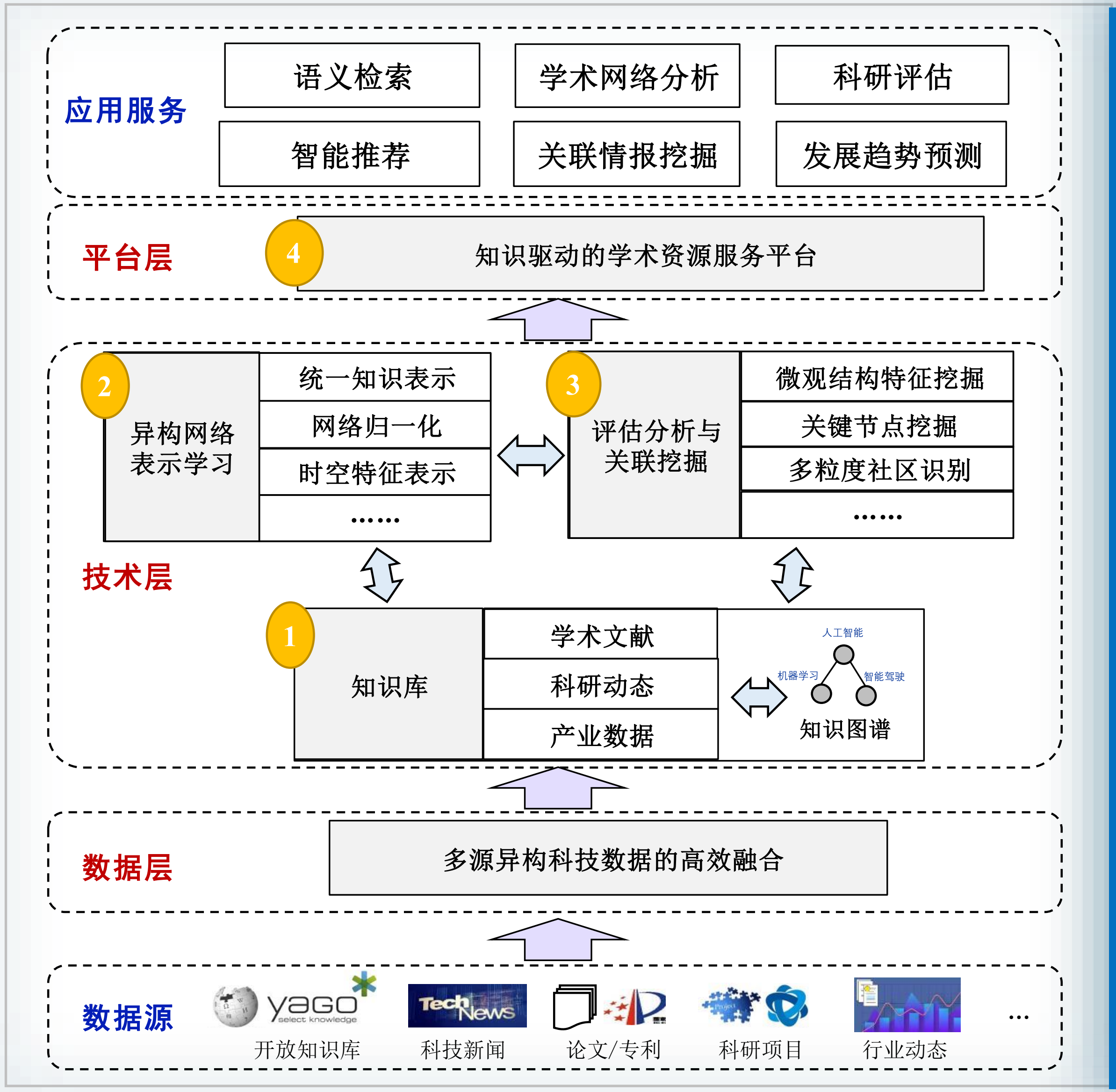
排名	会议名称	H5指数	上升指数	TK5指数	CCF等级	基础研究创新指数	THU等级	指标图
1	[CVPR] IEEE Conference on Computer Vision and Pattern Recognition 06/16/2020 Online	301	91	143	A	81.55	A	
2	[NeurIPS] Annual Conference on Neural Information Processing Systems 06/12/2020 Online	192	88	91.5	A	47.23	A	
3	[ICCV] International Conference on Computer Vision	172	80	104	A	56.67	A	
4	[ICLR] International Conference on Learning Representations	166	83		A	57.11	A	
5	[ICML] International Conference on Machine Learning	151	80	82	A	43.10	A	

