

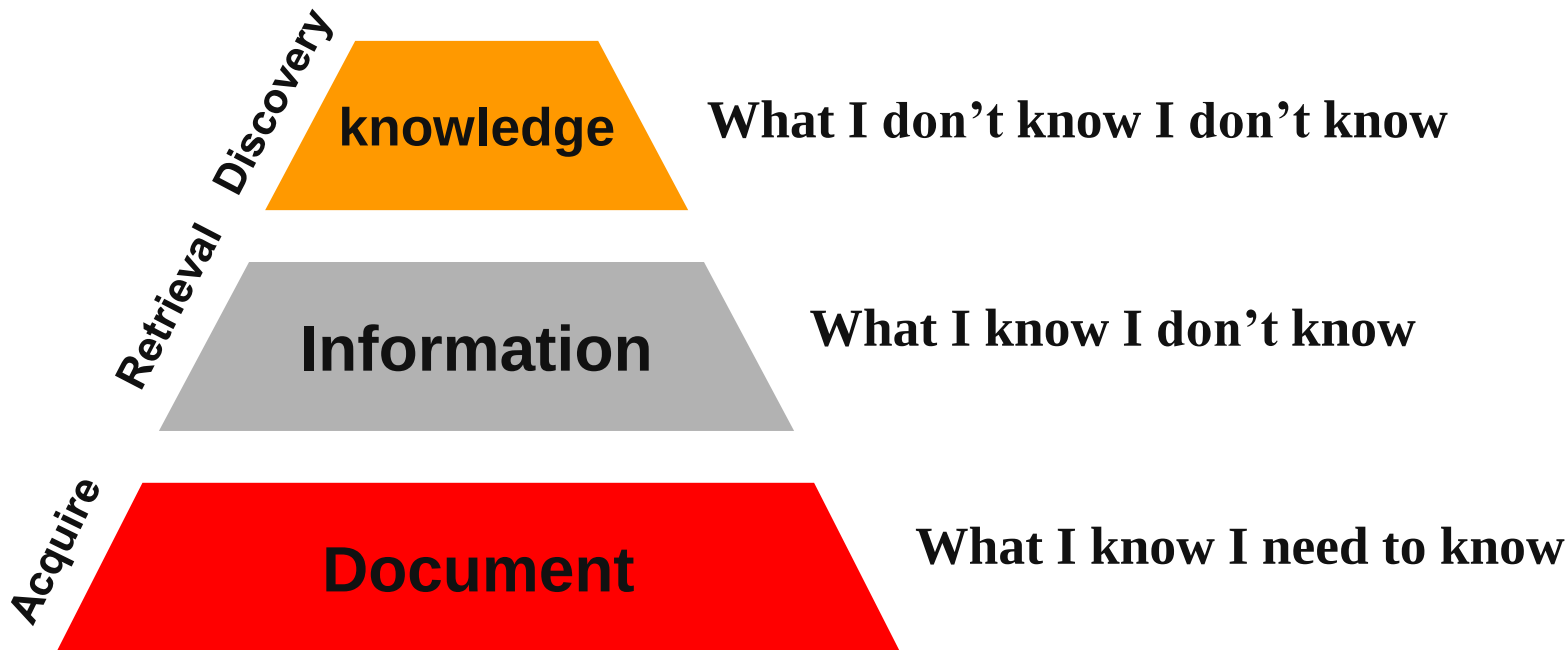


大数据时代学术图书馆数字资源建设与组织

——2018高校图书馆新任馆长高级研修班

孙 坦
中国农业科学院农业信息研究所

前言：学术图书馆信息资源建设本质与层次



第一节

学术图书馆的文献资源保障





馆藏体系VS保障体系

以藏为主，兼顾服务

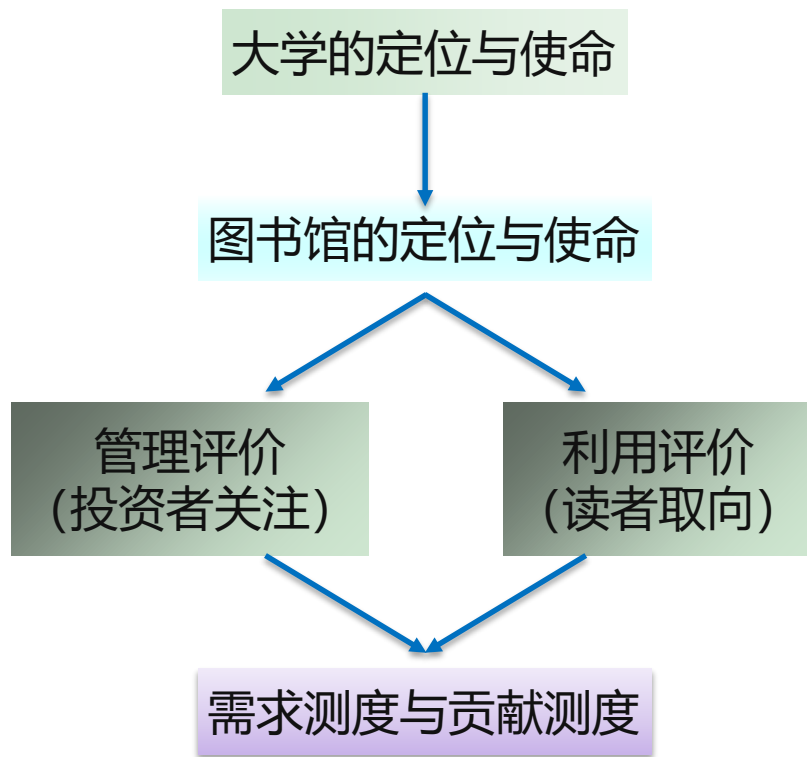
藏用并举，以用为主

不求所有，但求所用

以馆藏为中心VS以用户为中心

哪一个目标和原则是合适的，究竟要构建怎样的馆藏体系和资源保障体系？







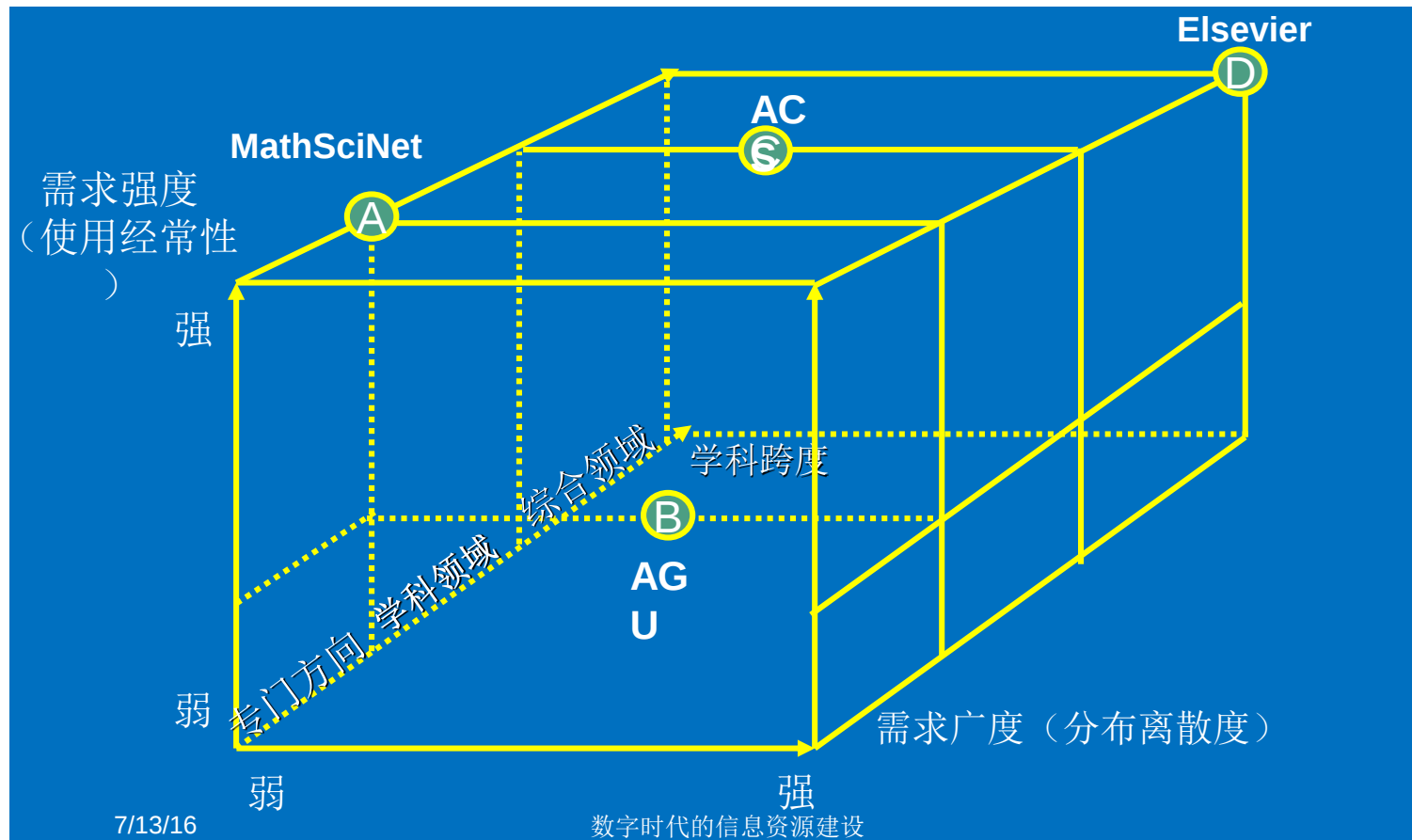
“读者满意”是图书馆保证生存的王道，“出乎意料”是图书馆提升地位的根本！

读者调查、流通统计、下载量分析是重要依据，
但更需要一套科学完整的需求测度和贡献测度体系

需求测度和贡献测度体系的建立不仅成为资源建设的依据，
更能成为资源建设创新的源泉！



学术图书馆文献资源保障





- 统筹分析用户需求各个角度（需求分析）

- 综合各个方面的用户需求信息

- 用户需求调查：通过问卷调查、读者推荐、互动访问、研究所图书馆推荐等方式获取的用户需求
 - 目标机构保障分析：通过对与我院研究领域相同的国内外同类的优秀科研机构的文献保障情况的统计分析获取的前沿用户需求
 - 用户引用行为统计分析：通过统计分析用户发表学术论文中的参考文献和引用行为而获取的用户需求
 - 用户使用统计：通过统计分析用户对本所和全院文献资源的借阅、复制、下载情况获取的用户需求



