

大规模在线教育知识图谱构建

李涓子

知识工程研究室

清华大学计算机科学与技术系

□ 智能学术搜索

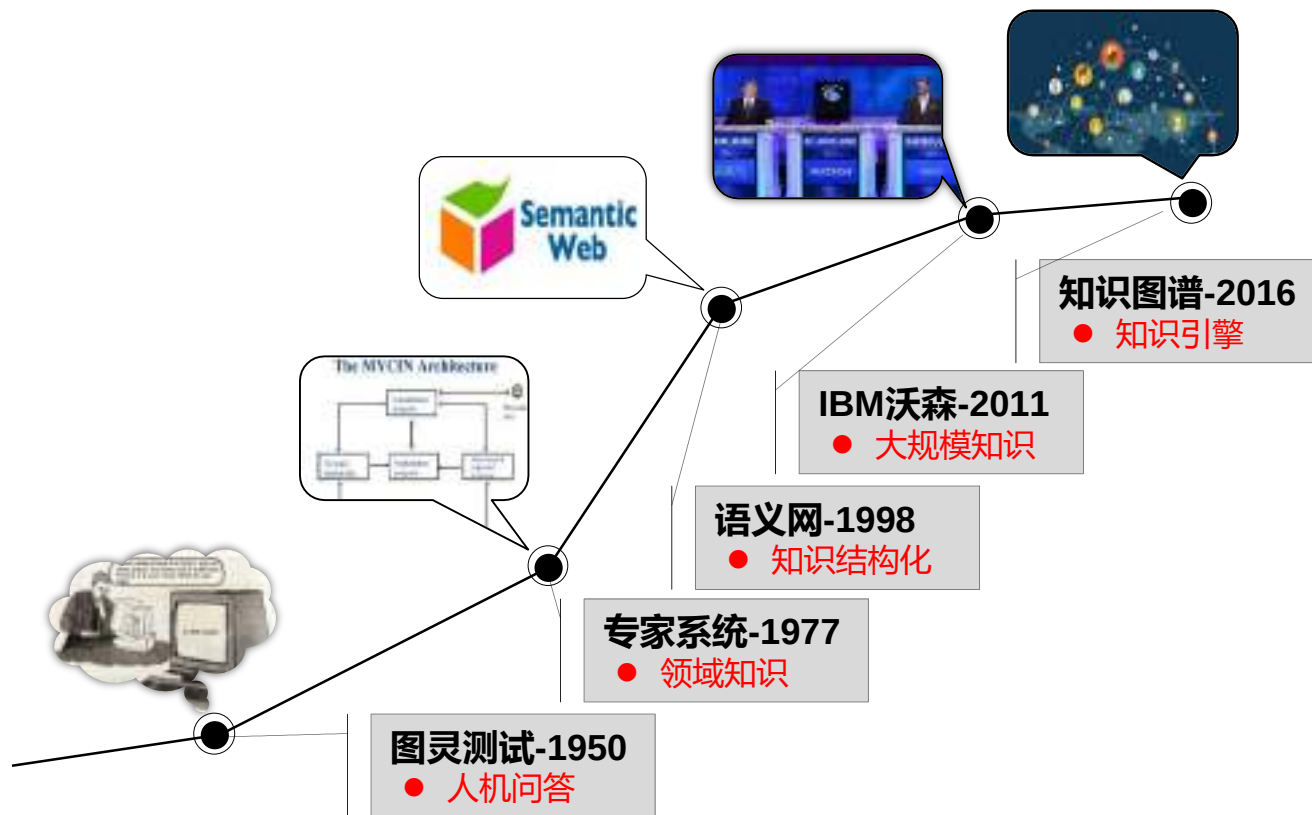
- 人工智能与知识
- 智能学术搜索

□ 大规模在教育知识图谱构建

- 以知识为中心的大规模在线教育资源组织
- 概念学习
- 概念扩展
- 概念先后修关系发现

□ 总结

人工智能对知识的需求

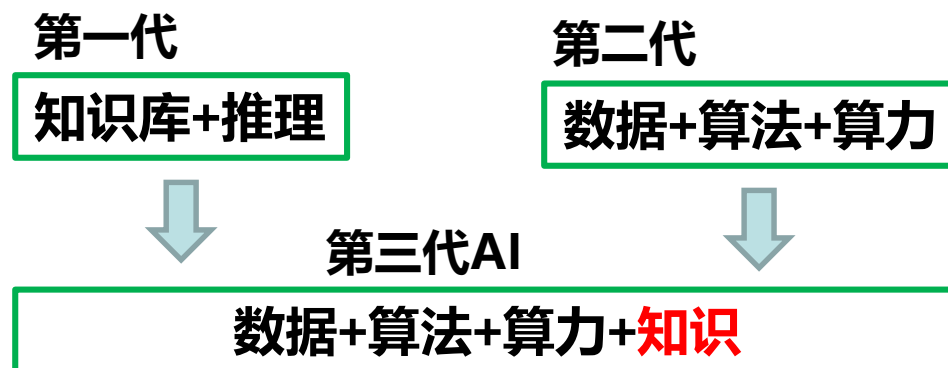


知识是人工智能的核心命题与关键基础

第三代人工智能

- **Explainable, robust, credible and safe AI**
可解释的、鲁棒的、可信安全的人工智能
- **From AI without understanding to AI with understanding**
从不带理解的人工智能成为带理解的人工智能

摘自张钹院士在YOCSEF精英大会上报告



实现从数据智能到知识智能的跃迁是人工智能的必由之路

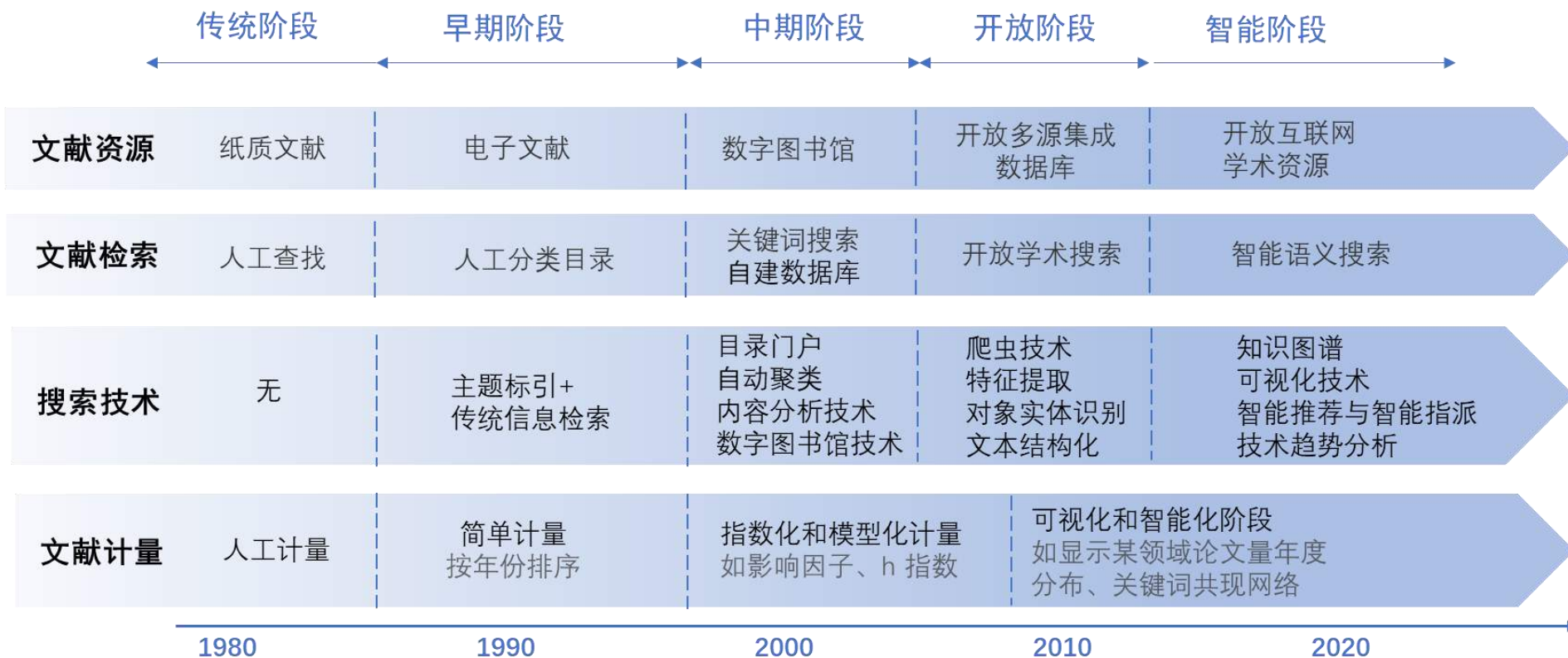
人工智能之学术搜索



扫码微信阅读

人工智能之学术搜索

学术搜索引擎发展历程



人工智能之学术搜索



创新应用



科技趋势分析



科技智库构建



科技成果评估



科技知识传播

科技知识智能助理

双引擎
自动驱动

人的判断
目标约束知识等

经验指导

状态反馈

知识引擎

知识表示
知识获取和融合
知识推理和计算

社会计算引擎

影响力分析
信息传播
行为演化

科技文献
专利
新闻

多源异构
知识融合

机器学习、自然语言处理、信息检索等

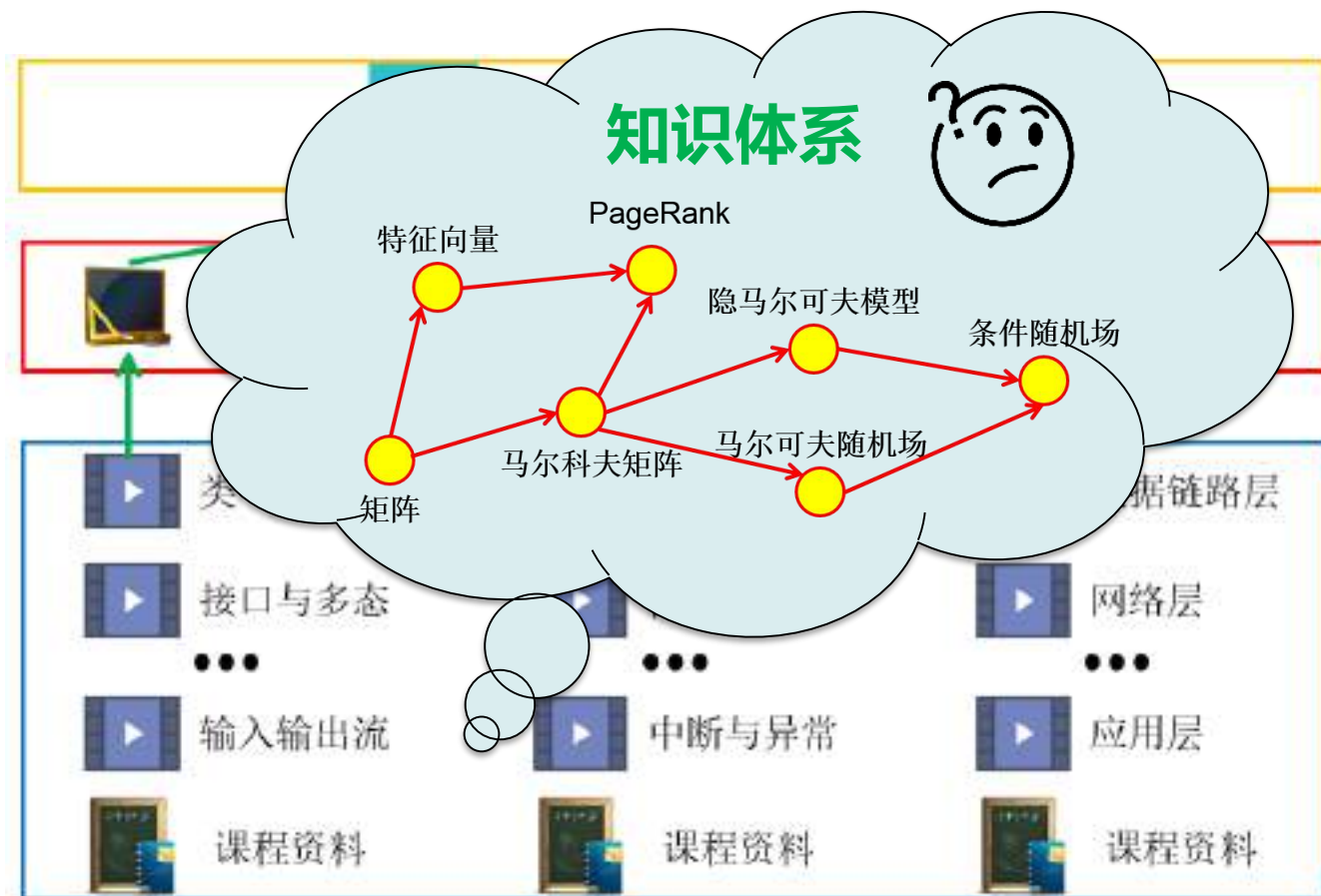
大规模在线教育知识图谱构建

领域—课程—视频 三级组织

1 学科

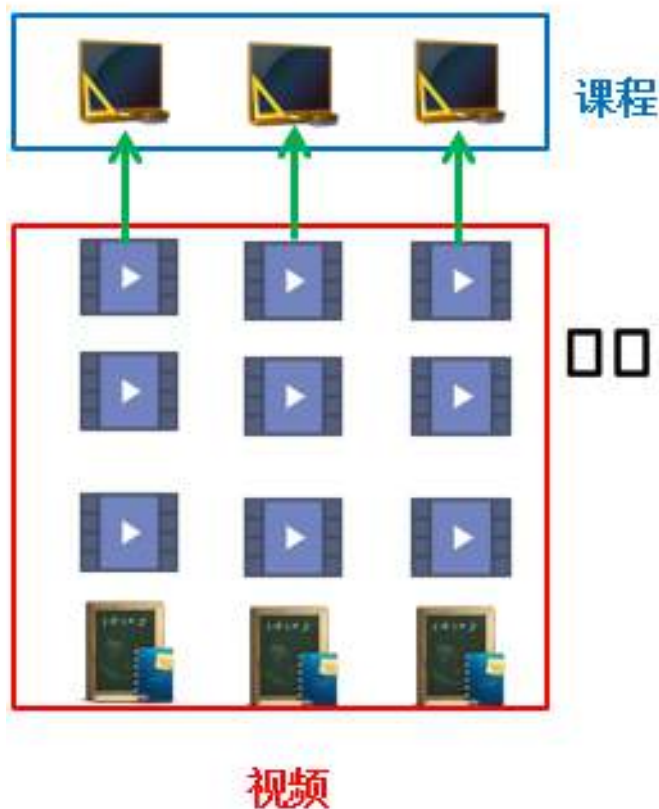
2 课程

3 视频

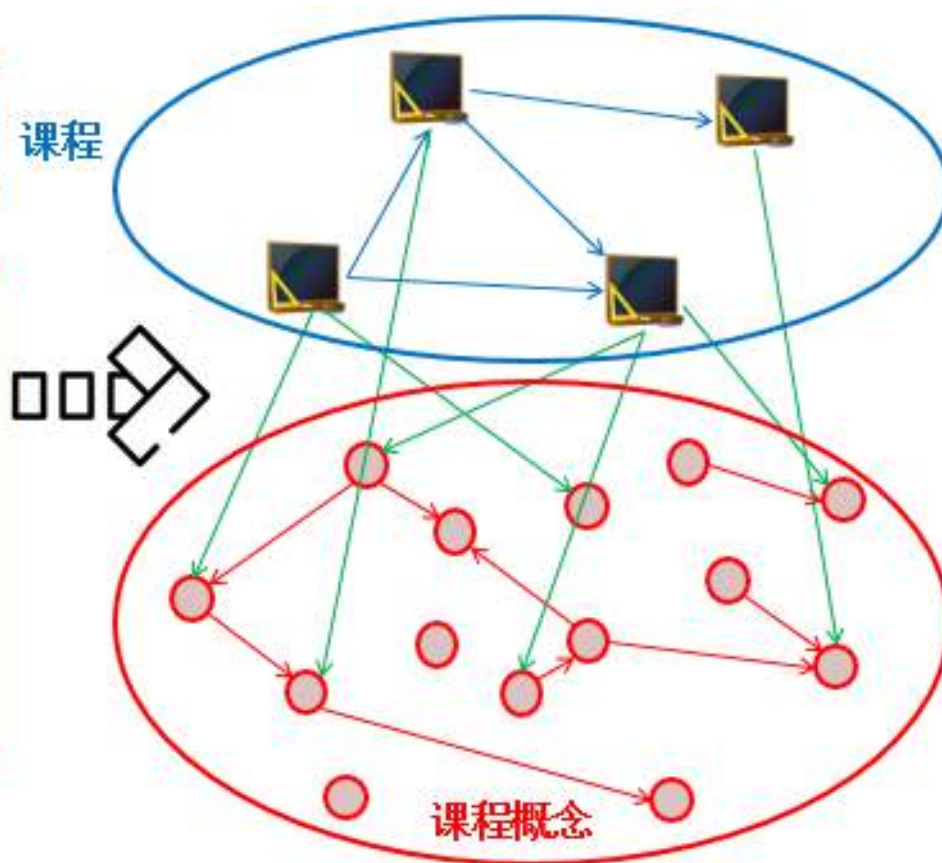


大规模在线教育数据：知识概念

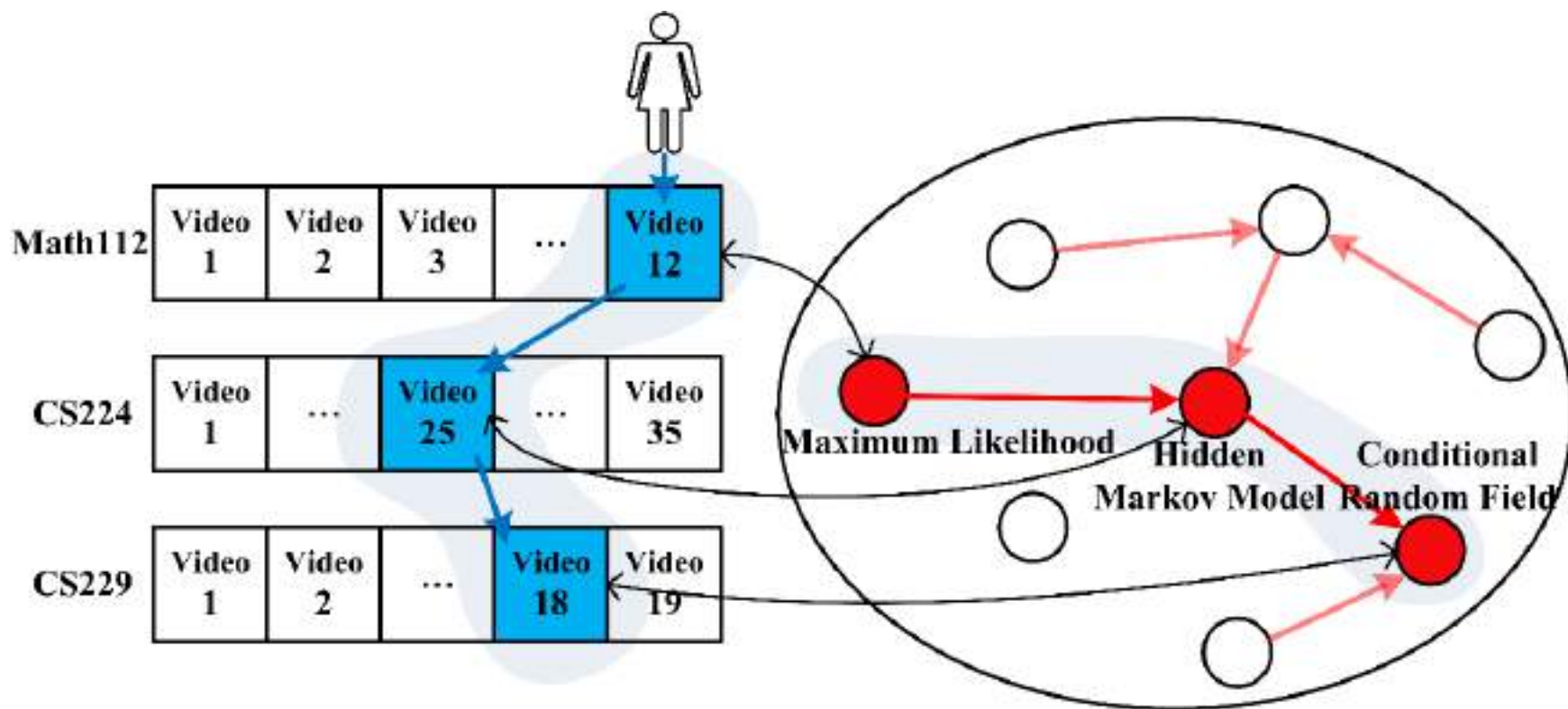
视频为中心



知识为中心



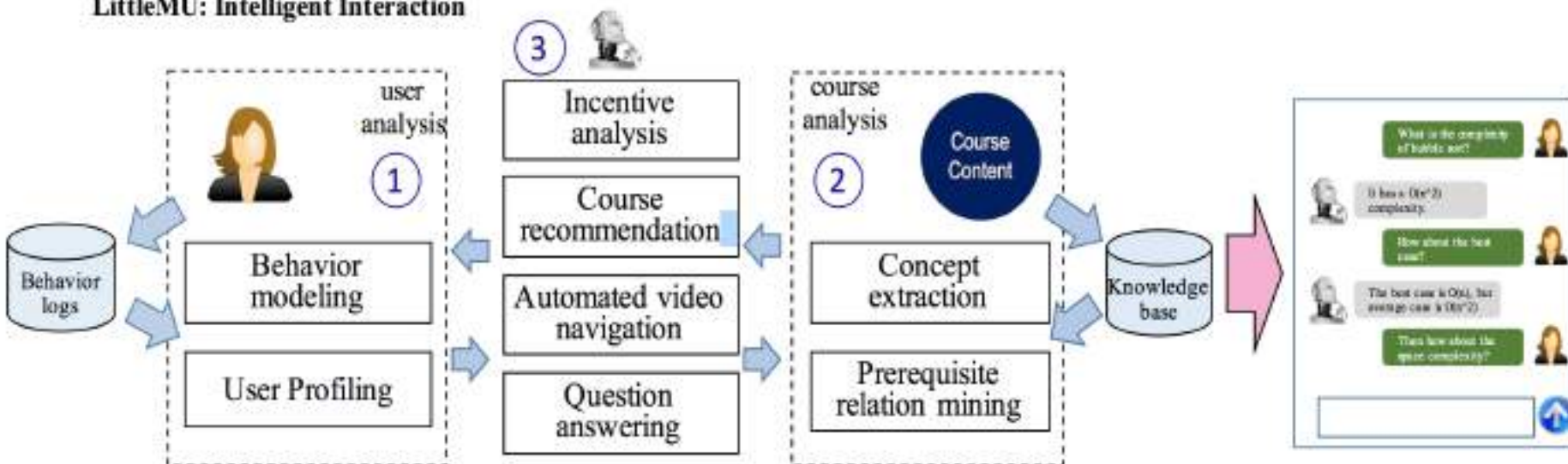
应用——智能学习路径规划



在线教育智能交互的知识服务

应用—智能助教

LittleMU: Intelligent Interaction



大规模在线教育知识图谱



以知识为核心大规模在线教育资源组织

基于置信度网络传播的课程概念抽取

利用外部知识库的概念概念扩展

基于多维度特征的课程概念先后序关系学习

[1] MOOCCube: A Large-scale Data Repository for NLP Applications in MOOCs, ACL2020