情报分析



- 关键词抽取
- 知识图谱构建
- 主题聚类
- 语义匹配
- 关键节点挖掘
- 社区发现
- 影响力评估

情报分析

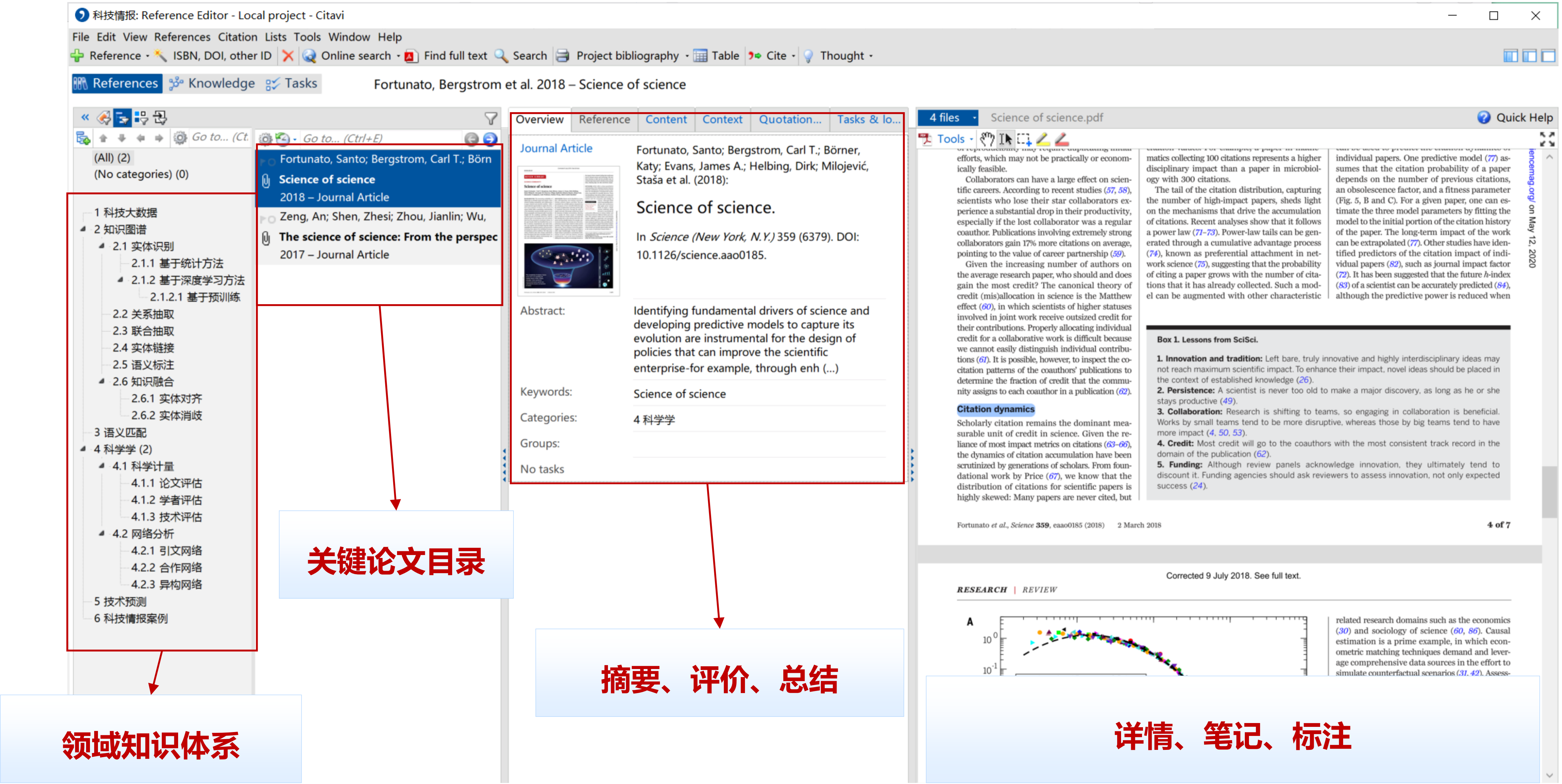


AMiner 平台架构

- **多源异构数据融合**：基于命名消歧与冲突检测、信息补齐等技术，解决多源异构学术资源命名错误和数据冲突、缺失的问题；
- **多维关联知识库**：基于科技知识图谱对论文、专利、学者、项目等科技资源进行语义关联及统一知识表示学习；
- **深度关联挖掘**：分析异构网络显式/隐式关系，网络传播关键节点/路径；
- **可视化分析**：自动生成可视化图表，从多个维度展示技术发展情况；
- **知识服务**：科技发展规划、科研评估、潜在合作分析等。

目 录

文献管理

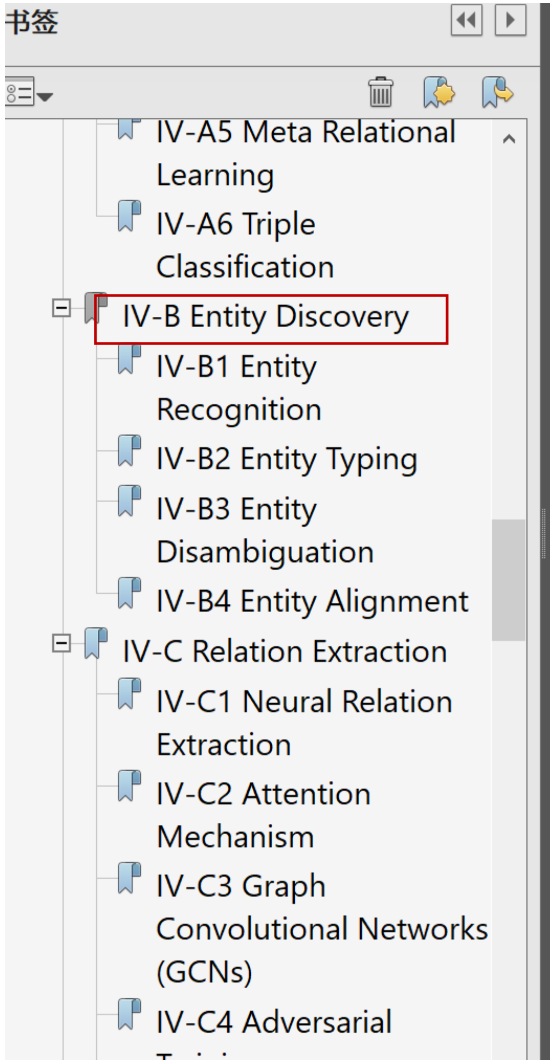


科技情报: Reference Editor - Local project - Citavi
File Edit View References Citation Lists Tools Window Help
Reference ISBN, DOI, other ID Online search Find full text Search Project bibliography Table Cite Thought
References Knowledge Tasks
Fortunato, Bergstrom et al. 2018 – Science of science
(All) (2)
(No categories) (0)
1 科技大数据
2 知识图谱
2.1 实体识别
2.1.1 基于统计方法
2.1.2 基于深度学习方法
2.1.2.1 基于预训练
2.2 关系抽取
2.3 联合抽取
2.4 实体链接
2.5 语义标注
2.6 知识融合
2.6.1 实体对齐
2.6.2 实体消歧
3 语义匹配
4 科学学 (2)
4.1 科学计量
4.1.1 论文评估
4.1.2 学者评估
4.1.3 技术评估
4.2 网络分析
4.2.1 引文网络
4.2.2 合作网络
4.2.3 异构网络
5 技术预测
6 科技情报案例
Fortunato, Santo; Bergstrom, Carl T.; Börn
Science of science
2018 – Journal Article
Zeng, An; Shen, Zhesi; Zhou, Jianlin; Wu,
The science of science: From the perspec
2017 – Journal Article
Overview Reference Content Context Quotation... Tasks & lo...
Journal Article
Fortunato, Santo; Bergstrom, Carl T.; Börner, Katy; Evans, James A.; Helbing, Dirk; Milojević, Staša et al. (2018):
Science of science.
In Science (New York, N.Y.) 359 (6379). DOI: 10.1126/science.aao0185.
Abstract: Identifying fundamental drivers of science and developing predictive models to capture its evolution are instrumental for the design of policies that can improve the scientific enterprise-for example, through enh (...)
Keywords: Science of science
Categories: 4 科学学
Groups:
No tasks
4 files Science of science.pdf Quick Help
Tools
Citation dynamics
Box 1. Lessons from SciSci.
Fortunato et al., Science 359, eaao0185 (2018) 2 March 2018 4 of 7
Corrected 9 July 2018. See full text.
RESEARCH | REVIEW
关键论文目录
摘要、评价、总结
领域知识体系
详情、笔记、标注