· 案例分析

主要学科方向：纳米科学的基础研究和应用研究

物质的研制

重点研究领域

纳米结构的系统和集成技术

纳米技术标准化和纳米标准

纳米结构的生物学效应和安全性研究

纳米制造的相关基础研究

纳米结构制备关键分析技术

学术图书馆文献资源保障



✓ 用户群体

科技人员89人

研究员18人

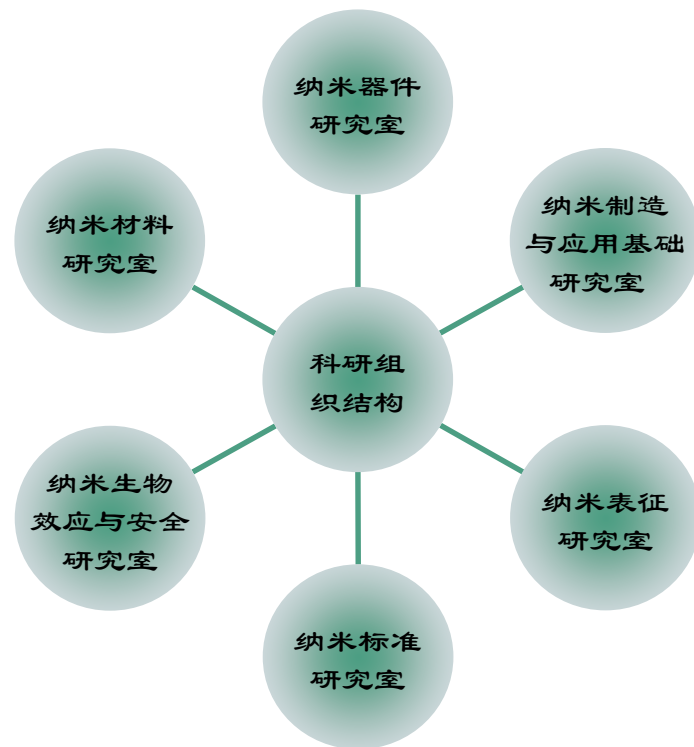
副研究员20人

高级工程师5人

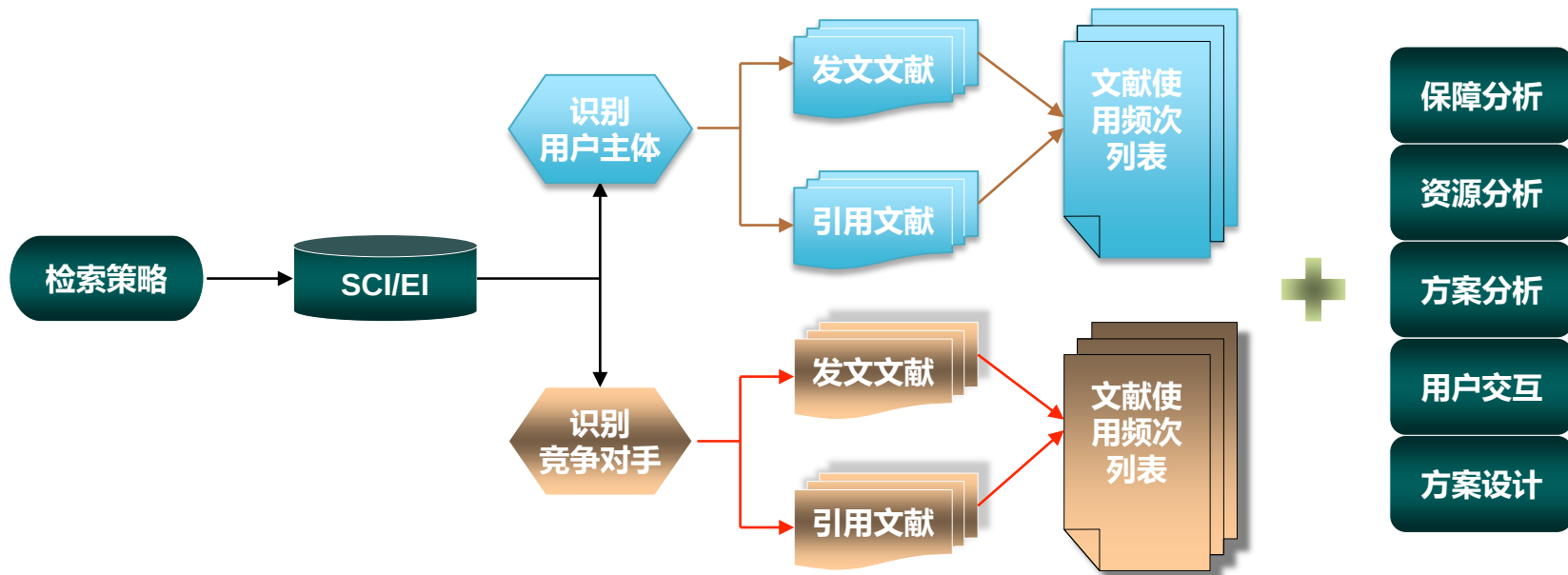