[2] Course Concept Extraction in MOOCs via Embedding-Based Graph Propagation, IJCNLP2017

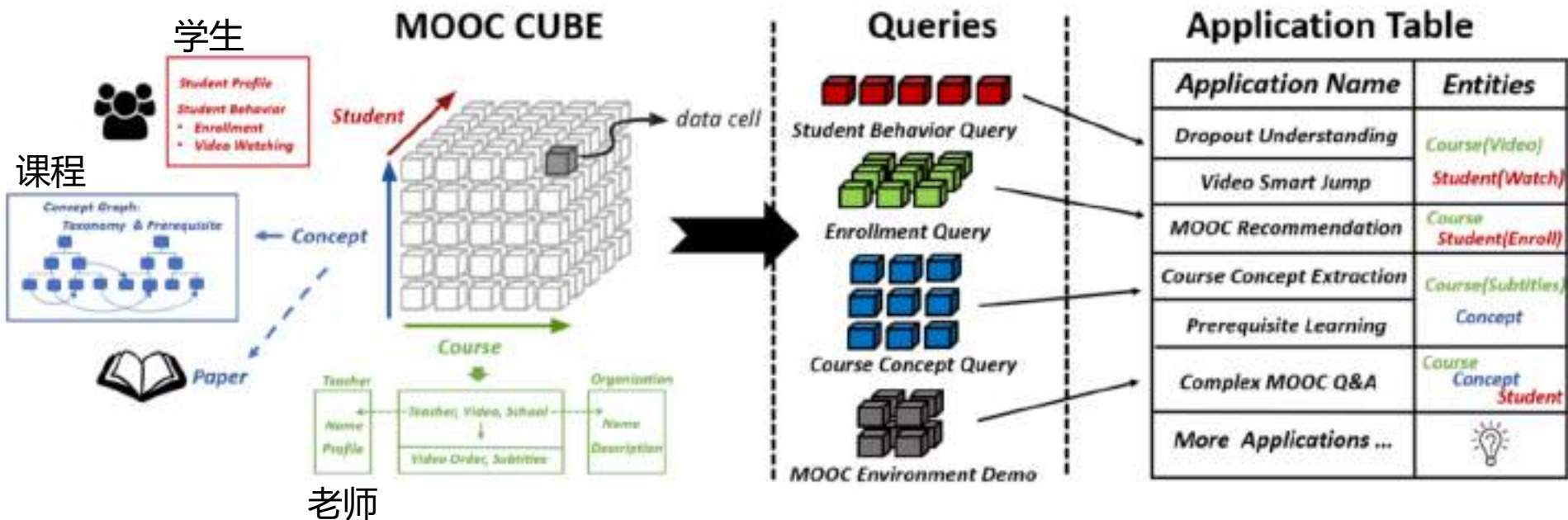
[3] Course Concept Expansion in MOOCs with External Knowledge and Interactive Game, ACL2019

[4] Prerequisite Relation Learning for Concepts in MOOCs. ACL2017

大规模在线教育数据立方体



- MOOCCube: 服务于MOOC相关研究的开源大规模数据仓库
学生行为记录, 课程视频资料, 知识概念体系
开源链接: <http://moocdata.cn/data/MOOCCube>



大规模在线教育数据：用户行为



MOOCCube 细粒度学生行为记录



Student Profile
Student Behavior
• Enrollment
• Video Watching

用户个人画像

性别，年龄，所在地...

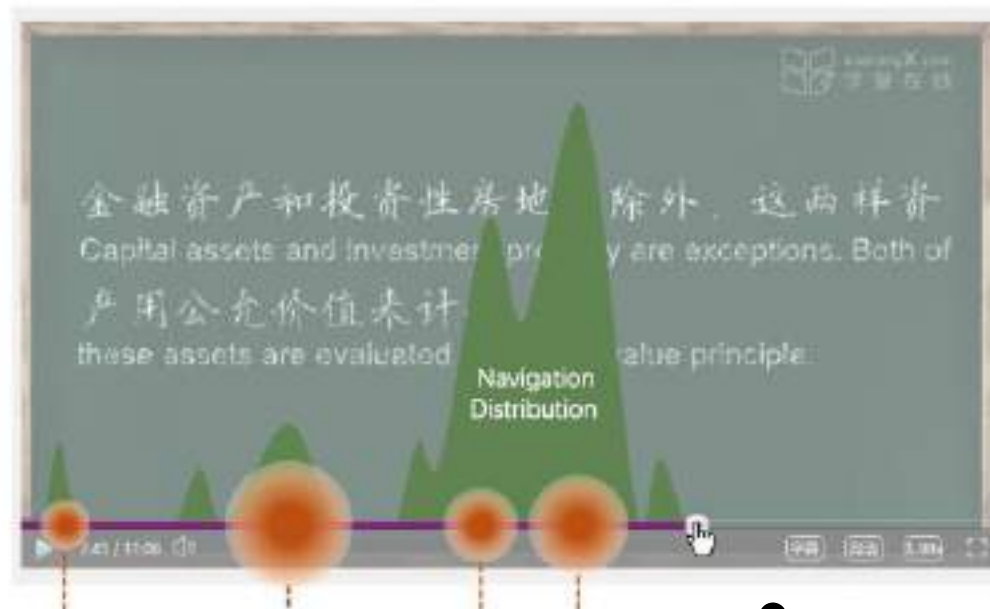
用户视频学习记录

(精确到毫秒)

学习时长

学习次数

观看视频的观看范围(t_1, t_2)



199,199名学生，4,874,298人次观看视频记录

大规模在线教育数据：课程信息

MOOCCube课程信息：

课程信息：

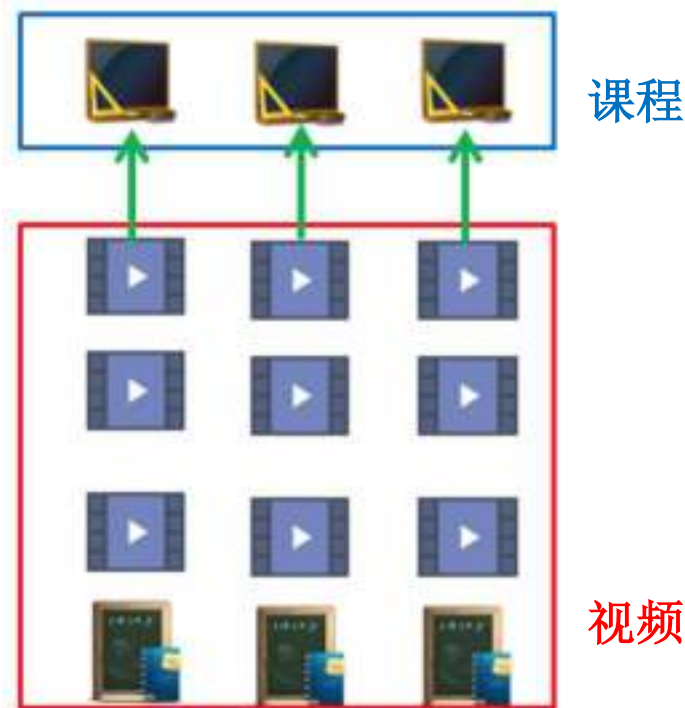
教师

开课机构

视频信息：

字幕文本

视频顺序

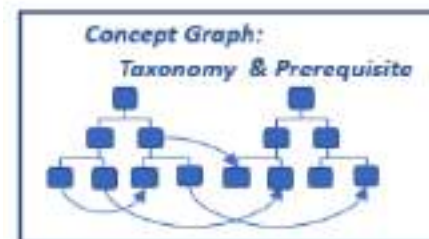


706门课程， 38,181个视频

大规模在线教育数据：知识概念

MOOCCube知识概念体系：

- 课程概念集
- 概念先后修关系
- 概念的相关资源(论文)



106,056个知识概念，并被进一步补充

课程概念 + 先后序关系 知识体系

定义

- 课程视频中所教授的知识概念

例子

You might learn how to write a **bubble sort** and learn why a **bubble sort** is not as good as a **heapsort**. Next, we are going to talk about the **quick sort** algorithm. **Quicksort** is an algorithm invented in the 1960s by doctor Tony Hoare. It is also called the **partition exchange sort**, and is a typical algorithm based on **divide-and-conquer**.

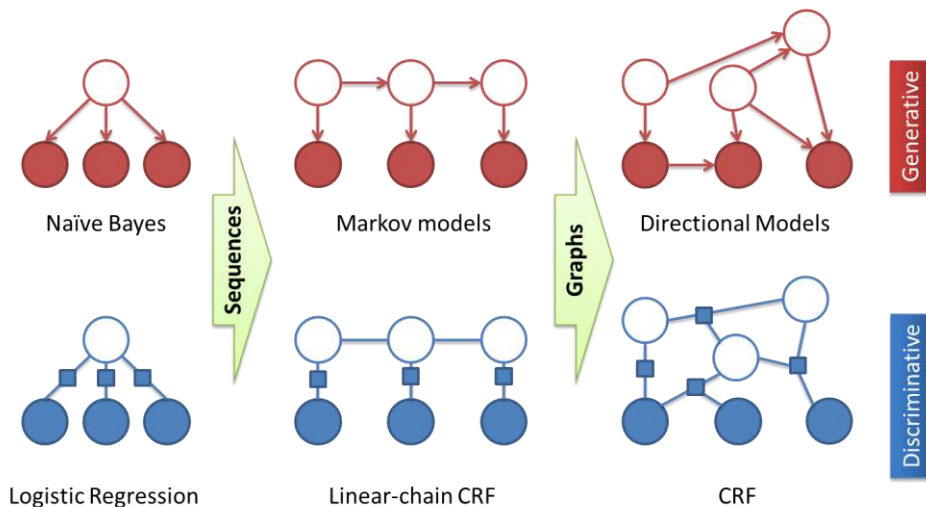
Now we have the first version of **Q sort**. After we make an analysis on its performance, performance, we will find that **quicksort** is an **unstable sorting algorithm**. Fortunately, the **quick sort** has an average **time complexity** of $n \log n$, and in most cases, it can achieve its optimal performance. We first estimate its performance under **independent uniform distribution**.

课程概念 + 先后序关系 知识体系

定义

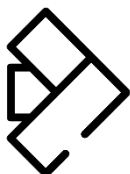
➤ 知识概念间的学习依赖关系

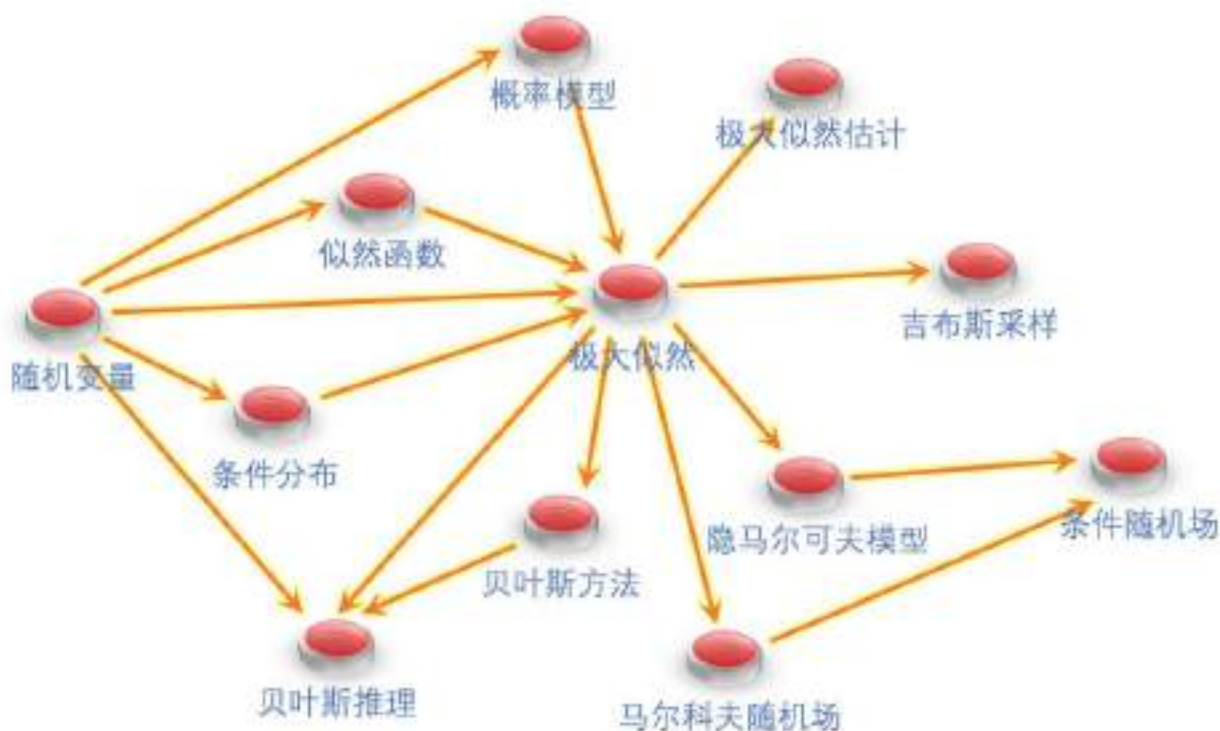
例子



Adapted from C. Sutton, A. McCallum, "An Introduction to Conditional Random Fields", ArXiv, November 2010

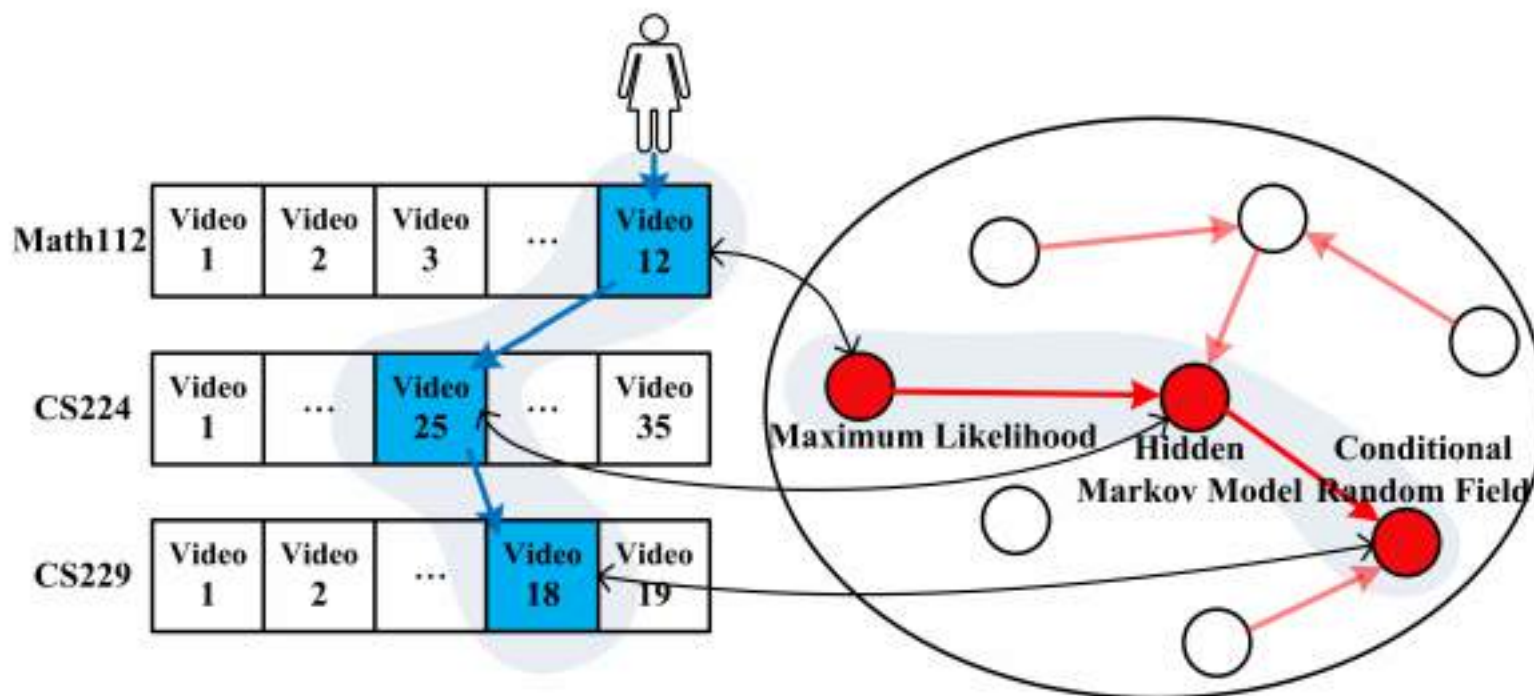
在线教育课程知识体系构建

课程概念 + 先后序关系  知识体系



课程概念图

在线教育课程知识体系构建



- How to extract concepts from course scripts?
- How to recognize (prerequisite) relationships between concepts?

课程概念学习

Candidate
Concept
Extraction



Semantic
Representation
Learning



Graph-
based
Ranking

In this course, we will teach some basic knowledge about **data mining** and its application in **business intelligence**.

Video script

data mining

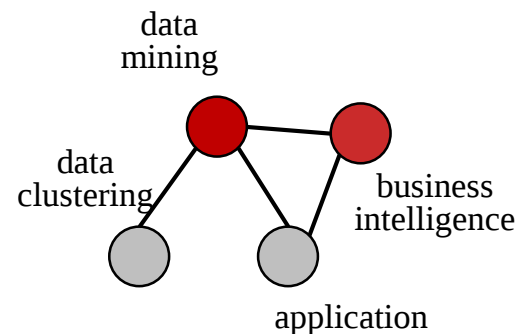
0.8	0.2	0.3	...	0.0	0.0
-----	-----	-----	-----	-----	-----

business intelligence

0.1	0.1	0.2	...	0.8	0.7
-----	-----	-----	-----	-----	-----

Vector representation

Learned via embedding or
deep learning



课程概念抽取

课程概念的三个特性

短语性

- 课程概念应该是一个**语法正确的词组或短语**

信息性

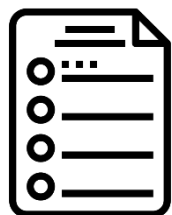
- 课程概念在语义上表示一个**科学或技术术语**

相关性

- 课程概念应该与MOOC课程语料中的**课程相关**

课程概念抽取

语义相关度

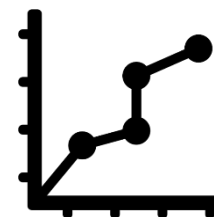


候选课程概念



维基百科语料

- 实体标注
 - 对维基语料进行**实体标注**
- 词嵌入表示学习
 - 学习维基语料的**词嵌入表示**
- 候选课程概念表示
 - 通过**向量加和**得到候选概念向量表示
- 语义相关度量
 - **余弦相似度**作为候选概念间语义相关度

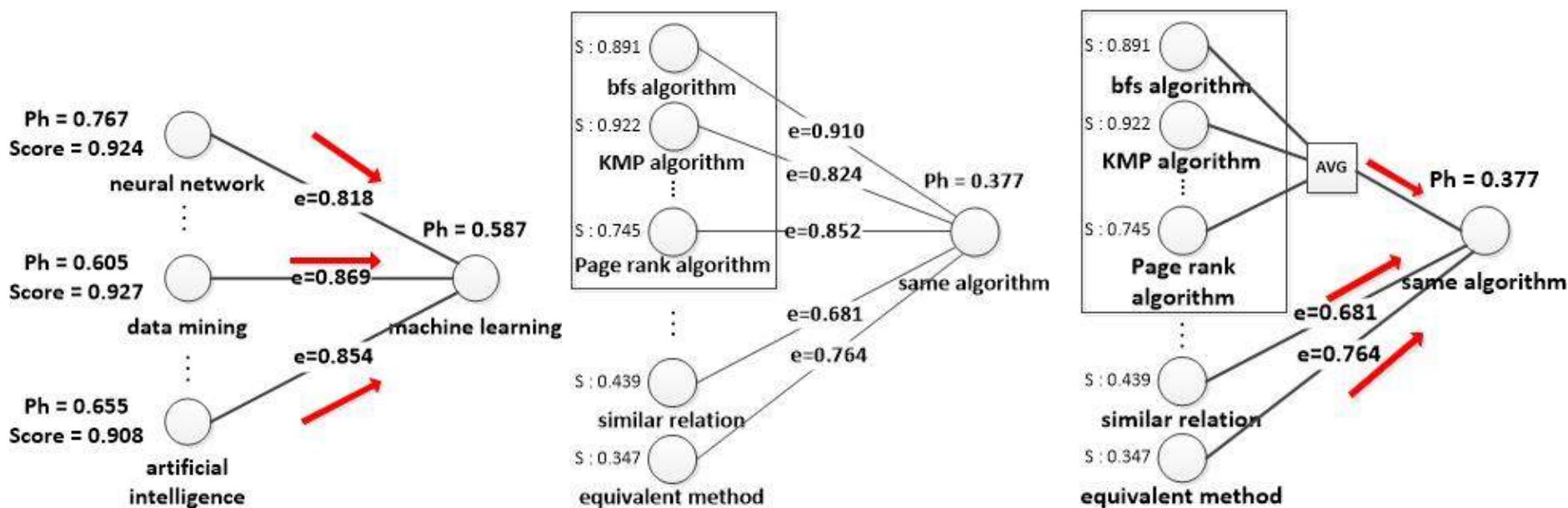


语义相关度

课程概念抽取

置信度图传播

- 课程概念图中，课程概念与其他课程概念以高语义相关度相连接
- 可以通过少数种子概念通过置信度传播找出更多课程概念



课程概念扩展

大量课程相关概念并不出现在课程文本中

Course subtitles

... there should be data structures that would give it to you; probably a **binary search tree** is the first thing you'd want to consider...