AI TIME 文献分析

Artificial Intelligence Time

智谱·AI



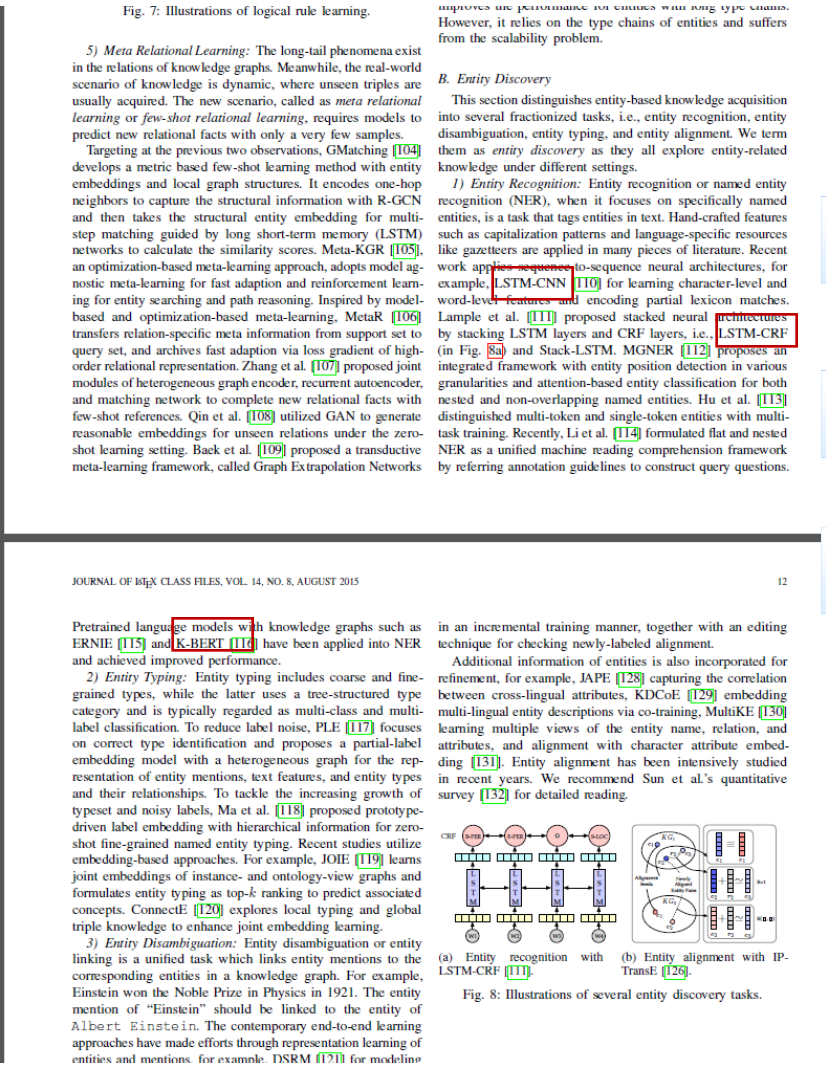


Fig. 7: Illustrations of logical rule learning.