客座研究员和

访问学者24人

在读研究生117人



学术图书馆文献资源保障

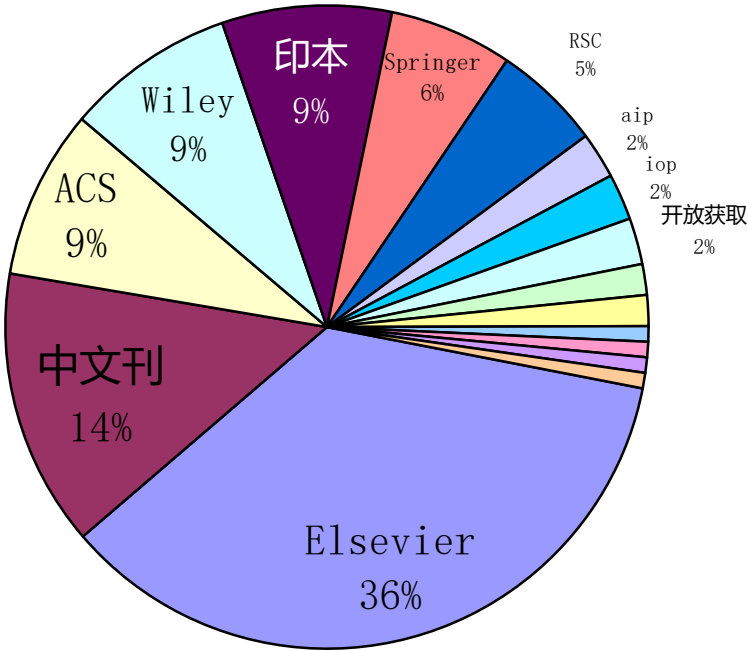




纳米中心发文期刊

- Elsevier
- 中文刊
- ACS
- Wiley
- 印本
- Springer
- RSC
- aip
- iop
- 开放获取
- aps
- Taylor & Francis
- Asme
- ieee
- Oxford
- Worldsci