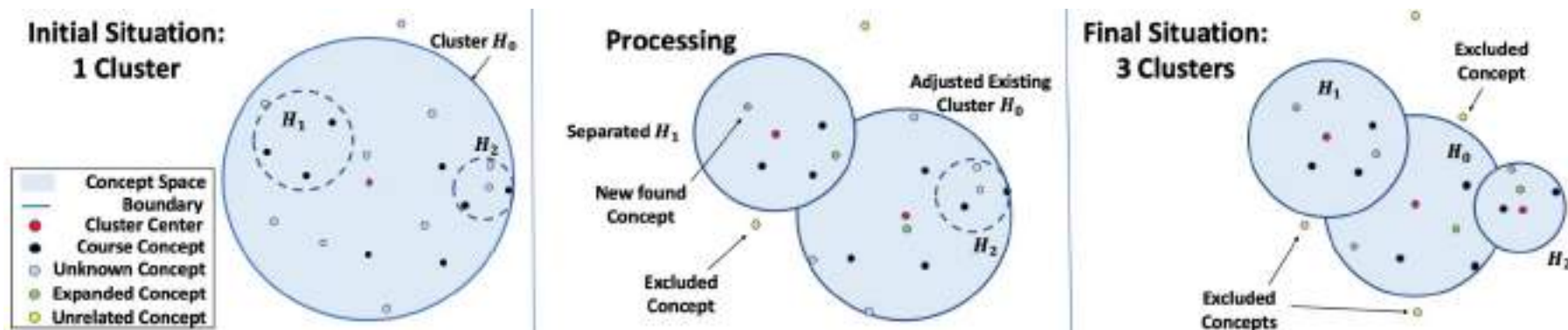
Hidden related concepts

Heap **Tango Tree**
Sorting Algorithms
Priority Queue

这些概念可能由于多种原因能够帮助当前学习

课程概念扩展

利用外部知识库中的概念，进行课程概念扩展



- 语义漂移是课程概念扩展任务中一个重要的挑战，因此使用了基于动态聚类的概念扩展方法

前后序概念关系学习

语义特征



上下文特征



结构特征



语义特征

语义相关度

- 语义相关度在判断概念对是否具有先后序关系上具有重要作用
- 若两个概念语义相关度很低，则它们之间具有学习依赖性的概率较低
- 采用与课程概念抽取任务中相同的方法度量语义相关度

矩阵  人类学

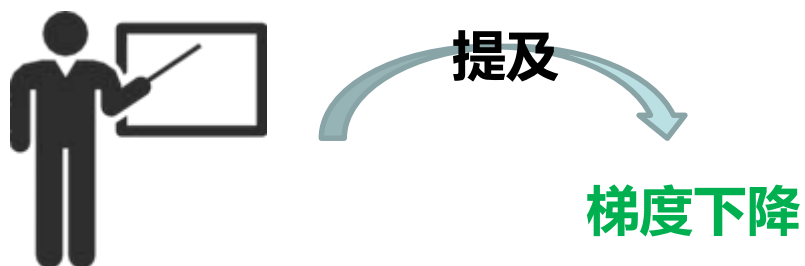
梯度下降  后向传播算法

上下文特征

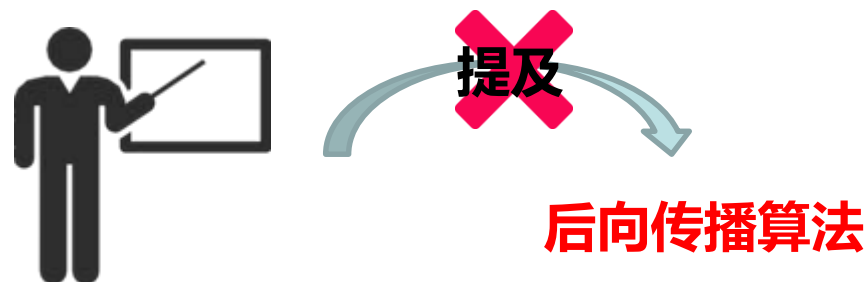
引用距离

- **核心假设：**在教师教授概念A时，如果教师经常提及概念B，反过来，在教师教授概念B时，概念A却很少被提及，则概念B很可能是概念A的先序概念

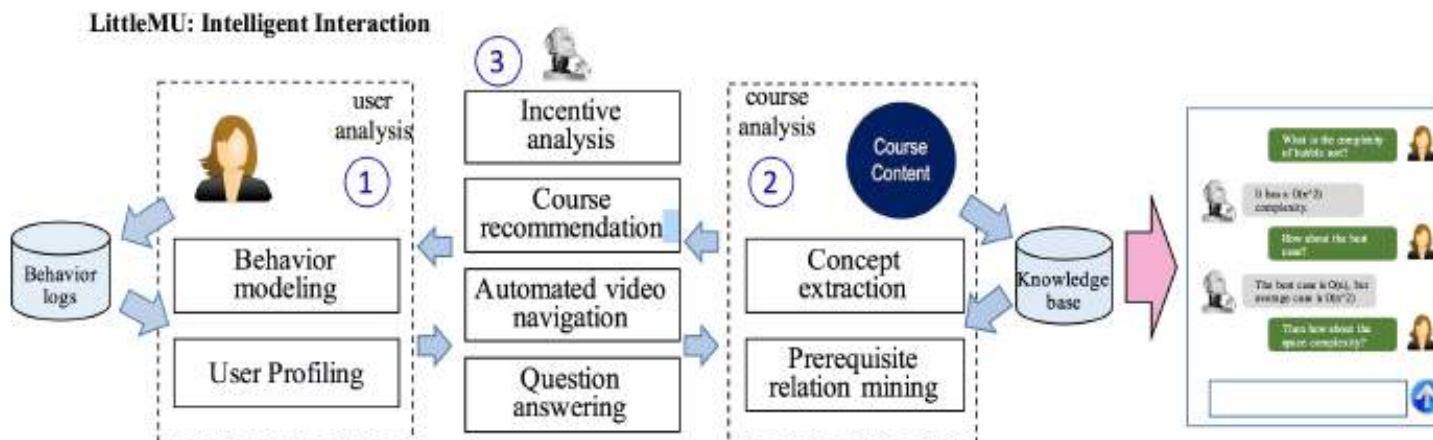
后向传播算法



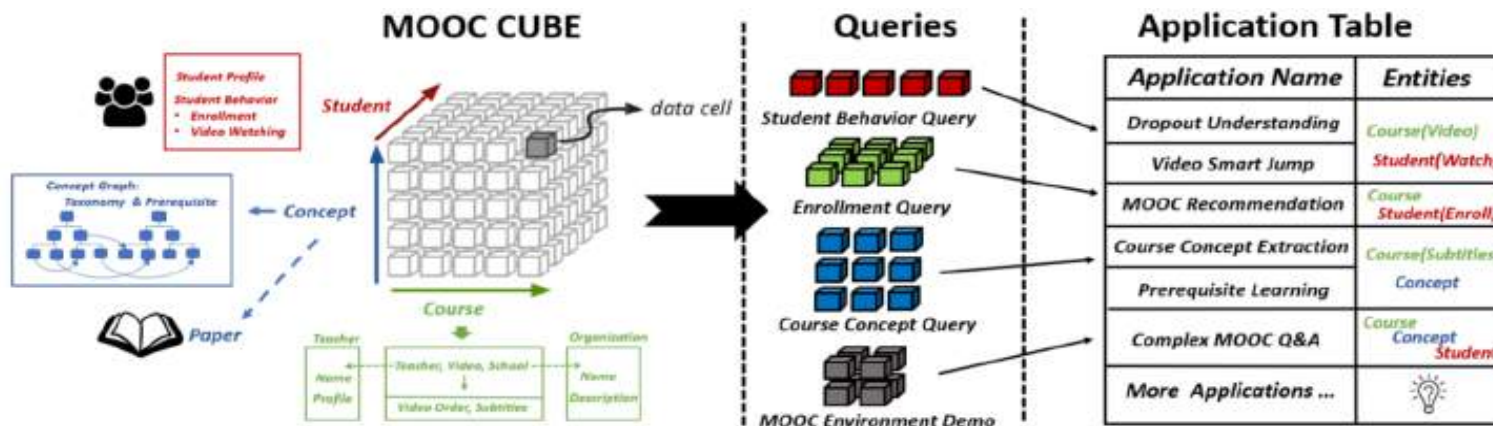
梯度下降



在线教育智能交互的知识服务



学堂智能知识机器人“小木”



大规模在线教育数据仓库 MOOCCube

总 结



- ❑ 知识图谱是人工智能的基础设施
- ❑ 知识图谱构建需要高质量大规模知识获取技术支持
- ❑ 知识驱动的智能学术搜索技术是学术搜索的新型态

谢谢大家!

大数据知识管理服务平台技术与实践

杜一

中国科学院计算机网络信息中心 大数据知识工程实验室

2020年6月2日

学术搜索

Search any topic, author, journal, etc. or any combination of these



Google Scholar



Microsoft[®] Academic Search Beta

Baidu 学术

学术搜索

Search any topic, author, journal, etc. or any combination of these



Google Scholar



Microsoft Academic Search Beta

Baidu 学术



Semantic Scholar



Academia.edu



Aminer

学术搜索

Search any topic, author, journal, etc. or any combination of these



Google Scholar



Microsoft Academic Search Beta

Baidu 学术



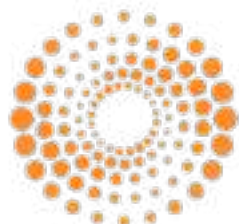
Semantic Scholar



Academia.edu



Miner



WEB OF SCIENCE

Scopus®



万方数据

WANFANG DATA



cnki 中国知网
www.cnki.net

中国知识基础设施工程

Explicit Semantic Ranking for Academic Search via Knowledge Graph Embedding

论文结论

(Zhongxin Wang, Russell Power, James P. Callan) • Published in WWW 2017 • Computer Science

Proceedings of the 2017 ACM Conference on World Wide Web

This paper introduces Explicit Semantic Ranking (ESR), a new ranking technique that leverages knowledge graph embedding. [...] Experiments demonstrate ESR's ability in improving Semantic Scholar's online production system, especially on hard queries where word-based ranking fails. [Expand Abstract](#)

87 Citations

5 Highly Influenced Papers
64 Cite Background
36 Cite Methods
3 Cite Results

引文分类

View On ACM

Alternate Sources

in Library

Create Alert

Like

In Feed



Semantic Scholar

ABSTRACT

FIGURES, TABLES, AND TOPICS

87 CITATIONS

7 REFERENCES

RELATED PAPERS

Figures, Tables, and Topics from this paper.

论文主题

Figures and Tables



Figure 1

Table 1



Figure 2

Table 2

SHOW MORE (4)

Explore Further: Topics Discussed in This Paper

- Knowledge Graph
- Graph embedding
- Academic Search
- Semantic Scholar
- Production systems (computer science)
- Web search engine

学术搜索

DOI: 10.1145/3038912.3062558 • Corpus ID: 1644335

Explicit Semantic Ranking for Academic Search via Knowledge Graph Embedding

(Zhong Wang, Russell Power, James P. Callan) • Published in WWW 2017 • Computer Science

Proceedings of the 2017 ACM Conference on World Wide Web

This paper introduces Explicit Semantic Ranking (ESR), a new ranking technique that leverages knowledge graph embedding. [...] Experiments demonstrate ESR's ability in improving Semantic Scholar's online production system, especially on hard queries where word-based ranking fails. [Expand Abstract](#)

论文结论

87 Citations

5 Highly Influential Papers
64 Cite Background
36 Cite Methods
3 Cite Results

引文分类

[View On ACM](#)

[Alternate Sources](#)

[in Library](#)

[Create Alert](#)

[Like](#)

[In Feed](#)



Semantic Scholar

ABSTRACT

FIGURES, TABLES, AND TOPICS

87 CITATIONS

7 REFERENCES

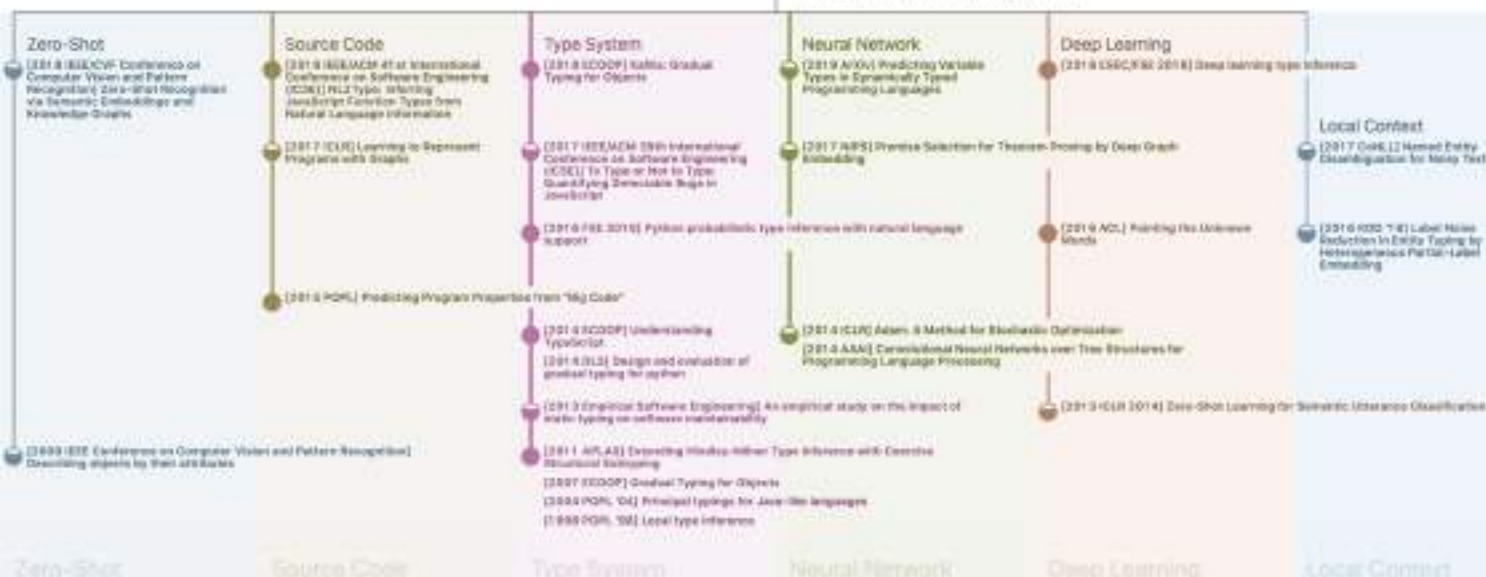
RELATED PAPERS

Figures, Tables, and Topics from this paper.

论文主题

Figures and Tables

[2020 (ICLR 2020) LambdaNet: Probabilistic Type Inference using Graph Neural Networks



Explore Further: Topics Discussed in This Paper

- Knowledge Graph
- Graph embedding
- Academic Search
- Semantic Scholar
- Production systems (computer science)
- Web search engine

Aminer 论文溯源树

Aminer 2020
Aminer Master

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



国家自然科学基金大数据知识管理服务平台



烟草科技知识图谱服务平台



Information Science Data Platform

大数据知识管理服务平台



大数据知识管理服务平台

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

机构画像

大数据知识管理服务平台

来自内部的需求逐渐增多

基金委

科学院

空间中心

烟草总局

出版社

...

交叉学科

学术团队

资助方向

专家画像

热词矩阵

人才合作

研究热点

机构画像

大数据知识管理服务平台

来自内部的需求逐渐增多

交叉学科

学术深度挖掘需求迫切

人员关系

科研社区

学术画像

人才引进

学科交叉

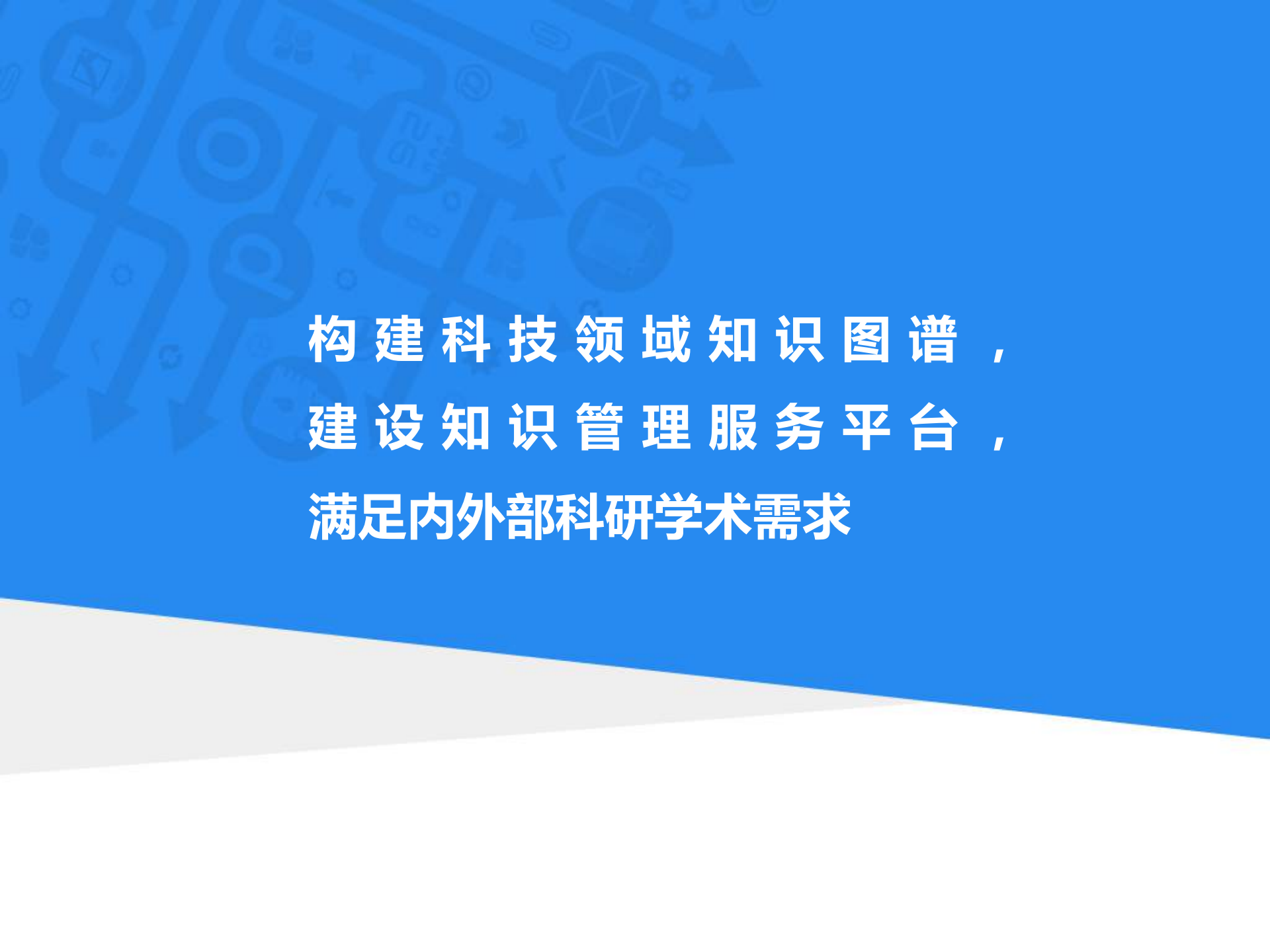
...