5) Meta Relational Learning: The long-tail phenomena exist in the relations of knowledge graphs. Meanwhile, the real-world scenario of knowledge is dynamic, where unseen triples are usually acquired. The new scenario, called as *meta relational learning* or *few-shot relational learning*, requires models to predict new relational facts with only a very few samples.

Targeting at the previous two observations, GMatching [104] develops a metric based few-shot learning method with entity embeddings and local graph structures. It encodes one-hop neighbors to capture the structural information with R-GCN and then takes the structural entity embedding for multi-step matching guided by long short-term memory (LSTM) networks to calculate the similarity scores. Meta-KGR [105], an optimization-based meta-learning approach, adopts model agnostic meta-learning for fast adaption and reinforcement learning for entity searching and path reasoning. Inspired by model-based and optimization-based meta-learning, MetaR [106] transfers relation-specific meta information from support set to query set, and archives fast adaption via loss gradient of high-order relational representation. Zhang et al. [107] proposed joint modules of heterogeneous graph encoder, recurrent autoencoder, and matching network to complete new relational facts with few-shot references. Qin et al. [108] utilized GAN to generate reasonable embeddings for unseen relations under the zero-shot learning setting. Baek et al. [109] proposed a transductive meta-learning framework, called Graph Extrapolation Networks

improves the performance for entities with long type chains. However, it relies on the type chains of entities and suffers from the scalability problem.

B. Entity Discovery

This section distinguishes entity-based knowledge acquisition into several fractionized tasks, i.e., entity recognition, entity disambiguation, entity typing, and entity alignment. We term them as *entity discovery* as they all explore entity-related knowledge under different settings.

1) Entity Recognition: Entity recognition or named entity recognition (NER), when it focuses on specifically named entities, is a task that tags entities in text. Hand-crafted features such as capitalization patterns and language-specific resources like gazetteers are applied in many pieces of literature. Recent work applies sequence-to-sequence neural architectures, for example, LSTM-CNN [110] for learning character-level and word-level features and encoding partial lexicon matches. Lample et al. [111] proposed stacked neural architectures by stacking LSTM layers and CRF layers, i.e., LSTM-CRF (in Fig. 8a) and Stack-LSTM. MGNER [112] proposes an integrated framework with entity position detection in various granularities and attention-based entity classification for both nested and non-overlapping named entities. Hu et al. [113] distinguished multi-token and single-token entities with multi-task training. Recently, Li et al. [114] formulated flat and nested NER as a unified machine reading comprehension framework by referring annotation guidelines to construct query questions.

JOURNAL OF LaTeX CLASS FILES, VOL. 14, NO. 8, AUGUST 2015 — 12

Pretrained language models with knowledge graphs such as ERNIE [115] and K-BERT [116] have been applied into NER and achieved improved performance.

2) Entity Typing: Entity typing includes coarse and fine-grained types, while the latter uses a tree-structured type category and is typically regarded as multi-class and multi-label classification. To reduce label noise, PLE [117] focuses on correct type identification and proposes a partial-label embedding model with a heterogeneous graph for the representation of entity mentions, text features, and entity types and their relationships. To tackle the increasing growth of typeset and noisy labels, Ma et al. [118] proposed prototype-driven label embedding with hierarchical information for zero-shot fine-grained named entity typing. Recent studies utilize embedding-based approaches. For example, JOIE [119] learns joint embeddings of instance- and ontology-view graphs and formulates entity typing as top-k ranking to predict associated concepts. ConnectE [120] explores local typing and global triple knowledge to enhance joint embedding learning.

3) Entity Disambiguation: Entity disambiguation or entity linking is a unified task which links entity mentions to the corresponding entities in a knowledge graph. For example, Einstein won the Noble Prize in Physics in 1921. The entity mention of "Einstein" should be linked to the entity of `Albert Einstein`. The contemporary end-to-end learning approaches have made efforts through representation learning of entities and mentions, for example, DSRM [121] for modeling

in an incremental training manner, together with an editing technique for checking newly-labeled alignment.

Additional information of entities is also incorporated for refinement, for example, JAPE [128] capturing the correlation between cross-lingual attributes, KDCoE [129] embedding multi-lingual entity descriptions via co-training, MultiKE [130] learning multiple views of the entity name, relation, and attributes, and alignment with character attribute embedding [131]. Entity alignment has been intensively studied in recent years. We recommend Sun et al.'s quantitative survey [132] for detailed reading.