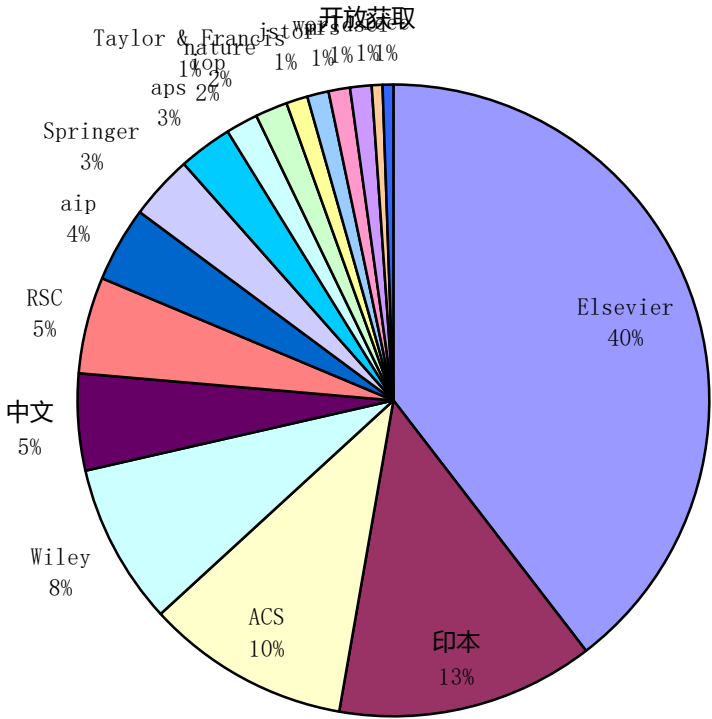
发文核心期刊（129种）分布状况（13个数据库）





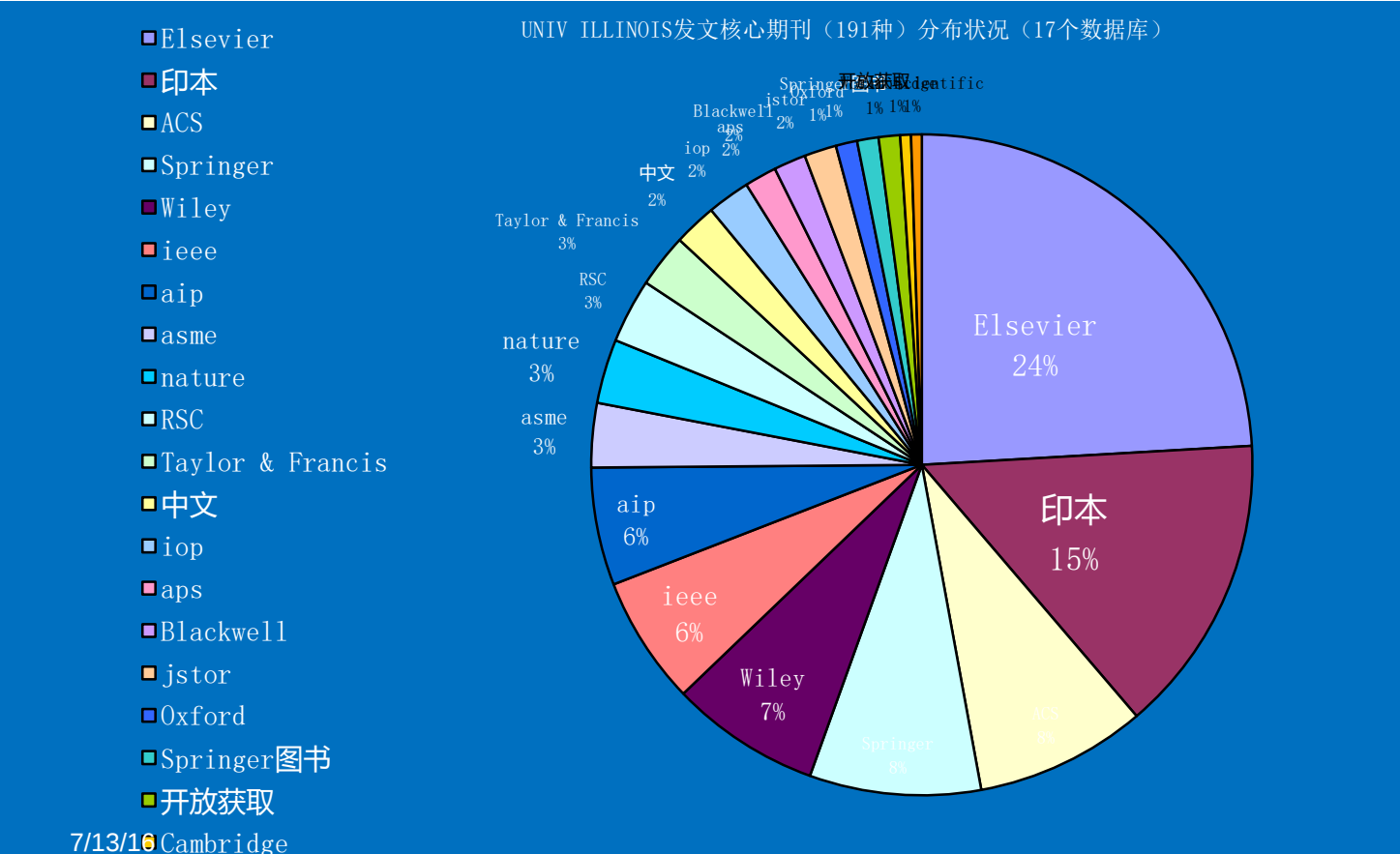
引文期刊（179种）数据库分布(14个)