资助方向

专家画像

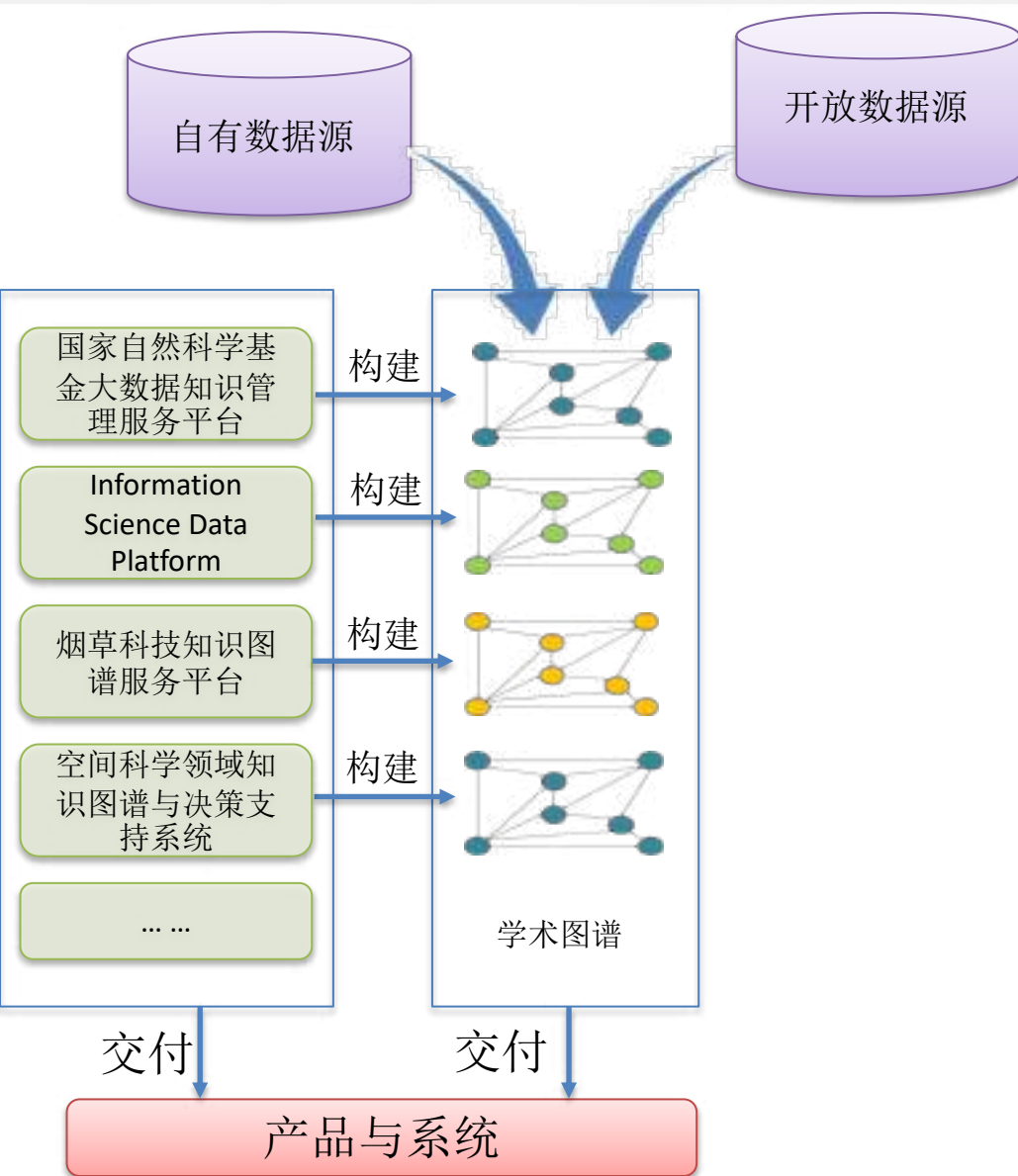
人才合作

机构画像

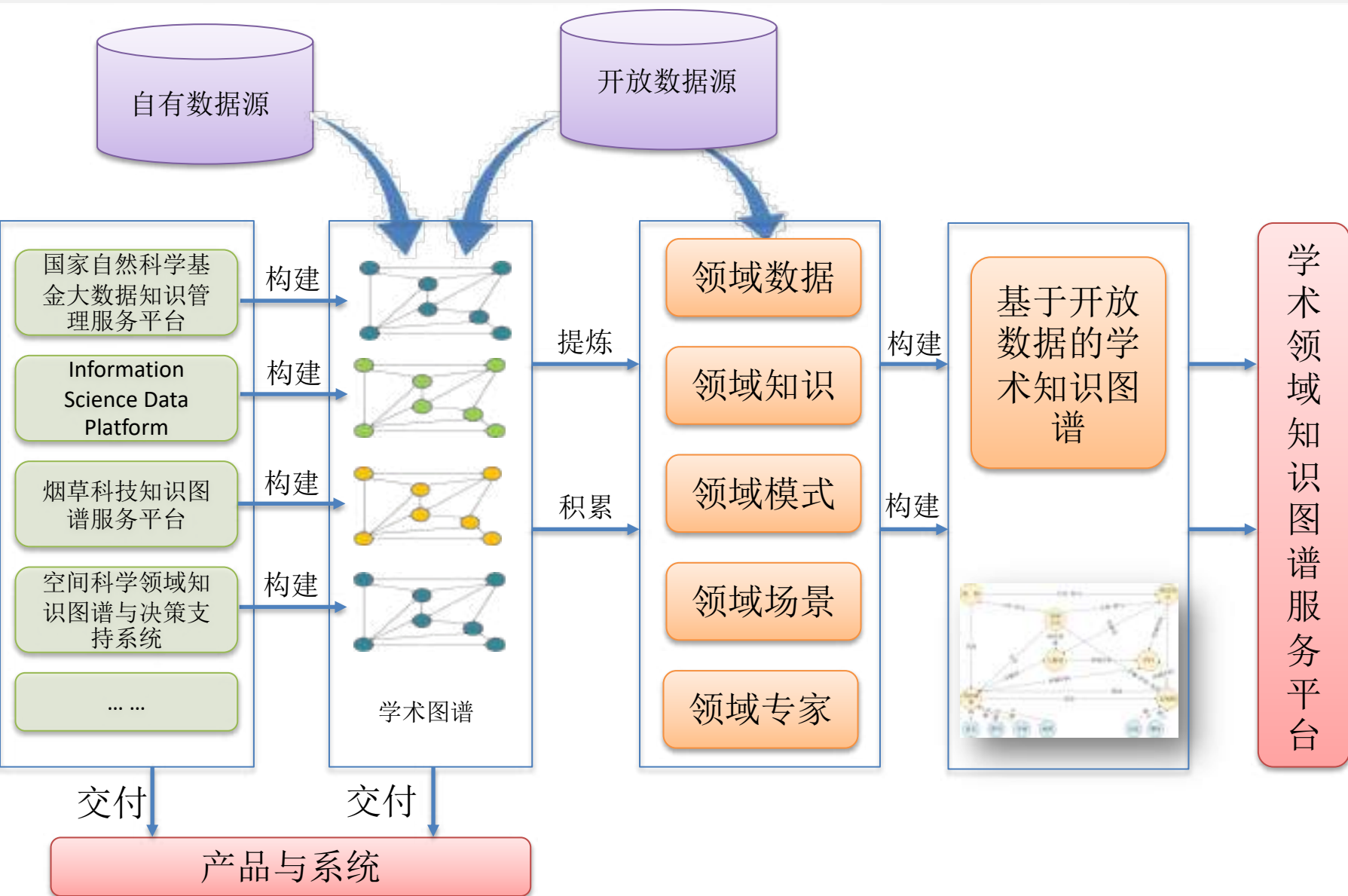


**构建科技领域知识图谱，
建设知识管理服务平台，
满足内外部科研学术需求**

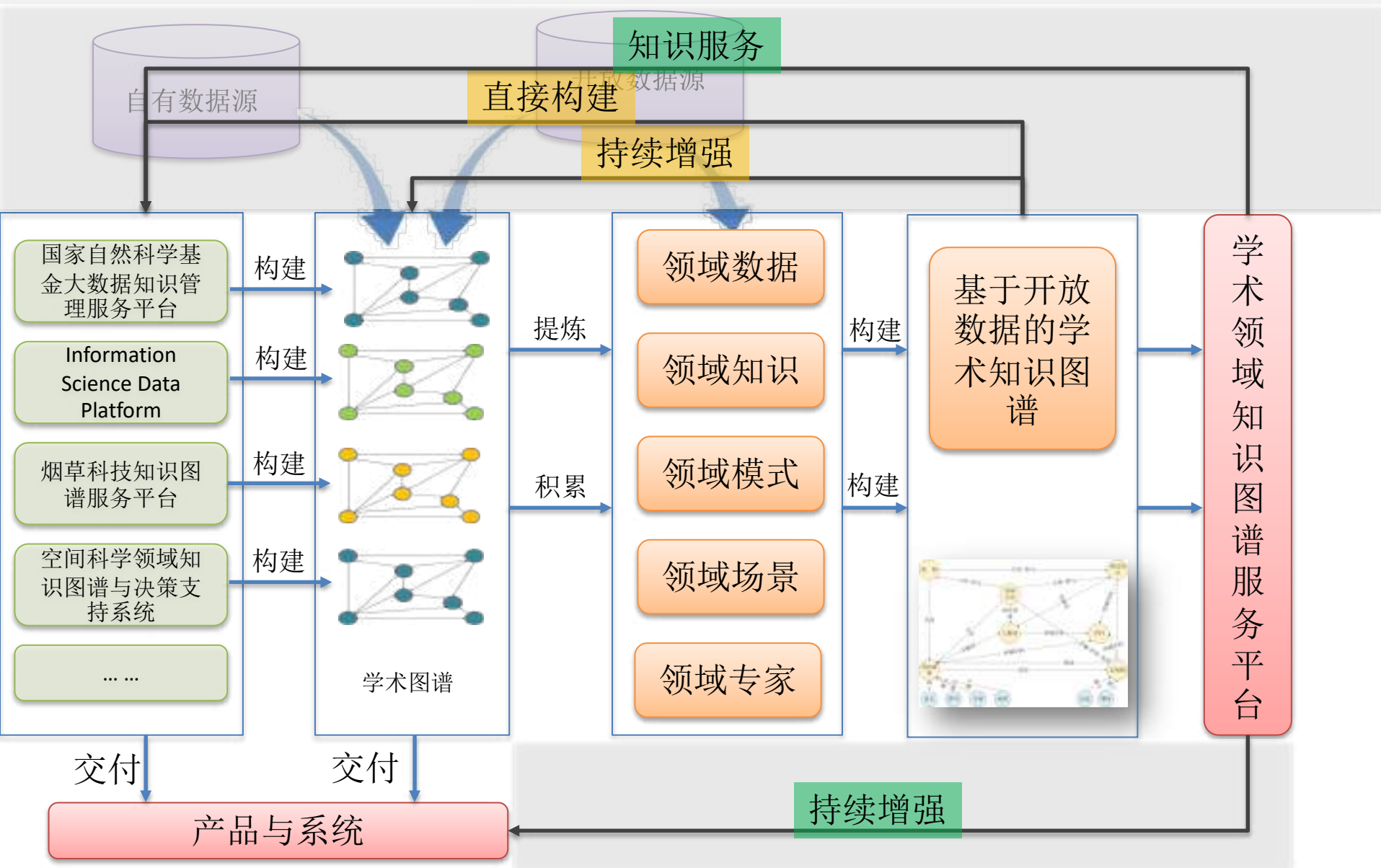
科技领域知识图谱



科技领域知识图谱

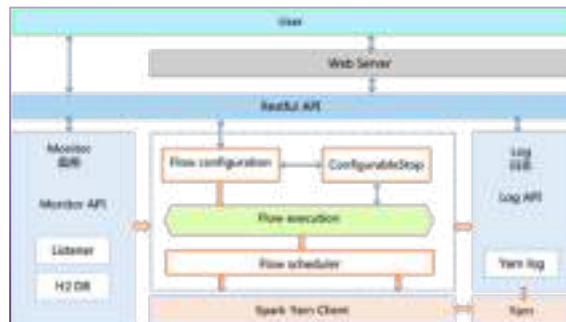


科技领域知识图谱



科技领域知识图谱：关键技术

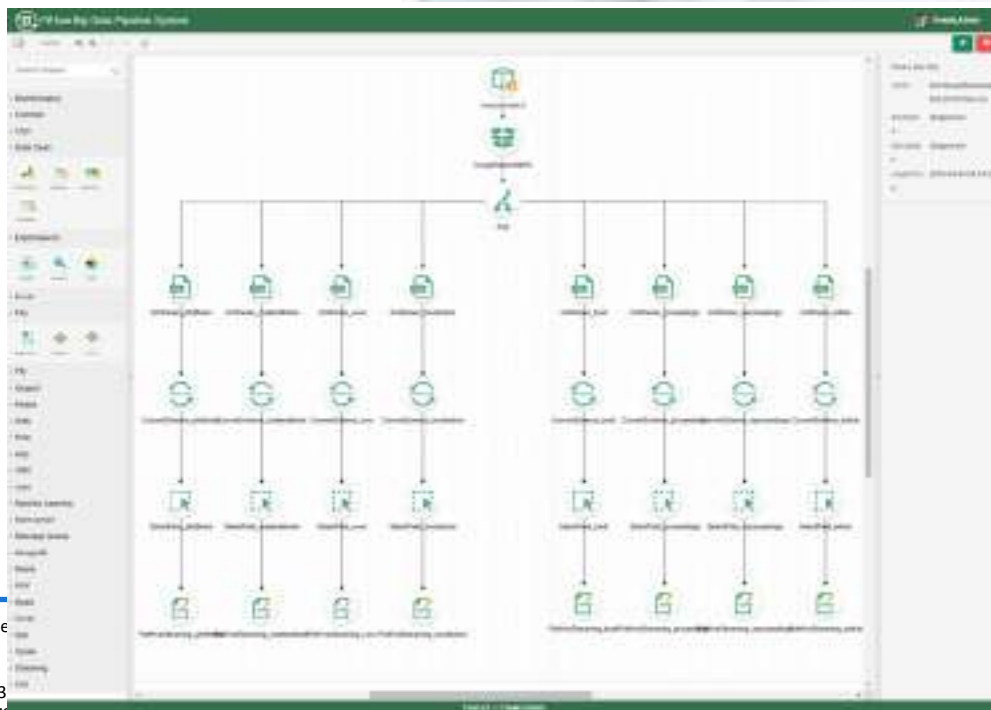
自动及半自动的数据采集、处理流水线机制



PiFlow 架构



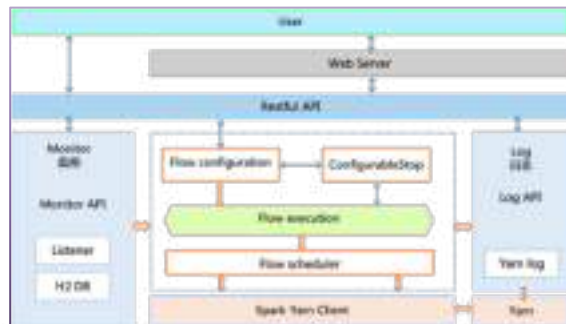
流水线实例



1. 乔子越,杜一*,傅衍杰,王鹏飞,周园春,Unsupervised Author Disambiguation using Heterogeneous Graphs
2. 2019年“智源-AMiner姓名排歧大赛”第一名
3. PiFlow: 码云2019年最有价值开源项目; 中国科技云首届开源大赛二等奖
4. 杜一,王寒雪,乔子越,周园春,一种基于网络表征和语义表征的同名作者消歧方法,20191113
5. 杜一,乔子越,周园春,一种基于异质图卷积神经网络嵌入的作者名字消歧方法,2019106357994,2019
6. 杜一,乔子越,周园春,一种基于异质网络嵌入的学者名字消歧方法,2018112671819,2018
7. 中国专利: 一种大数据ETL任务的编排方法与系统. 完成人: 朱小杰,沈志宏,杜一,赵子豪,周园春,2019
8. 中国专利: 一种大数据ETL任务的调度方法. 完成人: 朱小杰,沈志宏,杜一,赵子豪,周园春,2018

科技领域知识图谱：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

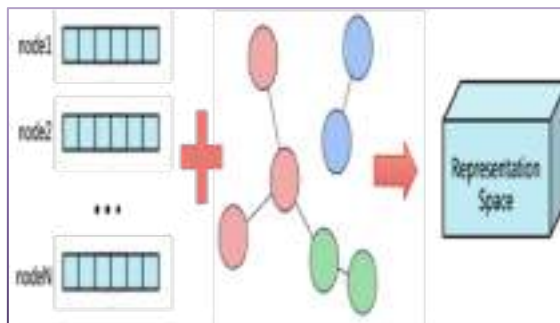


流水线实例

基于统计规则及深度学习方法的实体融合方法

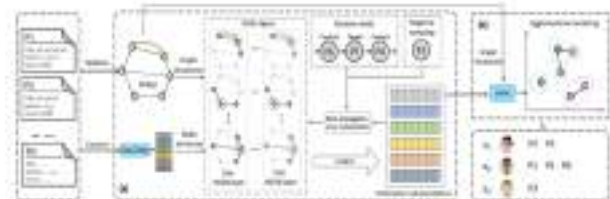


基于规则组合的实体融合



基于GraphEmbedding实体融合

Name	Our method	Component	Zhang et al. 2018	Ni et al. 2018	Zhang et al. 2017
Ajay Gupta	0.750	0.729	0.568	0.552	0.618
Alok Gupta	1	0.690	0.689	0.592	0.580
Ben Ya	0.696	0.762	0.444	0.585	0.610
David Cooper	0.900	0.127	0.737	0.884	0.931
David Nalson	0.944	0.219	0.750	0.735	0.556
Fai Su	1	0.648	0.932	0.630	0.941
Hao Wang	0.694	0.686	0.403	0.557	0.543
Jie Tang	0.982	0.883	0.657	0.522	0.910
Thomas Wolf	0.860	0.502	0.705	0.522	0.352
Yang Wang	0.548	0.118	0.273	0.874	0.409
Avg.	0.796	0.507	0.715	0.681	0.680



1. 乔子越, 杜一*, 傅衍杰, 王鹏飞, 周园春, Unsupervised Author Disambiguation using Heterogeneous Graph Convolutional Network Embedding, IEEE Big Data, 2019
2. 2019年“智源-AMiner姓名排歧大赛”第一名
3. PiFlow: 码云2019年最有价值开源项目; 中国科技云首届开源大赛二等奖
4. 杜一, 王寒雪, 乔子越, 周园春, 一种基于网络表征和语义表征的同名作者消歧方法, 2019113223833 | PCT/CN2019/128642, 2019
5. 杜一, 乔子越, 周园春, 一种基于异质图卷积神经网络嵌入的作者名字消歧方法, 2019106357994, 2019
6. 杜一, 乔子越, 周园春, 一种基于异质网络嵌入的学者名字消歧方法, 2018112671819, 2018
7. 中国专利: 一种大数据ETL任务的编排方法与系统. 完成人: 朱小杰, 沈志宏, 杜一, 赵子豪, 周园春, 2019
8. 中国专利: 一种大数据ETL任务的调度方法. 完成人: 朱小杰, 沈志宏, 杜一, 赵子豪, 周园春, 2018

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的高影响力人员关系网络挖掘

学科交叉性评估及学科趋势预测方法

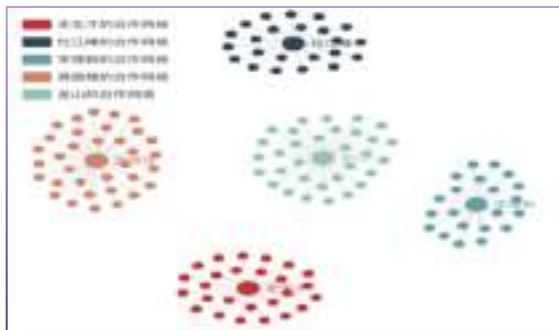


基于统计指标的学科交叉性评价

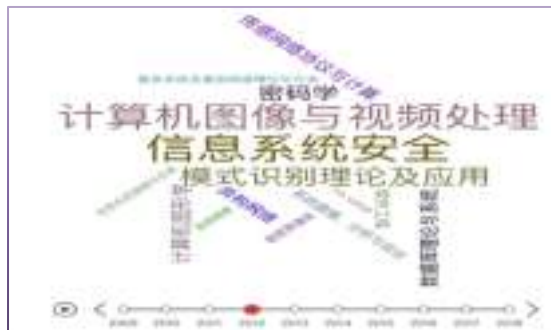
学术影响力评价方法



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+标签传播的科研社区发现



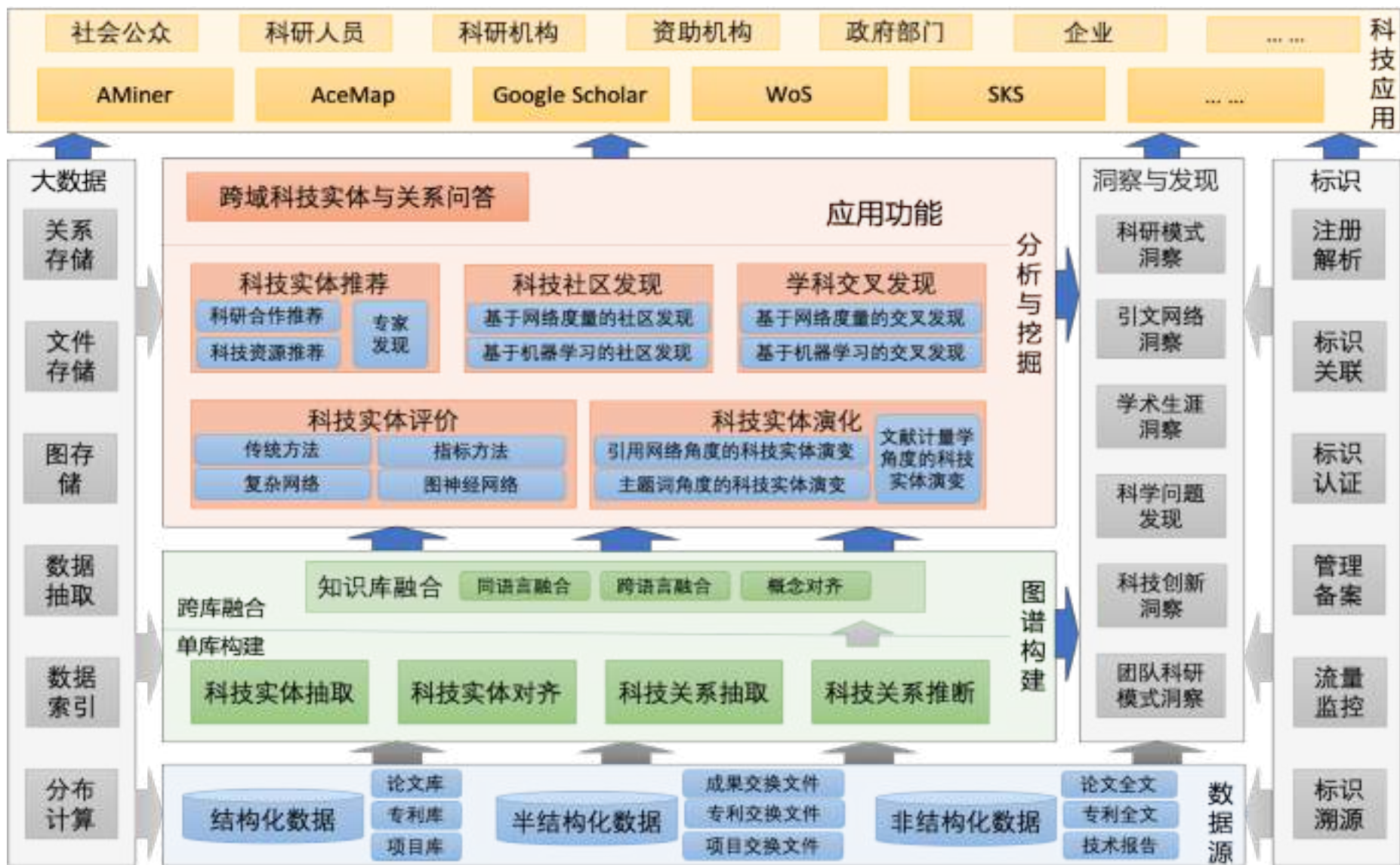
基于时序+图数据库的学科趋势评估



基于网络度量指标+计量学的科研项目影响力评价

1. 王卫军, 崔文娟, 杜一*, 周园春, 基于词嵌入的国家自然科学基金学科交叉知识点分析方法, 情报学报, 2020. (in submission)
2. 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, 中国科学基金, 2019.