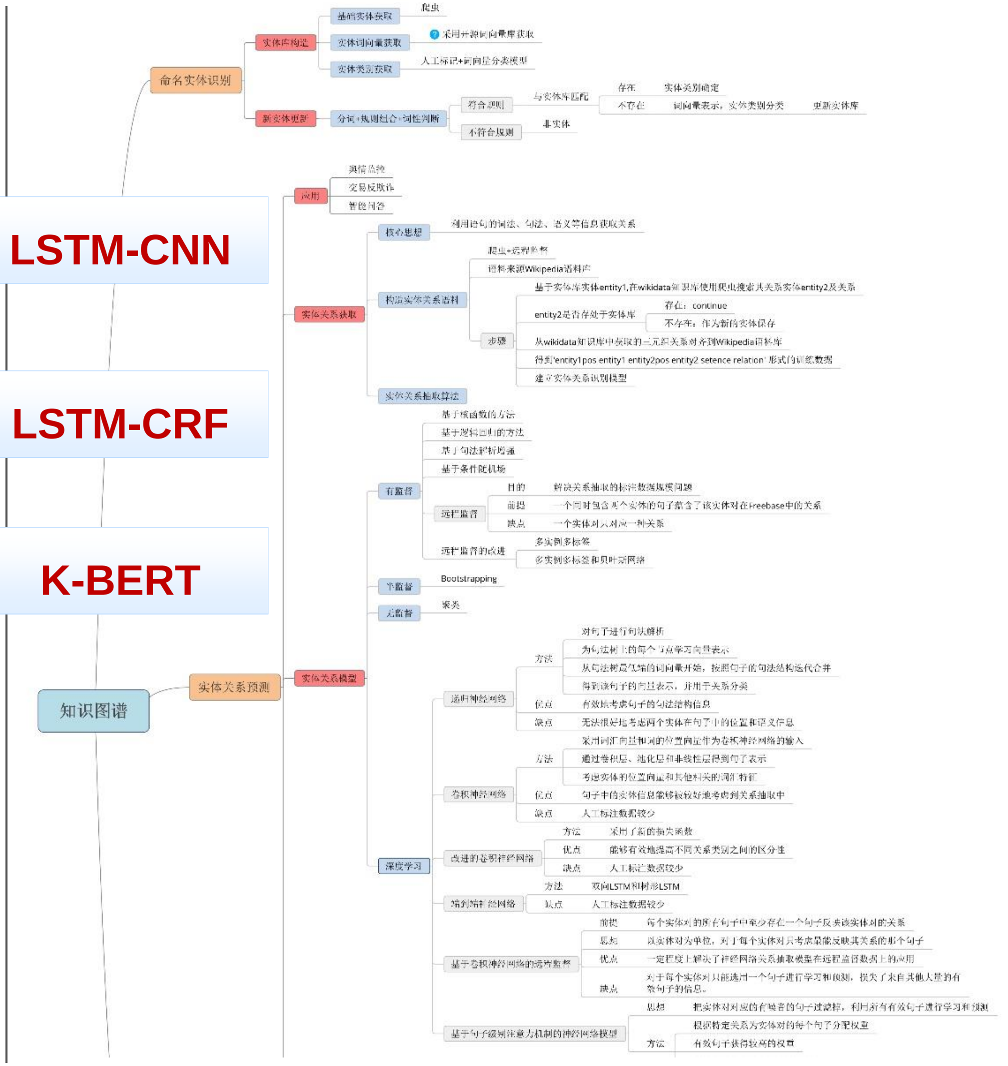
(a) Entity recognition with LSTM-CRF [111]. (b) Entity alignment with IP-TransE [126].

Fig. 8: Illustrations of several entity discovery tasks.

LSTM-CNN

LSTM-CRF

K-BERT



https://www.cnblogs.com/wkang/p/10211259.html

总结

检索需要耐心

注意经验积累

需要批判性思维

祝大家科研顺利，勇攀高峰！

感谢聆听！

qiuyu@mail.tsinghua.edu.cn