- Elsevier
- 印本
- ACS
- Wiley
- 中文
- RSC
- aip
- Springer
- aps
- iop
- nature
- Taylor & Francis
- jstor
- mrs
- worldsci
- 开放获取
- ieee



纳米中心引用期刊

学术图书馆文献资源保障



学术图书馆文献资源保障



需求测度 + 流通统计 + 读者数据
下载统计 + COUNTER数据

贡献测度
优化分析
专门知识服务系统开发

根据量化分析，精细化设计馆藏体系与保障体系，综合运用整库、学科包、集团、区域、单篇、快速馆际互借与文献传递、PDA订购、短期租借、OA等方式

第二节

大数据时代的文献资源建设与组织



数字时代的信息困境：信息过载与信息缺失纠缠



开放信息环境中海量异构信息与碎片化需求的矛盾：

- 海量数据广泛分散在各种书籍、报纸、不同格式的音视频媒体以及互联网和其他媒介中，并快速的数字化。**这些数据**的指数增长导致人们利用计算机查找**专门数据和有用的相关信息的能力迅速降低**。
- 搜索引擎的算法主要依赖关键词匹配，而不考虑关键词的各种不同含义以及位于短语中组合关键词的复杂含义，甚至是关键词和短语被用于不同上下文中的不同含义。
- 多数知识呈现出要么非结构化要么是异构化的，致使其不能被进一步处理

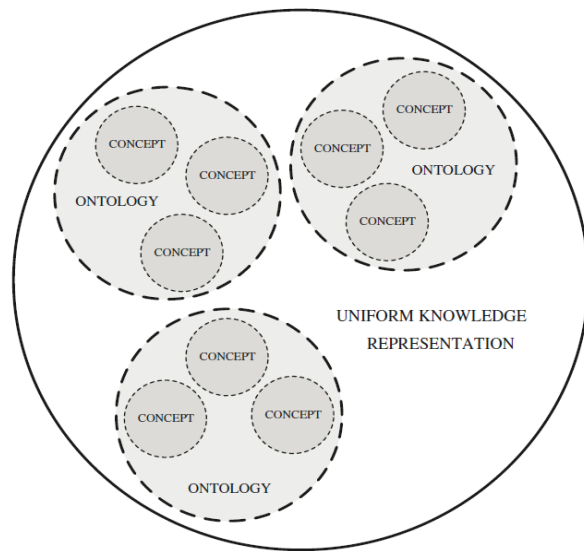
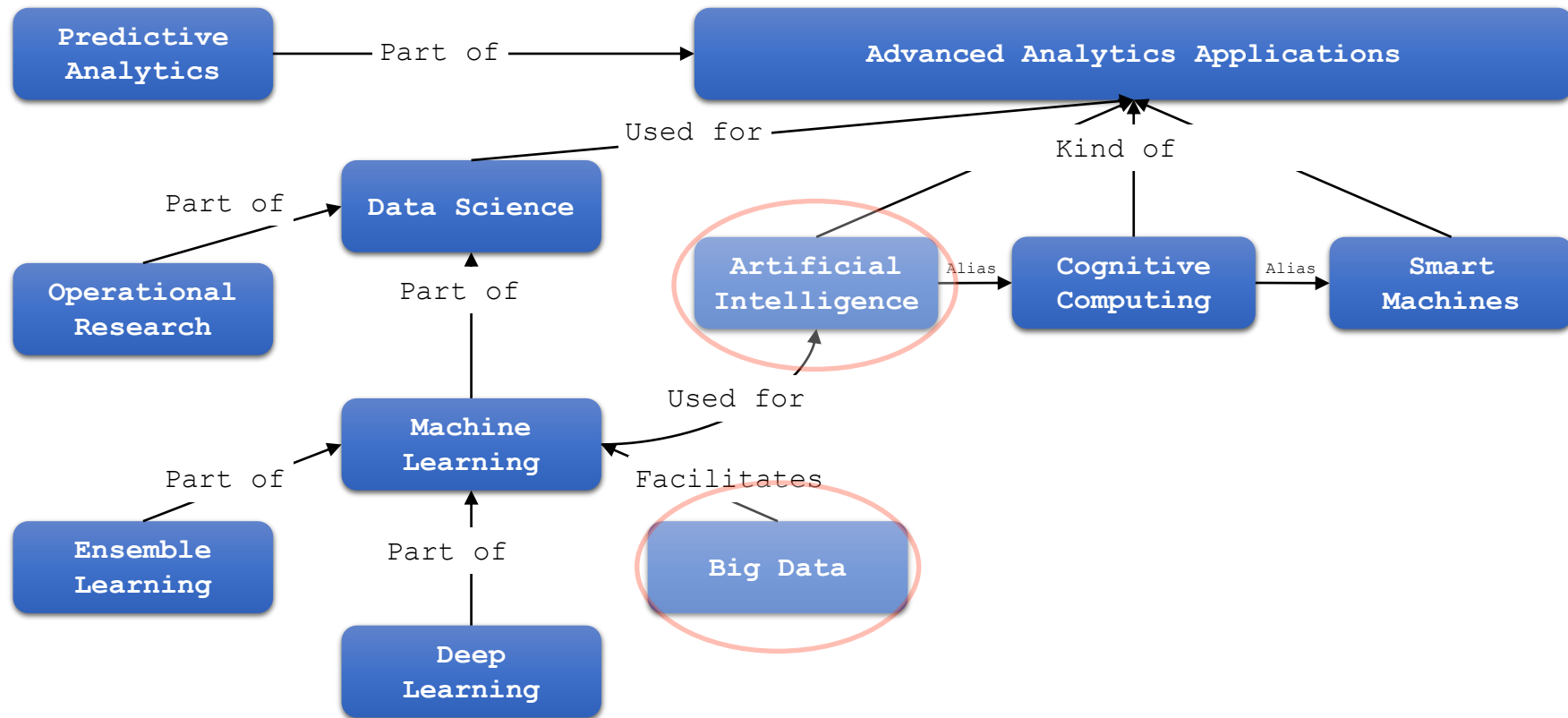