科技领域知识图谱：研究架构



请批评指正！



AMiner

— Author-centric academic search and mining

Peng Zhang
Tsinghua University

Academic search is far from sufficient...

A screenshot of a Google search results page for the query "data mining". The page shows a list of search results with titles, snippets, and links. The first result is "Advances in knowledge discovery and data mining" by L. De Raedt, J. G. Peacock, and J. G. Peacock, published in 1998. Other results include "Data Mining: Techniques" by Ian H. Witten and Mark A. H. Wright, and "Data Mining: Introduction to the Basic Concepts" by David Hand and Michael Karmel. The page also shows a sidebar with "Related searches" and "More results" links.

A screenshot of the Academic Search Engine results page for the query "data mining". The page shows a list of search results with titles, snippets, and links. The first result is "Data Mining: Introduction to the Basic Concepts" by David Hand and Michael Karmel, published in 1998. Other results include "Data Mining: Techniques" by Ian H. Witten and Mark A. H. Wright, and "Data Mining: Introduction to the Basic Concepts" by David Hand and Michael Karmel. The page also shows a sidebar with "Related searches" and "More results" links. A line graph is visible on the right side of the page, showing the number of publications over time from 1980 to 2010.

Academic search is far from sufficient...

A screenshot of a Google search for 'data mining'. The results page shows several links, including 'Advances in knowledge discovery and data mining', 'Data Mining: Concepts and Techniques', and 'The WEKA data mining software: an update'. The search bar at the top shows 'data mining' and the Google logo. The left sidebar has filters for 'All results', 'Images', 'Videos', 'Books', 'Maps', 'Shopping', 'More', 'My library', 'My history', 'My activity', 'My settings', 'My account', 'My data', 'My privacy', 'My security', 'My location', 'My device', 'My network', 'My contacts', 'My calendar', 'My tasks', 'My reminders', 'My notes', 'My documents', 'My photos', 'My videos', 'My files', 'My folders', 'My apps', 'My settings', 'My account', 'My data', 'My privacy', 'My security', 'My location', 'My device', 'My network', 'My contacts', 'My calendar', 'My tasks', 'My reminders', 'My notes', 'My documents', 'My photos', 'My videos', 'My files', 'My folders', 'My apps'.

A screenshot of Microsoft Academic search results for 'data mining'. The page shows a list of publications, including 'Advances in knowledge discovery and data mining', 'Data Mining: Concepts and Techniques', and 'The WEKA data mining software: an update'. The search bar at the top shows 'data mining' and the Microsoft Academic logo. The left sidebar has filters for 'All results', 'Images', 'Videos', 'Books', 'Maps', 'Shopping', 'More', 'My library', 'My history', 'My activity', 'My settings', 'My account', 'My data', 'My privacy', 'My security', 'My location', 'My device', 'My network', 'My contacts', 'My calendar', 'My tasks', 'My reminders', 'My notes', 'My documents', 'My photos', 'My videos', 'My files', 'My folders', 'My apps'.

A screenshot of Scopus search results for 'data mining'. The page shows a list of publications, including 'Advances in knowledge discovery and data mining', 'Data Mining: Concepts and Techniques', and 'The WEKA data mining software: an update'. The search bar at the top shows 'data mining' and the Scopus logo. The left sidebar has filters for 'All results', 'Images', 'Videos', 'Books', 'Maps', 'Shopping', 'More', 'My library', 'My history', 'My activity', 'My settings', 'My account', 'My data', 'My privacy', 'My security', 'My location', 'My device', 'My network', 'My contacts', 'My calendar', 'My tasks', 'My reminders', 'My notes', 'My documents', 'My photos', 'My videos', 'My files', 'My folders', 'My apps'.

A screenshot of IEEE Xplore search results for 'data mining'. The page shows a list of publications, including 'Advances in knowledge discovery and data mining', 'Data Mining: Concepts and Techniques', and 'The WEKA data mining software: an update'. The search bar at the top shows 'data mining' and the IEEE Xplore logo. The left sidebar has filters for 'All results', 'Images', 'Videos', 'Books', 'Maps', 'Shopping', 'More', 'My library', 'My history', 'My activity', 'My settings', 'My account', 'My data', 'My privacy', 'My security', 'My location', 'My device', 'My network', 'My contacts', 'My calendar', 'My tasks', 'My reminders', 'My notes', 'My documents', 'My photos', 'My videos', 'My files', 'My folders', 'My apps'.

AI Powered Research Career

133,208,365
RESEARCHERS

272,411,089
PUBLICATIONS

8,796,476
CONCEPTS

754,201,878
CITATIONS

Whatever comes to your mind



Advanced

An **author-centric** academic search
and mining system...

IC

COVID-19 Graph

Open Datasets

Knowledge Dashboard

Knowledge Graph

Experts

Academic Rankings

THU AI TR

Topic Trends

Open

Open Data



TOPIC Must Reading Papers...
COVID-19, Knowledge Graph



AI 2000
AI 2000 Most Influential
Scholars



Master Reading Tree
Tools to help scholars study the
evolution of papers



Xisomai's Star Talents
Predict Stars of Tomorrow with
AI

Experts

A.M. Turing Award Winners

Chinese Academy of Sciences Talents

Academic Rankings

Input name to check one's position

Women in AI **new**

THU AI TR

THU AI TR: 人工智能之机器学习 **Hot**

ICLR 2020 1 反事实因果理论如何帮助深度学习? **new**

Expert Search

Finding experts, for “data mining”

with more semantic constraints, e.g., from **USA**, speak **Chinese**, **gender**, and **age**...

The screenshot displays the iMinter Expert Search interface. At the top, a search bar contains the text 'data mining'. Below the search bar, there are filters for 'Expert' and 'Paper'. The 'Expert' filter shows 'Search Results: 4'. The 'Paper' filter shows 'H-index: 1'. The 'Expert' filter also includes a 'Relevance' dropdown set to '17'. The 'Paper' filter includes a 'H-index' dropdown set to '17'. The 'Expert' filter also includes a 'Gender' dropdown set to 'Male (94)' and 'Female (46)'. The 'Paper' filter includes a 'Language' dropdown set to 'Chinese (288)', 'English (194)', 'Greek (27)', 'German (27)', 'French (22)', 'Japanese (26)', 'Korean (14)', and 'Italian (12)'. The 'Paper' filter also includes a 'Location' dropdown set to 'USA (216)', 'China (141)', 'Taiwan (34)', 'Australia (28)', 'Canada (26)', 'Japan (38)', 'Germany (24)', 'Italy (20)', 'Hong Kong (26)', and 'Singapore (26)'. The 'Expert' filter also includes a 'Similar' button. The 'Paper' filter also includes a 'Follow' button. The 'Expert' filter also includes a 'Similar' button. The 'Paper' filter also includes a 'Follow' button. The 'Expert' filter also includes a 'Similar' button. The 'Paper' filter also includes a 'Follow' button.

Demographics: gender, language, location, etc.

Similar Authors

Knowledge about “data mining”

Data mining

Data mining (the analysis step of the “Knowledge Discovery in Databases” process, or KDD), a relatively young and interdisciplinary field of computer science, is the process that attempts to discover patterns in large data sets. It utilizes methods at the intersection of artificial intelligence, machine learning, statistics, and database systems. The overall goal of the data mining process is to extract information from a data set and transform it into an understandable structure for further use. Aside from the raw analysis step, it involves database and data management aspects, data preprocessing, model and inference considerations, interestingness metrics, complexity considerations, post-processing of discovered structures, visualization, and online updating.

Super Concepts:
Data analysis, Data mining, Formal sciences, Applied sciences, Networks, Artificial intelligence

Related Concepts:
Data compression, Data visualization, Natural language processing, Data cleansing, Distributed computing, Informationization, Speech recognition, Business intelligence, Pattern recognition, Spatial databases, Full text search, Metadata, Computer vision, ISAM, Biological neural network, Database, Grid computing, Database marketing, Parallel computing

数据挖掘

数据挖掘(Data Mining)是指从分析每个数据,从大量数据中寻获其规律的技术。主要有数据挖掘、数据挖掘和规律表示三个步骤。数据挖掘的任务有关联分析、聚类分析、分类分析、异常分析、特异组分析和离群点分析等。

上拉词:
资料分析, 计算机科学基础理论, 决策支持系统, 信息管理术语, 数据挖掘, 形式科学

The screenshot displays a researcher's profile page with several key sections and annotations:

- Basic Info:** Located at the top, it includes the researcher's name (Jiawei Han), affiliation (Department of Computer Science, University of Illinois at Urbana-Champaign), title (Professor), contact information (phone, email, website), and a 'Follow' button.
- Research Interests:** A section below the basic info listing areas like Data Mining, Information Extraction, Data Analysis, Machine Learning, and Text Mining, each with a corresponding colored dot.
- Citation Statistics:** A section showing a line graph of citation trends over time (1995-2015) and a list of papers with their citation counts (e.g., 790, 15, 5).
- Ego Network:** A circular network diagram showing connections between the researcher and other authors, with nodes representing authors and edges representing citations or collaborations.
- Background and Experience:** A section detailing the researcher's education (PhD from MIT, MS from UIUC) and professional experience (Professor at UIUC since 2001).
- Publications:** A list of research papers, including 'Entity Linking with a Knowledge Base' and 'Troubleshooting interactive computing'.

Annotations include red boxes highlighting specific areas (Basic Info, Research Interests, Citation Statistics, Ego Network, Background and Experience, Publications) and a red arrow pointing from the Research Interests section to the Ego Network.

AMiner

- ❑ Academic search and mining system — <https://aminer.cn>
 - ❑ Online since 2006
 - ❑ **>136 million** researcher profiles
 - ❑ **>267 million** publication papers
 - ❑ **>341 million** requests
 - ❑ **20M IP access from 220 countries** per month
- ❑ Deep analysis, mining, and search



Technology Overview



Data->Semantics
Search->Intelligence

Tim Berners Lee
WWW Inventor
Turing Award Winner

Intelligence

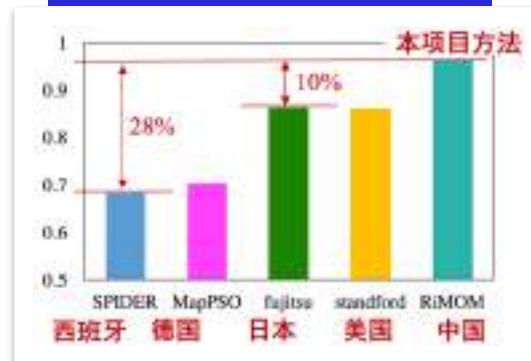
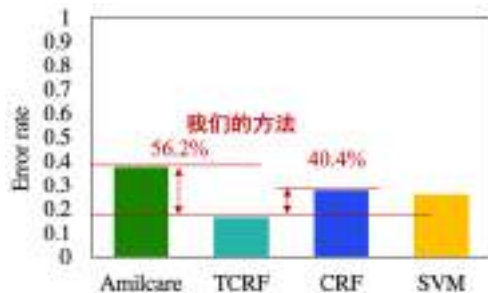
Linking

**Recommendation
accuracy +161%**

Extraction

**15 champions in
the past 7 years**

**Error rate reduced
40-56%**



Apply to
Tencent
Networks

+161%

Prior

Our approach

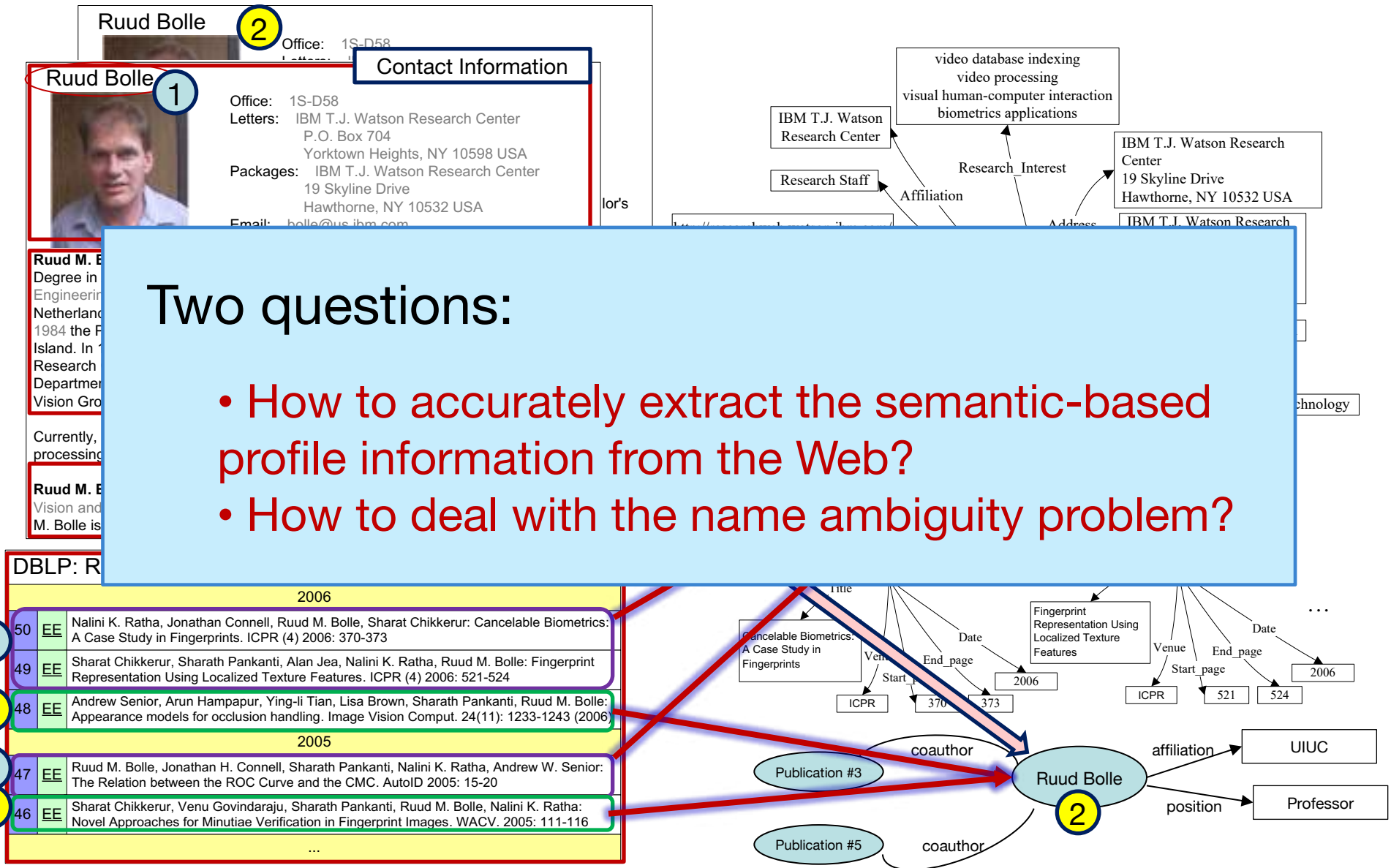
Reported by UN



#citation>8,000, published ~20 papers on KDD

Extracting Profile Semantics from the Web

(ACM TKDD, WWW'12, ISWC'06, ICDM'07, ACL'07)



A Big Data Approach

- Search Engine as Data Source

The image displays two side-by-side screenshots of Google search results, illustrating a 'Big Data Approach' by aggregating information from multiple sources.

Left Screenshot: Search for 'philip s. yu email'

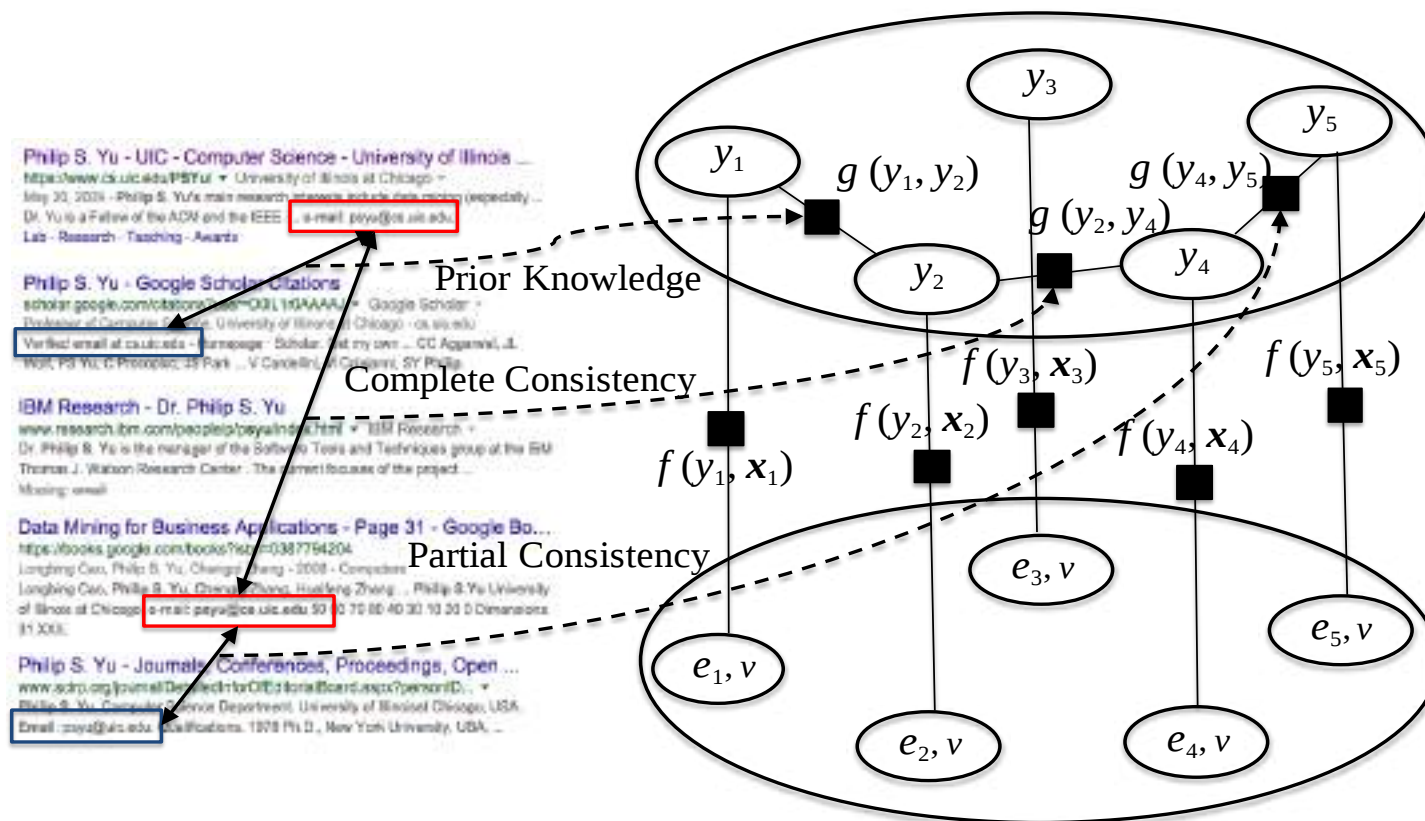
- Search results for 'philip s. yu email' (About 951,000 results).
- Results include links to Philip S. Yu's University of Illinois page, Google Scholar Citations, Wikipedia, dblp, and Jiawei Han's page.
- Red boxes highlight the email address `psyu@cs.uic.edu` in the University of Illinois page, the Google Scholar Citations page, the Wikipedia page, and the dblp page.

Right Screenshot: Search for 'philip s. yu affiliation'

- Search results for 'philip s. yu affiliation' (About 516,000 results).
- Results include links to dblp: Philip S. Yu, Google Scholar Citations, Wikipedia, Link Mining: Models, Algorithms, and Applications, Discovering High-Order Periodic Patterns, On clustering massive text and categorical data streams, and An Introduction to Privacy-Preserving Data Mining.
- Red boxes highlight the affiliation 'University of Illinois at Chicago' in the dblp: Philip S. Yu page, the Google Scholar Citations page, the Wikipedia page, the Link Mining page, the Discovering High-Order Periodic Patterns page, the On clustering massive text and categorical data streams page, and the An Introduction to Privacy-Preserving Data Mining page.

MagicFG: Markov Logic Factor Graph

- ✓ Depict and utilize correlations between possible candidates from redundant data.
- ✓ Incorporate human knowledge to guide and amend the classification model.





Jeannette Wing

Computer Science Department Carnegie Mellon University
 Professor of Computer Science
 412-260-8926 (cell)
wing@cs.cmu.edu
<http://www.cs.cmu.edu/~wing/>

External Links

Follow

Update



Research Interests

Formal Methods Software Engineering Formal Verification
 Embedded Systems Specification Languages

1982 1985 1990 1995 2000 2005 2009



Similar Authors

Ego Network

Overview

Papers 24

Merge

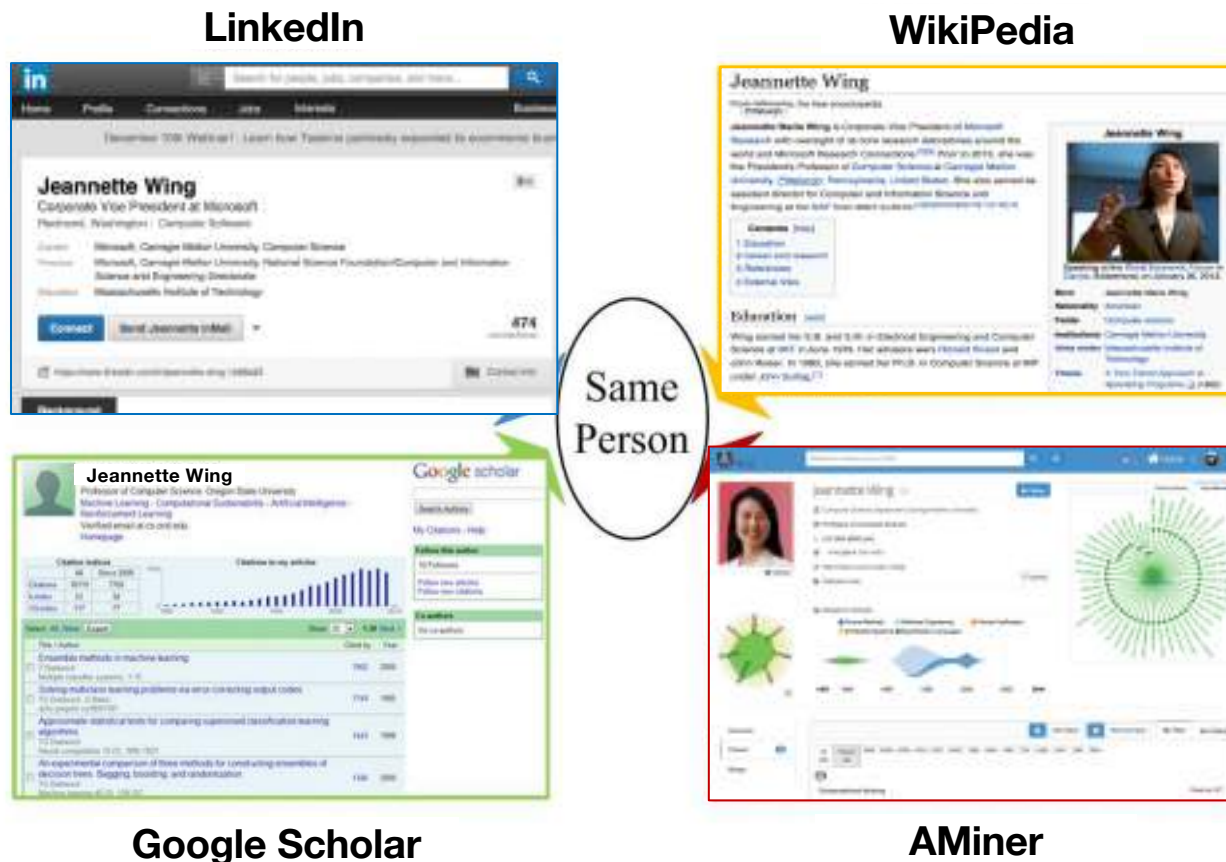
24

Computational thinking

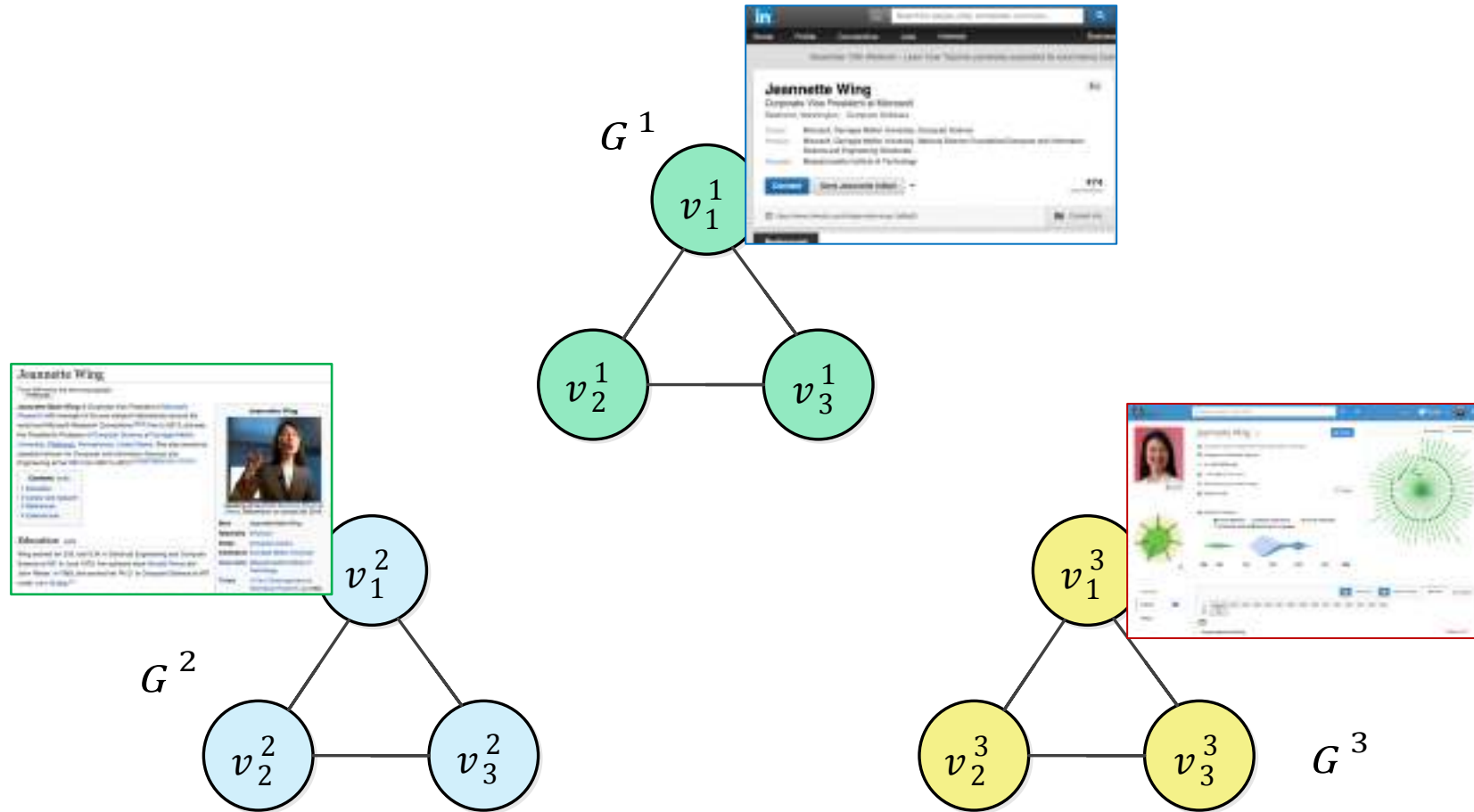
Cited by 1377

Linking Semantics across Networks

- Identifying users from multiple heterogeneous networks and integrating semantics from the different networks together.

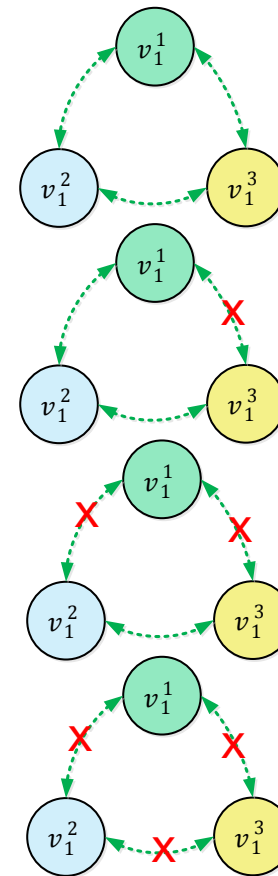
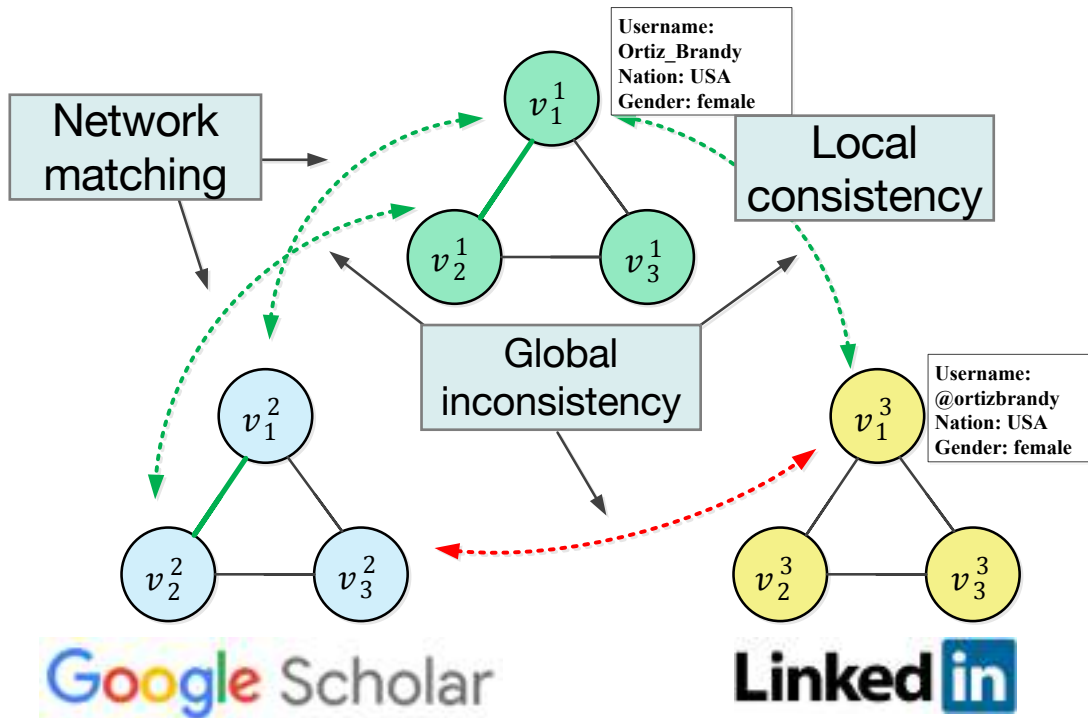


Considering the networks...



Local + Global consistency

AMiner



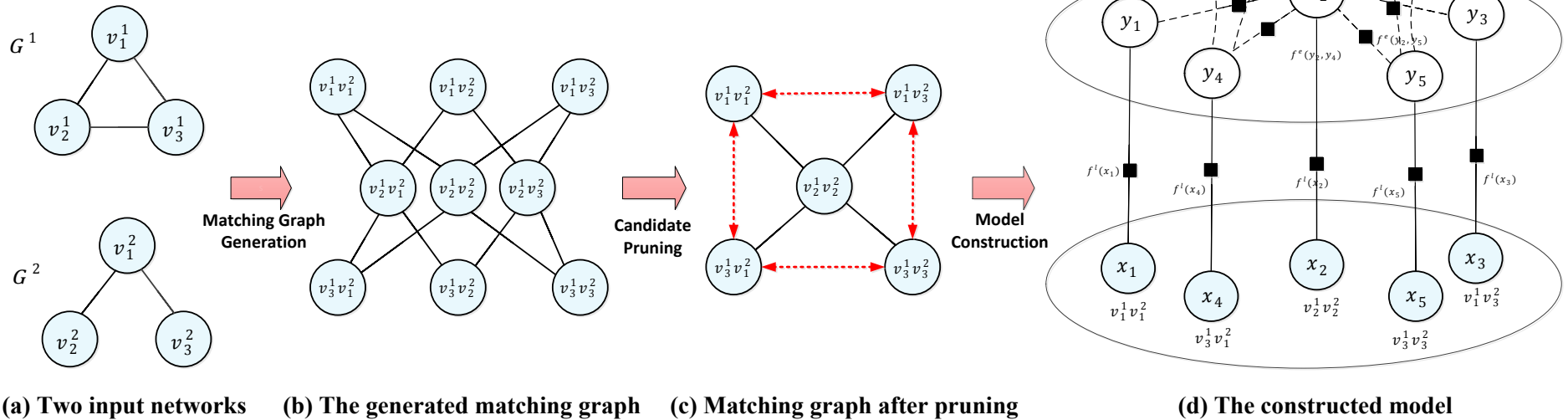
Inconsistent!



COSNET: Connecting Social Networks with Local and Global Consistency

- **Input:** $\mathbf{G} = \{G^1, G^2, \dots, G^m\}$, with $G^k = (V^k, E^k, R^k)$
- **Formalization:** $\mathbf{X} = \{x_i\}$, all possible pairwise matchings and each corresponds to $y_i \in \{1, 0\}$
- **COSNET:** an energy-based model
$$Y^* = \arg \max E(Y, X)$$

Model Construction



Objective function by combining all the energy functions

$$\begin{aligned}
 E(Y, X) = & \sum_{\mathbf{x}_i \in V_{MG}} \mathbf{w}_l^T \mathbf{g}_l(\mathbf{x}_i, y_i) + \sum_{\langle \mathbf{x}_i, \mathbf{x}_j \rangle \in E_{MG}} \mathbf{w}_e^T \mathbf{f}_e(y_i, y_j) \\
 & + \sum_{c \in T_{MG}} \mathbf{w}_t^T \mathbf{f}_t(Y_c)
 \end{aligned} \tag{2}$$

AMiner

Jiawei Han (韩家炜)
 Department of Computer Science, University of Illinois at Urbana-Champaign
 Professor
 (617) 333-6903
 hanj@cs.uiuc.edu
 http://www.cs.uiuc.edu/~hanj/

Google Scholar
 CiteSpace Index: 130, 500, 1000
 H-index: 130, 500, 1000
 i10-index: 130, 500, 1000
 h-index: 130, 500, 1000
 i10-index: 130, 500, 1000

Background
 Experience
 Professor
 UIUC
 August 2001 – Present (14 years 7 months)

Research Interests
 Data Mining, Information Extraction, Machine Learning, Text Mining

Papers
 790
 Recent: 20, 1994, 1990, 1990

Bringing Structure to Text: Mining Phrases, Entity Concepts, Topics, and Hierarchies
 20th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD), New York 2014

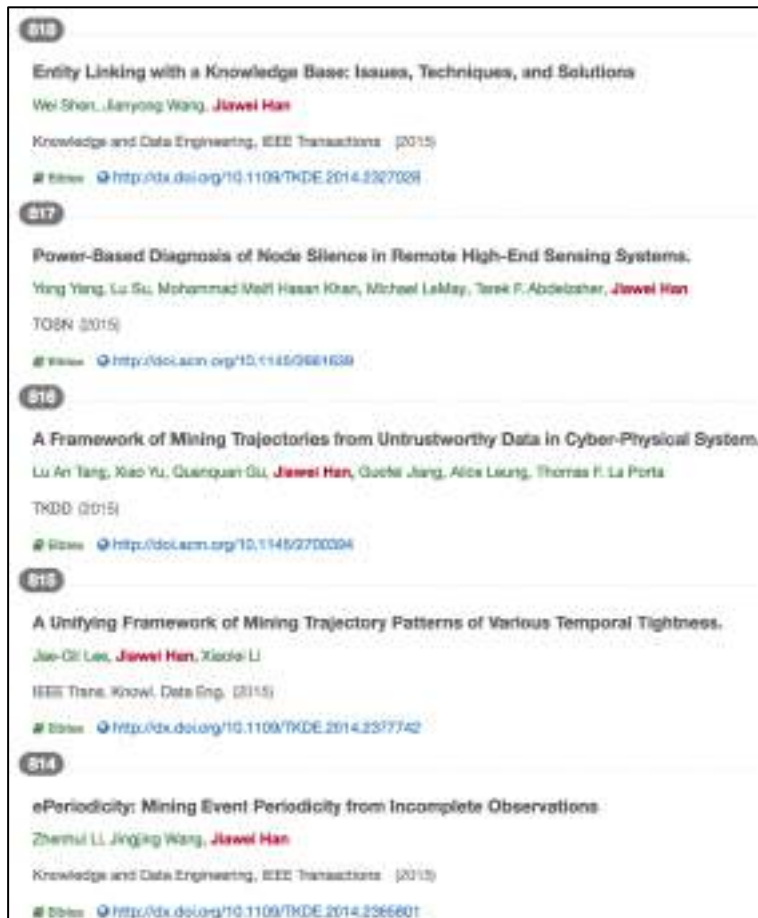
Mining Massive RFID, Trajectory, and Traffic Data Sets
 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD), Las Vegas 2008

Systems and Methods for Detecting a Novel Data Class
 Mohammad Mehedy Masud, Latifur Rahman Khan, Bhavani Marianne Thurasingham, Qing Chen, Jing Gao, Jiawei Han
 Publication-date: 2012-03-01 Application-date: 2011-08-22

Troubleshooting interactive complexity bugs in wireless sensor networks using data mining techniques
 Mohammad Moll Hasan Khan, Hieu Khan Lu, Hossain Ahmad, Tarek F. Abdelzaher, Jiawei Han
 ACM Transactions on Sensor Networks (TOSN) (2014)

Discovering Expertise Semantics

- Quantifying researchers' expertise using publications



Entity Linking with a Knowledge Base: Issues, Techniques, and Solutions
Wei Shen, Jianping Wang, **Jiawei Han**
Knowledge and Data Engineering, IEEE Transactions (2015)
BibTeX <http://dx.doi.org/10.1109/TKDE.2014.2327029>

Power-Based Diagnosis of Node Silence in Remote High-End Sensing Systems.
Ying Ying, Lu Su, Mohammad Walei Hassan Khan, Michael LeMay, Tarek F. Abdelzaher, **Jiawei Han**
TOSN (2015)
BibTeX <http://dx.doi.org/10.1145/2661639>

A Framework of Mining Trajectories from Untrustworthy Data in Cyber-Physical Systems.
Lu An Tang, Xiao Yu, Quanquan Gu, **Jiawei Han**, Guohui Jiang, Alice Luang, Thomas F. La Porta
TKDD (2015)
BibTeX <http://dx.doi.org/10.1145/2700094>

A Unifying Framework of Mining Trajectory Patterns of Various Temporal Tightness.
Jiao-Gil Lee, **Jiawei Han**, Xiaohu Li
IEEE Trans. Knowl. Data Eng. (2015)
BibTeX <http://dx.doi.org/10.1109/TKDE.2014.2327742>

ePeriodicity: Mining Event Periodicity from Incomplete Observations
Zhenhui Li, Jingjing Wang, **Jiawei Han**
Knowledge and Data Engineering, IEEE Transactions (2015)
BibTeX <http://dx.doi.org/10.1109/TKDE.2014.2366601>



Similar

Jiawei Han (韩家炜)

✓ Following | 87

H-Index: 133 | #Paper: 818 | #Citation: 111184

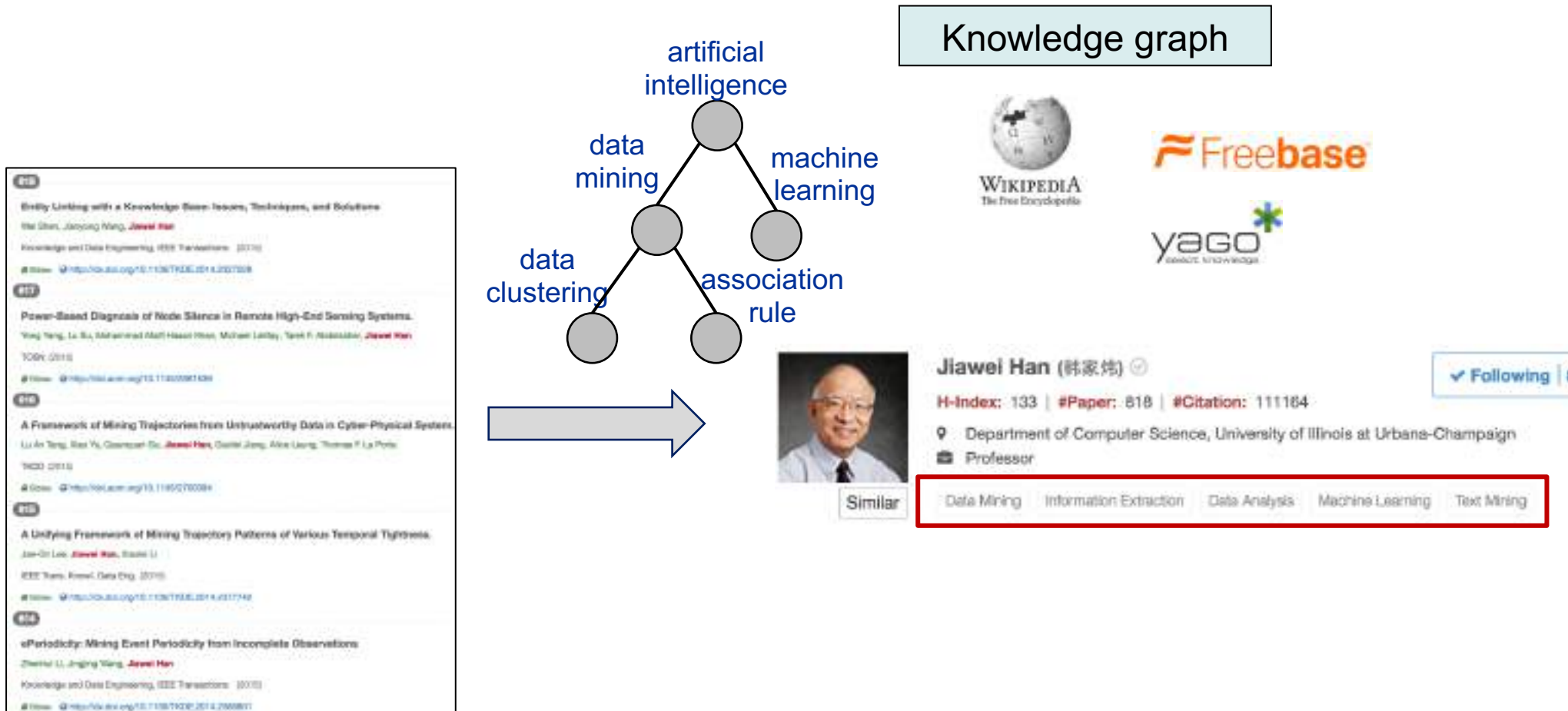
Department of Computer Science, University of Illinois at Urbana-Champaign

Professor

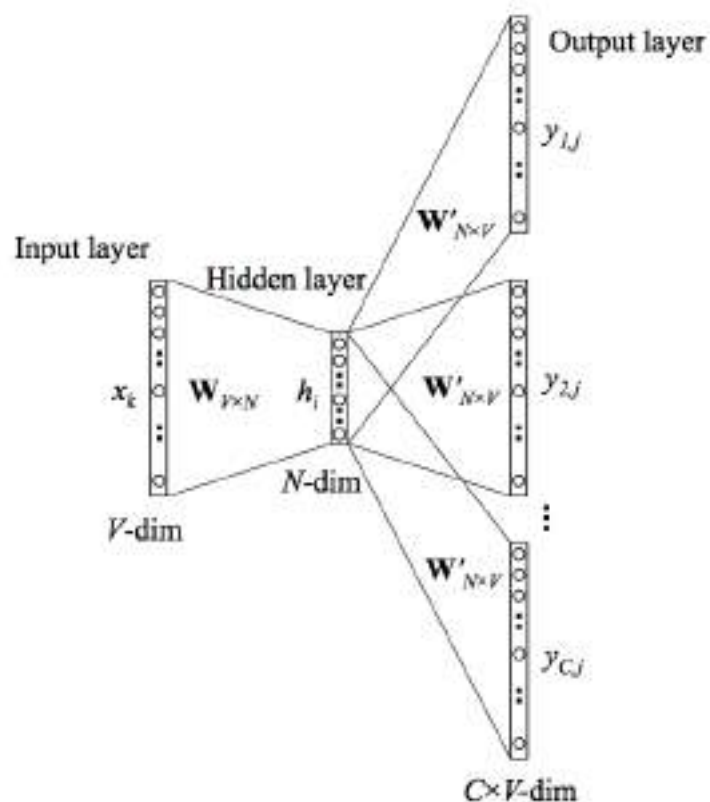
Data Mining | Information Extraction | Data Analysis | Machine Learning | Text Mining

The straightforward method is to extract high frequent terms as expertise; however, the extracted terms may be not good terms to represent expertise.

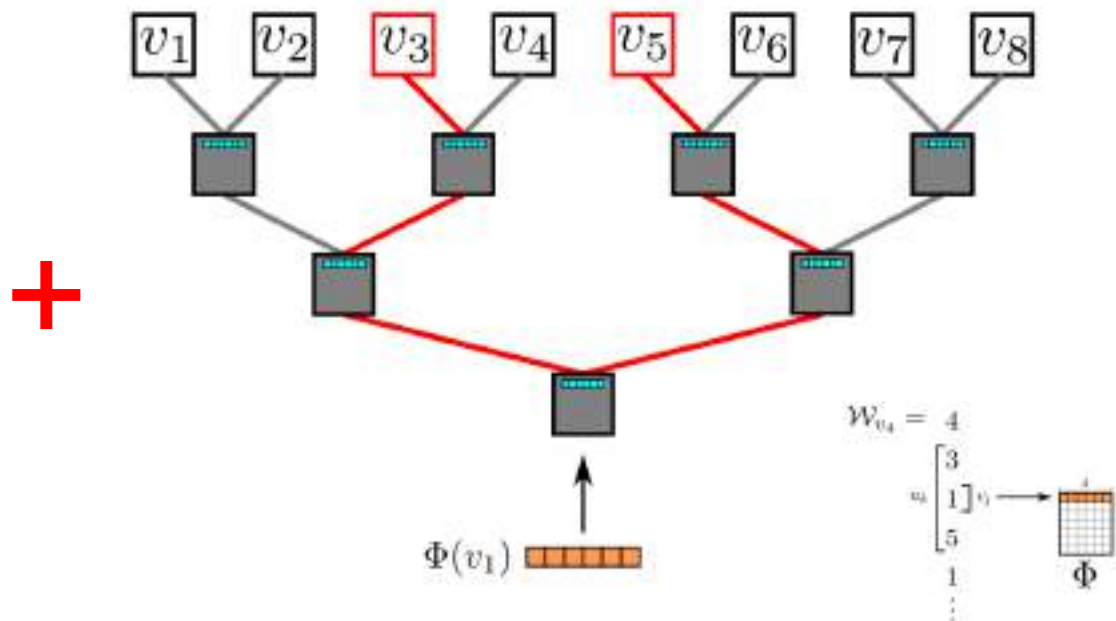
Knowledge-driven Semantic Mining



Embedding Knowledge and Networks

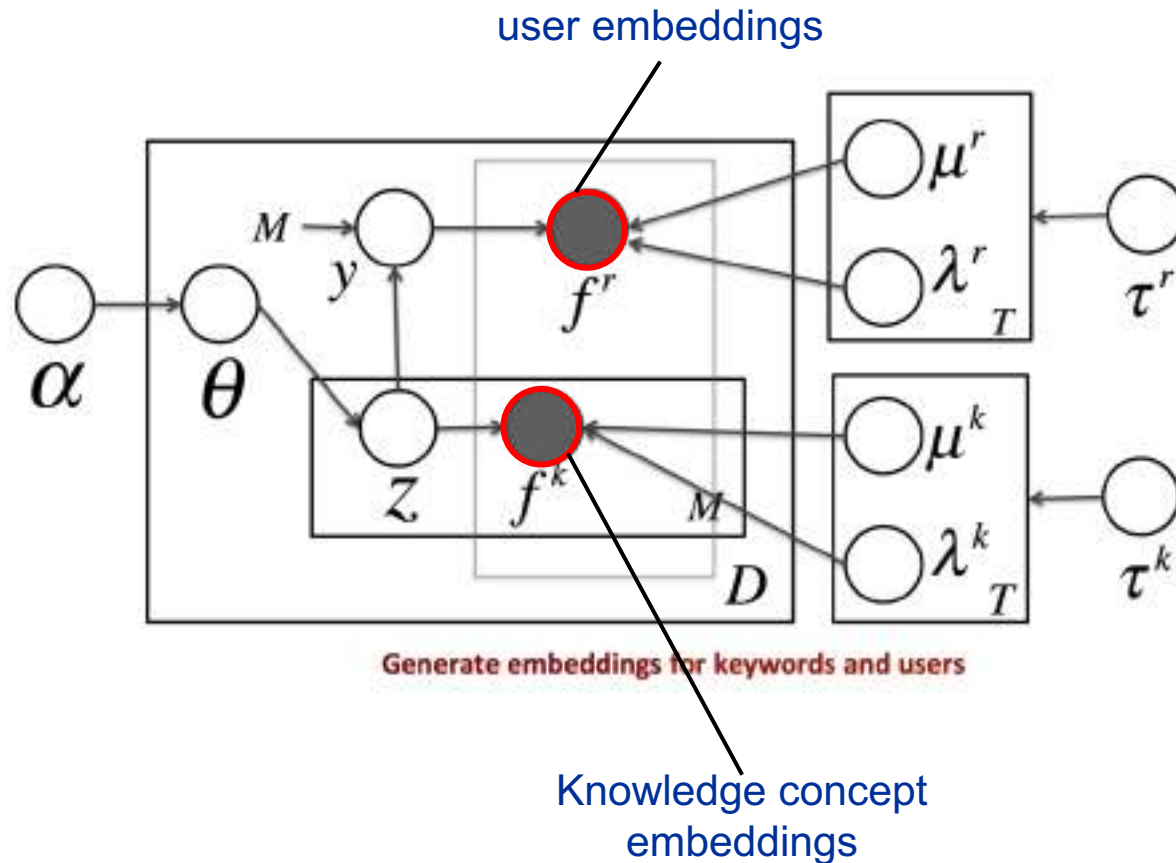


Knowledge concept embedding



Network-based researcher embedding

GenVector—bridging researcher network and knowledge graph



Examples

Case study: Dan Klein

GenVector	AM-base
Language models	Machine translation
Markov models	Word alignment
Probabilistic models	Bleu score
Natural language	Best result
Coreference resolution	Language model

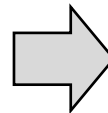
Case study: Xiaoou Tang

GenVector	AM-base
Feature extraction	Face recognition
Image segmentation	Face image
Image matching	Novel approach
Image classification	Line drawing
Face recognition	Discriminant analysis

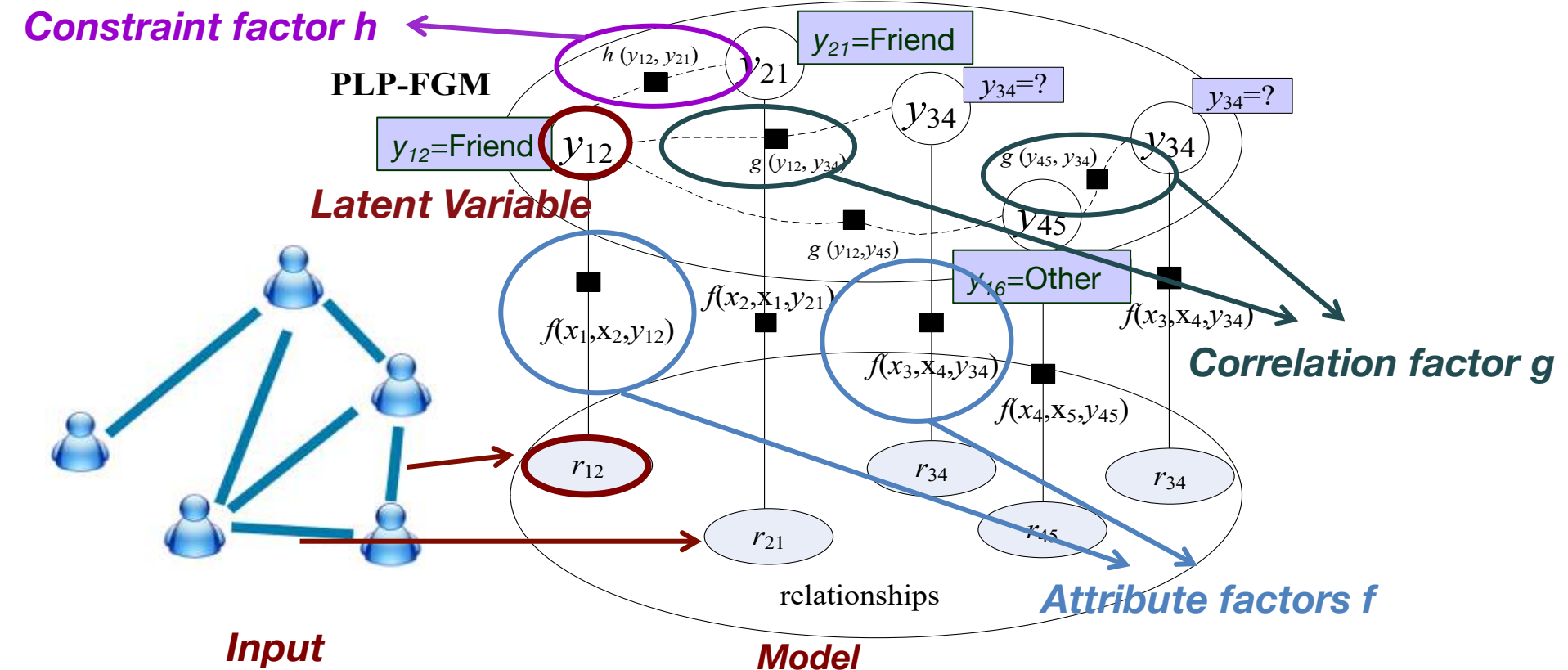
GenVector—the proposed Bayesian embedding method;
AM-base—previous language modeling method.

Discovering Network Semantics

- How to mine semantics within networks?
 - Who are Jiawei Han's students and advisor?



Partially Labeled Pairwise Factor Graph Model (PLP-FGM)



Map relationship to nodes in model

Example	Example:
#Coauth	Author A has a longer publication history than both B and C.

AMiner

The screenshot displays the AMiner profile for Jiawei Han (韩家炜). The profile includes a header with the AMiner logo, a search bar, and navigation links. The main content area features a profile picture, a bio, and a list of research interests. A circular chart visualizes these interests, with segments for Data Mining, Information Extraction, Data Analysis, Machine Learning, and Text Mining. Below the chart is a timeline showing the evolution of these interests from 1985 to 2015. To the right, a red-bordered box highlights the 'Ego Network' visualization, which shows a radial network of connections between Jiawei Han and other researchers. The bottom section of the page lists publications, including 'Entity Linking with a Knowledge Base: Issues, Techniques, and Solutions' (2015), 'Mining Graph Patterns' (2016), and 'Troubleshooting interactive complexity bugs in wireless sensor networks using data mining techniques' (2014).

AMiner

Whatever comes to your mind

Home

Jiawei Han (韩家炜)

Follow

Department of Computer Science, University of Illinois at Urbana-Champaign

Professor

(617) 333-6903

hanj@cs.uiuc.edu

<http://www.cs.uiuc.edu/~hanj/>

External Links

Research Interests

Data Mining Information Extraction Data Analysis Machine Learning Text Mining

1985 1990 1995 2000 2005 2010 2015

Similar Authors

Ego Network

Ego Network

About

Papers 790

Lectures 15

Patents 5

790

Entity Linking with a Knowledge Base: Issues, Techniques, and Solutions

Hui Shen, Jinyong Wang, Jiawei Han

Knowledge and Data Engineering, IEEE Transactions (2015)

Item <http://dx.doi.org/10.1109/TKDE.2014.2321526>

799

Mining Graph Patterns

Hong Chang, Meng Yan, Jiawei Han

Frequent Pattern Mining (2016)

Item http://dx.doi.org/10.1007/978-1-4939-6045-0_13

798

Troubleshooting interactive complexity bugs in wireless sensor networks using data mining techniques

Mohammad Mufi Hasan Khan, Hieu Khan Lu, Hossain Ahmad, Tanek F. Abdelzaher, Jiawei Han

ACM Transactions on Sensor Networks (TOSN) (2014)

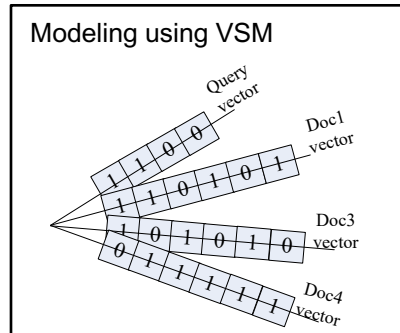
Item <http://dx.doi.org/10.1145/2552990>

797

Search in AMiner



Search with keyword



Return

[Principles of Data Mining.](#)

DJ Hand - Drug Safety, 2007 - [drugsafety.adisonline.com](#)

[Advances in Knowledge Discovery and Data Mining](#)

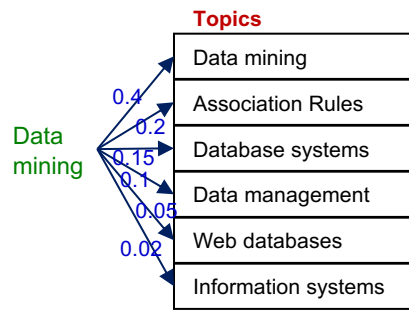
UM Fayyad, G Piatetsky-Shapiro, P Smyth, R...

[Data Mining: Concepts and Techniques](#)

J Han, M Kamber - 2001...

Search with semantic modeling

Modeling using semantic topics



Return

Experts

Search Results for: data mining

Experts

Jiawei Han (H-index: 125 | iPaper: 790 | iCitation: 50401) Department of Computer Science, University of Illinois at Urbana-Champaign Professor Data Mining Information Extraction Data Analysis Machine Learning

Philip S. Yu (H-index: 104 | iPaper: 838 | iCitation: 65438) Department of Computer Science, University of Illinois at Chicago Professor Distributed System Query Optimization Query Processing Database Systems

Hilol Kargupta (H-index: 42 | iPaper: 141 | iCitation: 9150) Department of Computer Science and Electrical Engineering University of Maryland Baltimore County Associate Professor Data Mining Machine Learning Data Analysis Knowledge Discovery Genetic Algorithms

Xindong Wu (H-index: 48 | iPaper: 591 | iCitation: 3944) Department of Computer Science, University of Wisconsin Professor Machine Learning Information Extraction Bayesian Networks Data Mining Supervised Learning

Data mining

Data mining (the analysis step of the "Knowledge Discovery in Databases" process, or KDD), a relatively young and interdisciplinary field of computer science, is the process that attempts to discover patterns in large data sets. It utilizes methods at the intersection of artificial intelligence, machine learning, statistics, and database systems. The overall goal of the data mining process is to extract information from a data set and transform it into an understandable structure for further use. Aside from the raw analysis step, it involves database and data management aspects, data preprocessing, model and inference construction, interpretability, complexity, model and inference construction, post-processing of discovered structures, visualization, and online updating.

Super Concepts:

Data analysis Data mining Formal sciences Applied sciences Networks Artificial intelligence

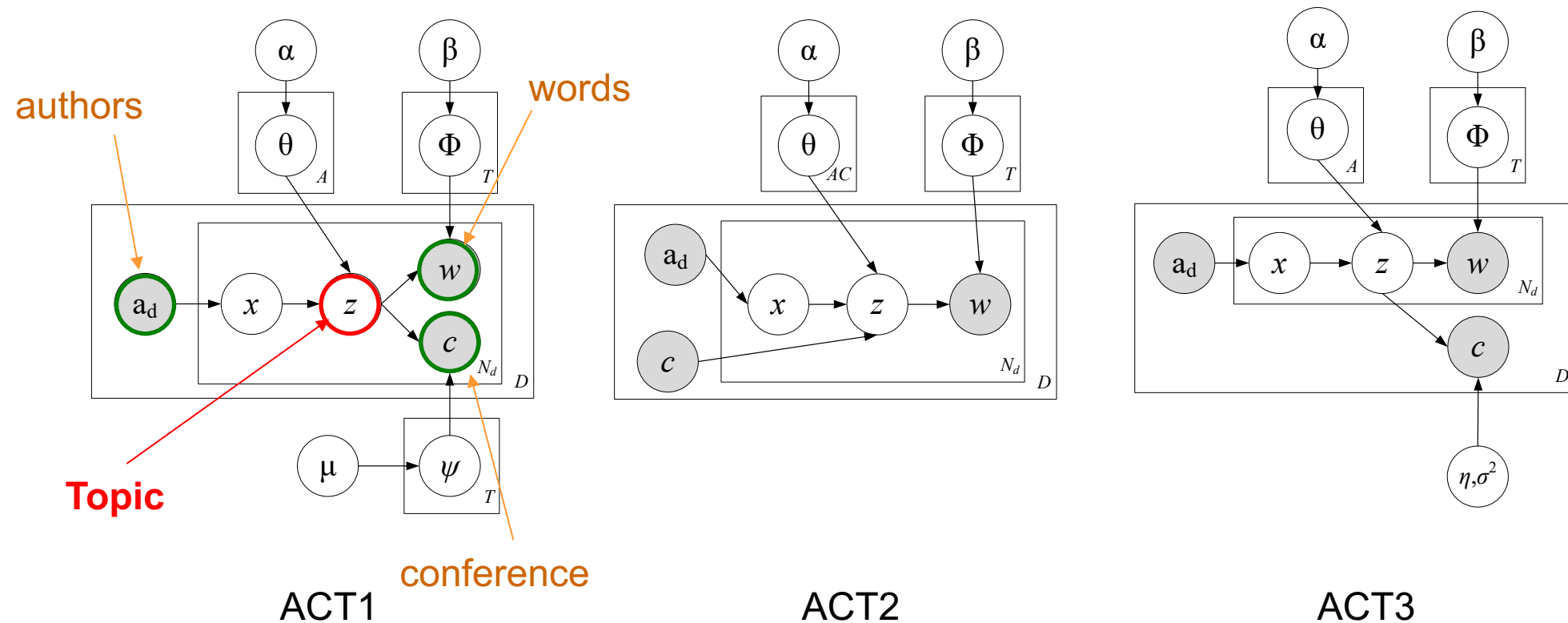
Related Concepts:

Data compression Data visualization Natural language processing Data cleansing Distributed computing Informationization Speech recognition Business intelligence Pattern recognition Spatial databases Full text search Metadata Computer vision ISAM Biological neural network Database Grid computing Database marketing Parallel computing

数据挖掘

数据挖掘(Data Mining)是指从海量数据中, 提取出隐含且先前未知的, 对决策有潜在价值的信息的过程。数据挖掘技术, 主要指数据库系统、统计方法和机器学习三个分支。数据挖掘的世界有两大分支: 分类分析、聚类分析、异常分析、特征提取分析和决策分析等。

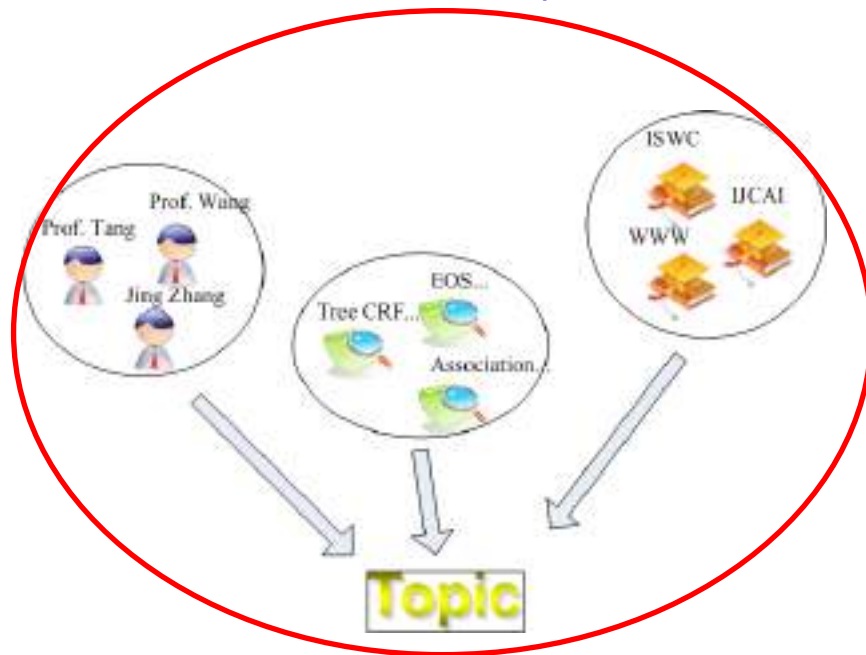
Modeling the Academic Network



Author-Conference-Topic Model [Tang et al., 08]

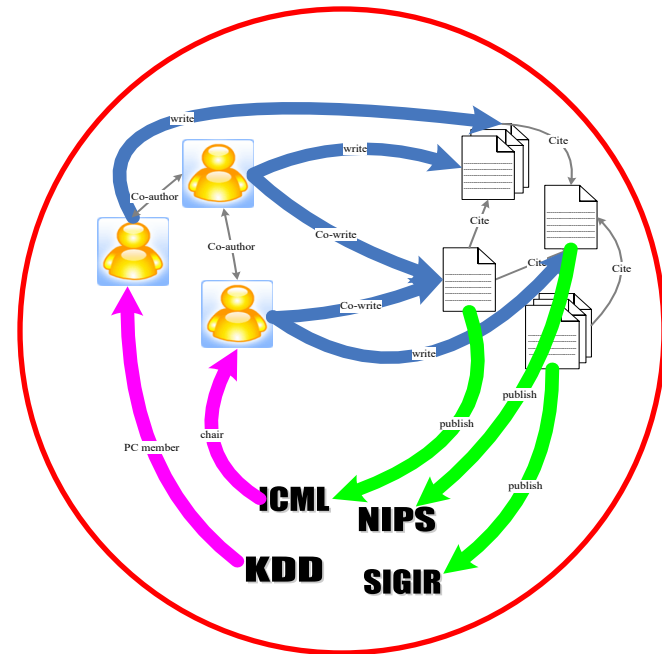
Integrating Topic Model into Random Walk

Modeling academic network with topics

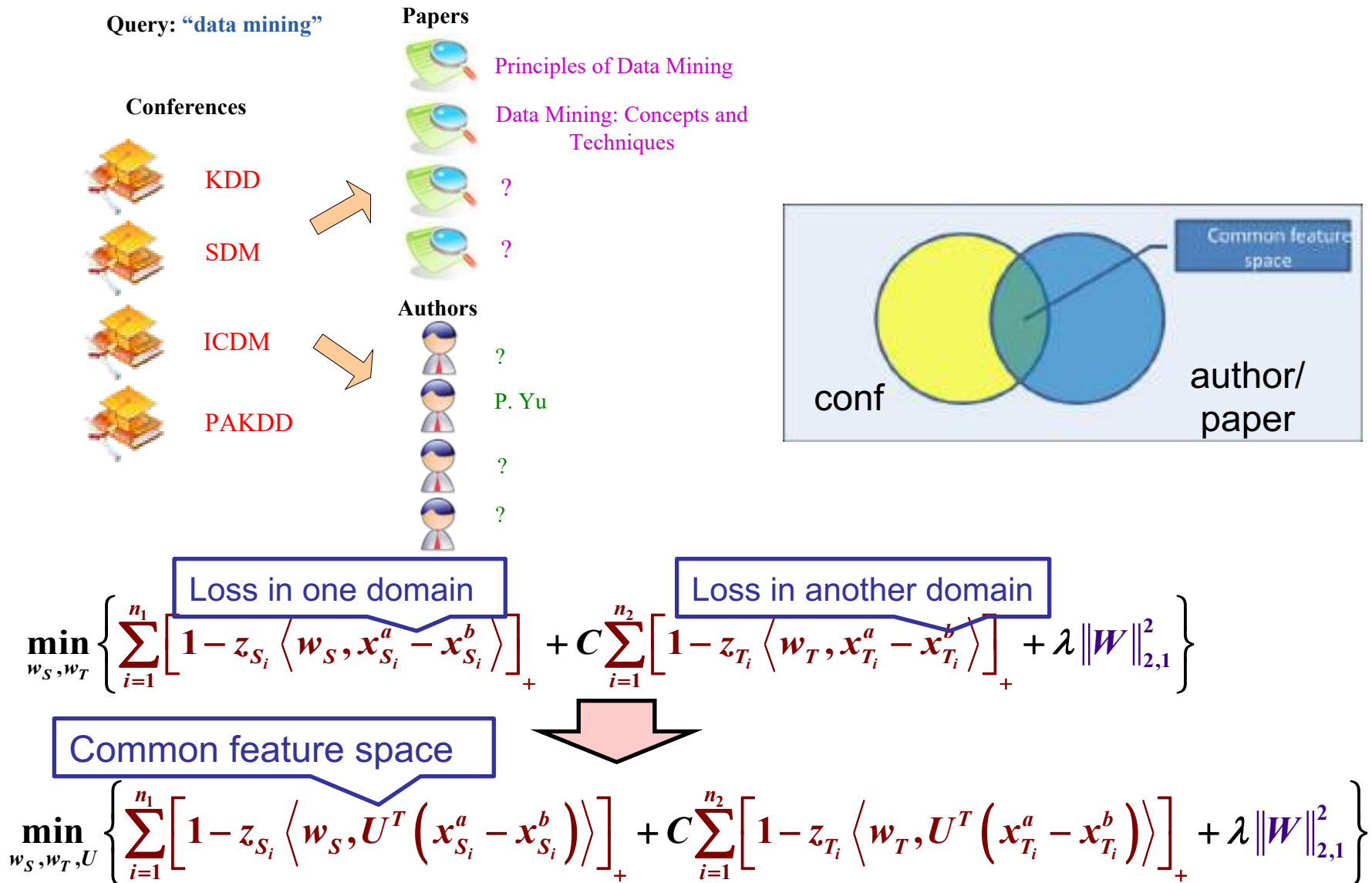


Author-Conference-Topic Model [Tang et al., 08]

Random walk over the academic network



Heterogeneous Cross-domain Ranking



Results on Heterogeneous Tasks

- Expert finding verse Bole search (finding best supervisor)
- To obtain ground truth of bole for each query
 - We sent emails to 50 senior researchers and 50 junior researchers (91.6% are post doc or graduates)
 - Average their feedbacks

Table **Results on heterogeneous tasks.**

Approach	P@5	P@10	P@15	MAP	N@5	N@10
RSVM	.7714	.8429	.8285	.7756	.5545	.5947
RSVMt	.8000	.8286	.8476	.7837	.5923	.5999
MTRSVM	.8000	.8286	.8476	.7875	.6140	.6075
HCDRank	.8285	.7857	.8571	.7971	.6189	.6112
Language model	.6250	.6875	.6500	.6726	.3343	.3809



Why you will love AMiner?

- finding collaborations, applying for PhD, proposal reviewing, recruiting...

Related publications: ACM TKDD (2), IEEE TKDE (3), KDD'08-15, WWW'12-15, J. Informetrics, SIGMOD'09, IJCAI'09-15, ICML'14

Examples – Expert search

- When starting a Work on a new research topic;
- Or brainstorming for novel ideas.

Researcher

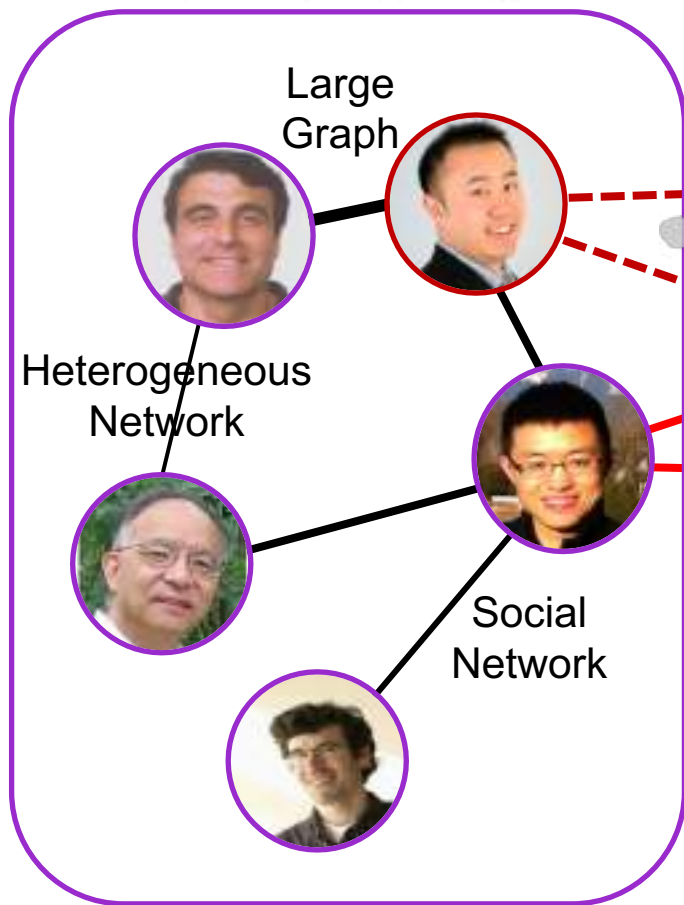


- **Who are experts in this field?**
- **What are the top conferences in the field?**
- **What are the best papers?**
- **What are the top research labs?**

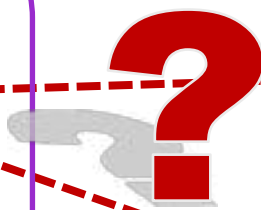
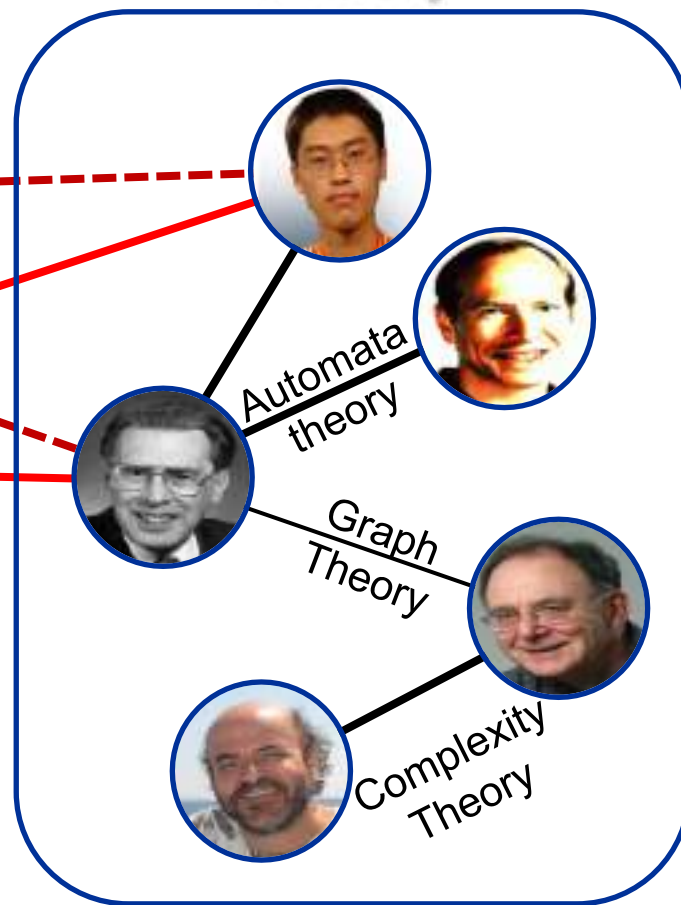


Examples – Collaboration Recommendation

Data Mining



Theory



Examples – More Challenging Questions...

Organization
or Agency



- Looking for researchers
from the U.S. who can
speak Chinese...



- Want to recruit female
researchers working on
WSDM

Data Mining Experts

We use this tool to help MOST (Ministry of Science and Technology of the People's Republic of China) to identify experts in different research fields.

Global Data Mining Experts

This collection includes experts in data mining field. The basic metric is that if a researcher served as DPMC at KDD (the best conference in Data Mining) for more than once, s/he will be considered to be included in this collection.

Jin Tang · 2018-02-18

0 ~ 20 of 310 results

Statistics Extract emails Export Members

Keywords: Name: Organization: Find All Persons

H-index: >=50 (13) 50-59 (20) 40-49 (29) 30-39 (62) 20-29 (71) 10-19 (106) <10 (96)

Gender: Male (362) Female (35)

Language: Chinese (113) English (99) Indian (21) Greek (10) German (12) Japanese (8) French (3) Korean (4)

Location: USA (170) China (26) Hong Kong (14) Germany (11) Canada (11) Australia (10) Singapore (3) Belgium (7)

Preference: H-index: Activity Diversity Rising Star #Citation #Paper AS



Jiawei Han (韩家炜)

Department of Computer Science, University of Illinois at Urbana-Champaign

Professor

Flags: Male Chinese



Philip S. Yu

Department of Computer Science, University of Illinois Chicago

Professor and Weiser Chair in Information Technology

Flags: Male Chinese



Christos Faloutsos

Dept. of Computer Science Carnegie Mellon University

Professor

Flags: Male English



Vipin Kumar

University of Minnesota

Professor

Flags: Male Indian



Rajeev Motwani

Stanford University

Professor and Director of Graduate Studies

Flags: Male English



George Karypis



Jian Pei



Gerhard Weikum



Charu C. Aggarwal

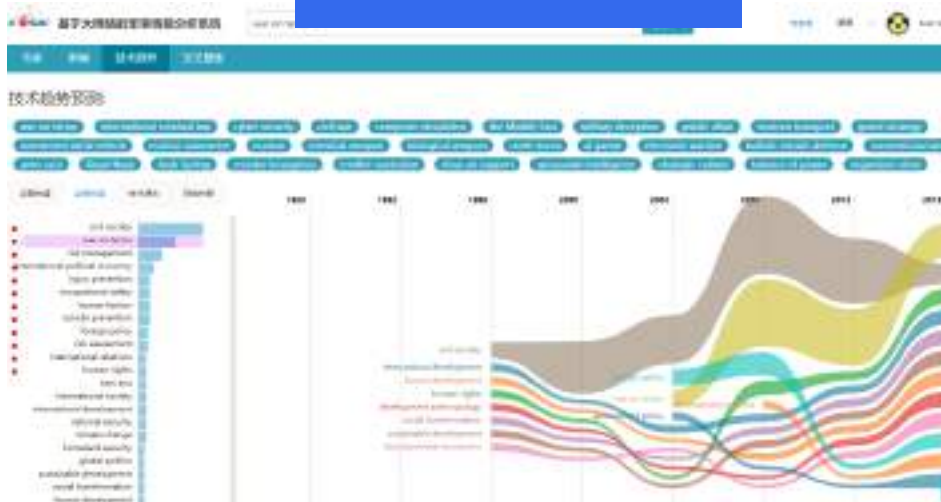


Zhi-Hua Zhou (周志华)

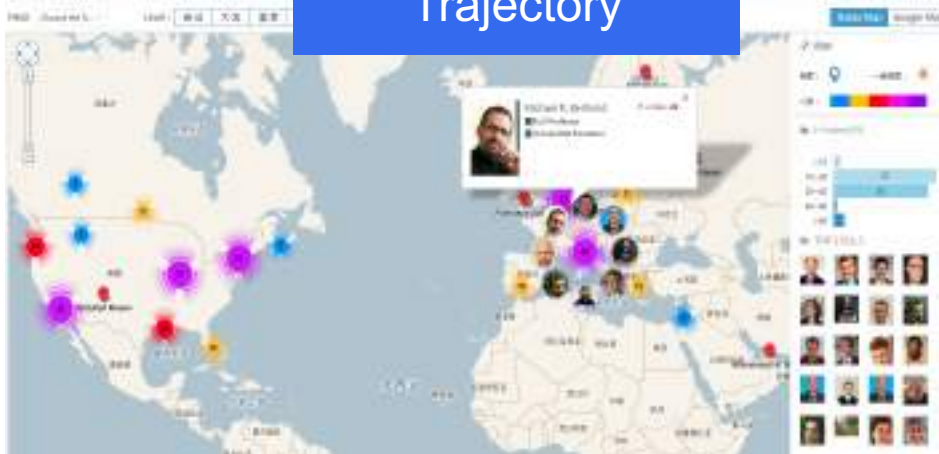
<https://aminer.org/datamining-experts/>

AMiner Services

Trend



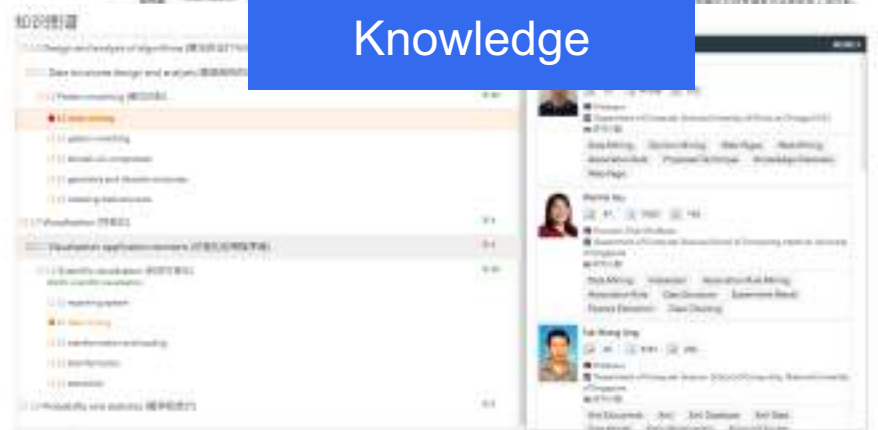
Trajectory



Expert Finding



Knowledge



Talent Map

学术资源地图 Data Mining

搜索

专家 地图

推荐国家: **ALL** 2013-2016 Fellow - ACM Fellow 美国国家科学院 - Research Fellows Directory 美国国家科学院 - Fellow Directory 中国科学院院士

通过条件: **ALL**

国籍: >=50 (64) 30-50 (43) 40-49 (75) 30-39 (137) 20-29 (222) 10-19 (257) <10 (96)

语言: Chinese (259) English (196) Greek (29) French (29) German (28) Japanese (26) Indian (24) Korean (14) Italian (11)

国家: USA (229) China (111) Taiwan (41) Australia (22) Canada (20) Japan (20) Italy (27) Singapore (22) Hong Kong (20) United Kingdom (19) Germany (18) India (18)

语言: H index 学术活跃度 被引用数 引用数 论文数

导出



韩国筠 (Jierwei Hong)

H index: 152, 430.96, 129.62, 139268, 933

Abel Bliss Professor
Department of Computer Science, University of Illinois at Urbana-Champaign
T: (217) 333-6900
jhong@illinois.edu



俞士伦 (Philip S. Yu)

H index: 144, 452.58, 123.46, 81671, 1872

Distinguished Professor & Wexler Chair in Information Technology
Department of Computer Science, University of Illinois at Chicago
T: (312) 996-0490
psyu@uic.edu



Mohammed J. Zaki

H index: 81, 22.81, 9.42, 25220, 280

Professor
Department of Computer Science, Rensselaer Polytechnic Institute
T: (518) 276-6360
zaki@cs.rpi.edu



Jon Pei

H index: 75, 61.60, 30.96, 36,955, 218

Professor
School of Computing Science, Simon Fraser University
T: 1-778-782-6851, 1-778-782-7197
jpei@cs.sfu.ca, jpei@cs.sfu.ca



Career Trajectory of Top Experts



AMiner

—AI-Powered Academic Search

<https://aminer.cn>

Thanks to our partners



AMiner

—AI-Powered Academic Search

<https://aminer.cn>





谢谢观看！



人工智能是学术搜索的未来吗？

AI Time 第十二期

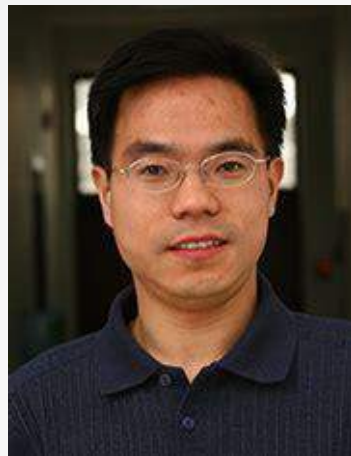
AI Time是一群关注人工智能发展，并有思想情怀的青年人创办的圈子。AI Time旨在发扬科学思辨精神，邀请各界人士对人工智能理论、算法、场景、应用的本质问题进行探索，加强思想碰撞，打造成为北京乃至全国知识分享的聚集地。



■ 嘉宾介绍



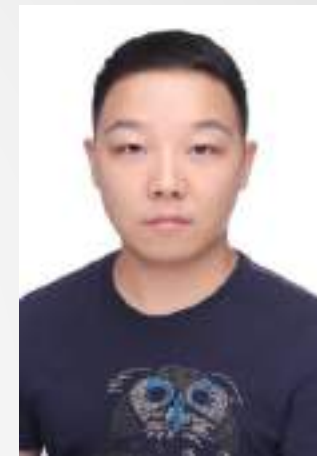
唐杰
清华大学计算机系教授
副系主任



周园春
中科院网络信息中心研究员,
大数据技术与应用发展部主任



刘筱敏
中科院文献情报中心研究馆员,
知识服务创业中心数据产品与
管理服务事业部主任



刘雪峰
百度学术搜索产品负责人

■ 主持人



李文珏
中科创星投资总监
CCF YOCSEF学术秘书福
布斯30Under30



杜一
中科院网络信息中心大数
据知识工程实验室主任
北京市科技新星



■ 目录

1. 学术搜索的内涵与外延
2. 思辨及讨论
 - a. 技术篇
 - b. 应用篇
 - c. 产业篇
3. 总结与畅想



■ 技术篇

- 学术搜索与普通搜索有什么区别，是更难还是更容易？
- 在学术搜索领域，我国现状是处在领跑、并跑还是跟跑？
- 关于学术搜索中的实体、知识融合，学术界已经有很多方法，但是在实践上并不能完全满足要求。如何应对这类问题？
- 在下列五个场景中：自动摘要生成、交叉学科发现、趋势分析与预测、学术影响力评价、专家&机构画像，还有哪些技术可以赋能？





■ 应用篇

- 从用户角度（例如学者、期刊/会议、高校、基金资助机构等），AI在哪些场景已经发挥了作用？哪些场景还需要改进？还有哪些需求没有被满足？
- 在学术搜索领域还有哪些场景可以用AI赋能？
- 学术搜索+AI对国内外学术环境有哪些影响？





■ 产业篇

- 目前谷歌、百度、微软等大型企业都持续在学术搜索上发力，作为Semantic Scholar、Aminer这样的学术搜索平台，在哪些方向会有所突破？
- 面对Aminer这样的平台的冲击，百度学术这样的传统学术搜索引擎，应该如何响应？
- 学术搜索对学术出版，科技情报产业的意义在哪里，能否突破产业局限？还有哪些商业上的想象空间？
- 自内而外 VS 自外而内的模式，这两种模式在应用场景、构建模式上有何不同？

■ 畅想未来

- 作为计算机、文献情报界以及产业界的代表，各位对学术搜索的未来怎么看？
- 为学术搜索领域从业者（科研人员、项目研发、产品设计等）提供一些建议？



■ It Is Your Turn



Q&A



谢谢大家



AI Time欢迎每一位有情怀的AI爱好者加入！我们需要您的思辨和碰撞！