Fig. 1.1 Uniform knowledge representation consisting of ontologies populated by concepts

- **根本性问题：**如何从海量数据源中有效提取需要的数据？如何找到那些代表着问题答案或有助于解决问题的必要和潜在可用的却又未知的信息片段？如何集成那些潜在的记录（信息片段）？



数据科学全景图 (Data Science Ontology)



问题的提出



挖掘 计算学术发现

关联 探索学术发现 交叉性知识需求

精准 专业学术发现 系统性知识需求

泛在 大众学术发现 即时性知识需求

- 数据的指数增长
- 知识表示非结构化与异构化
- 搜索算法为词匹配, 非语境语义

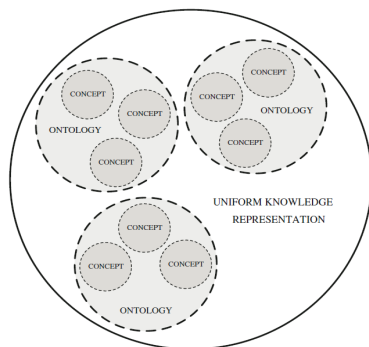
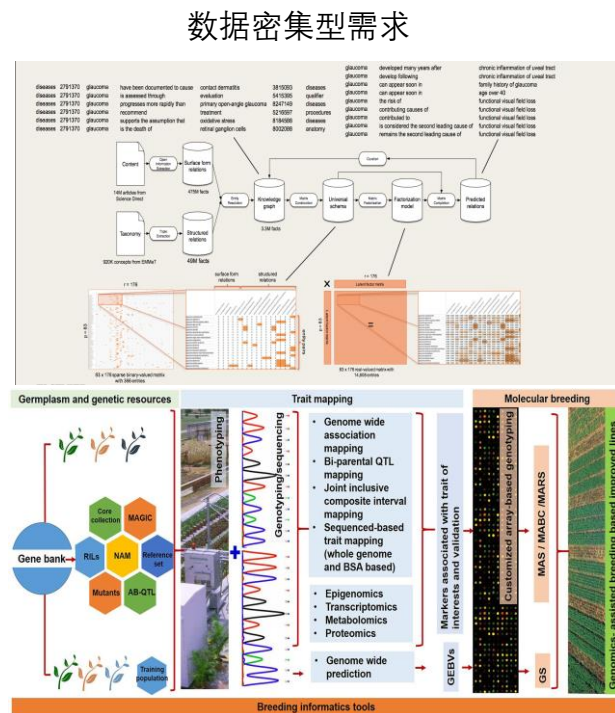
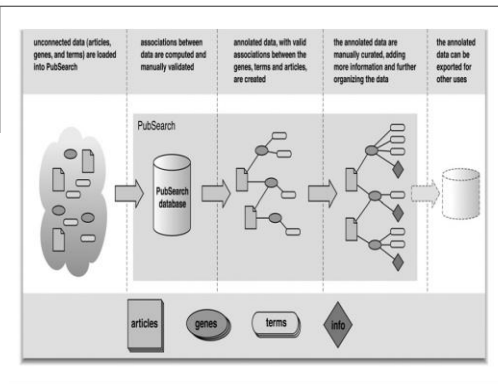
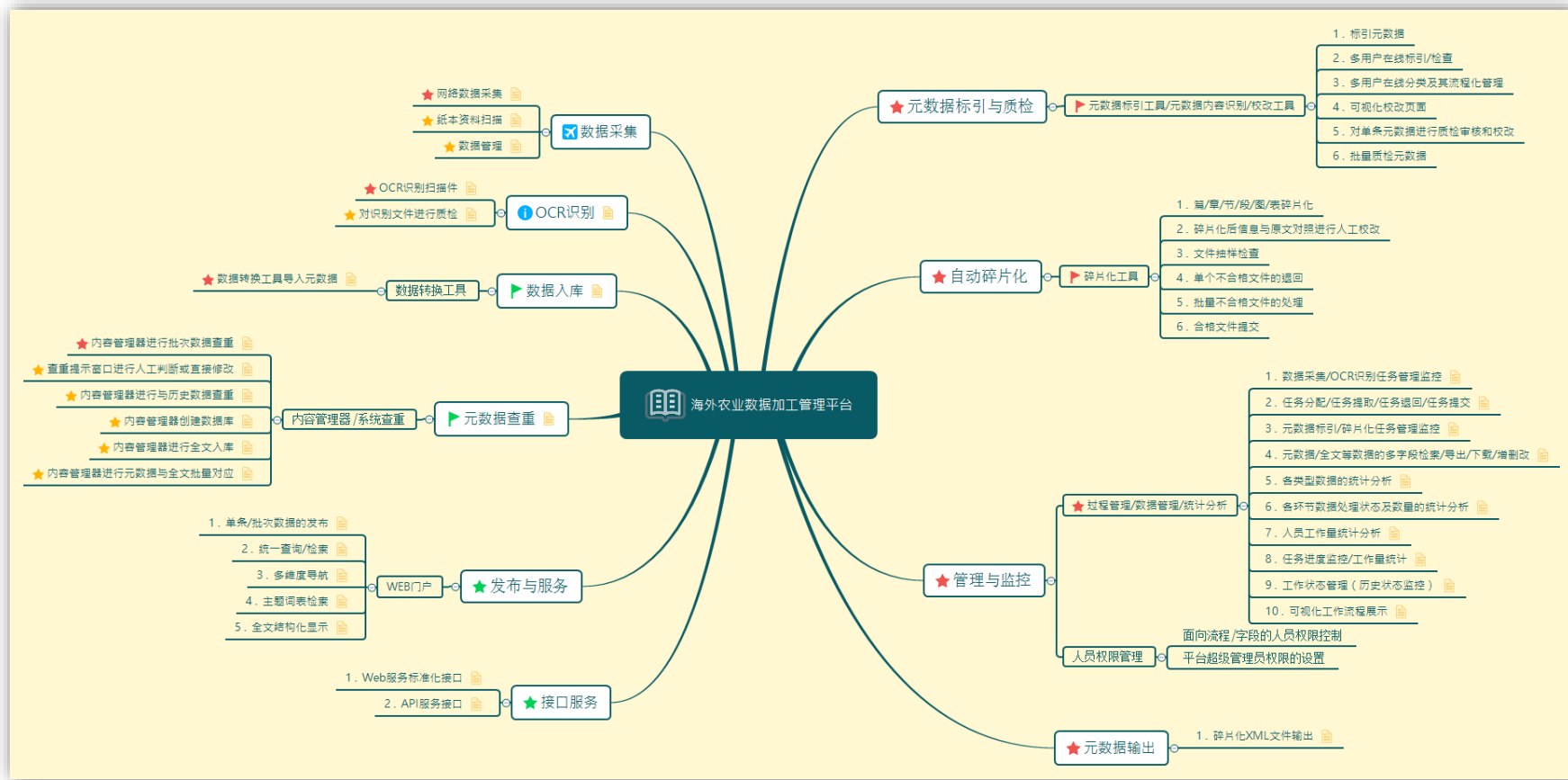


Fig. 1.1 Uniform knowledge representation consisting of ontologies populated by concepts



结构化、语义化知识表示



结构化、语义化知识表示



遴选分析

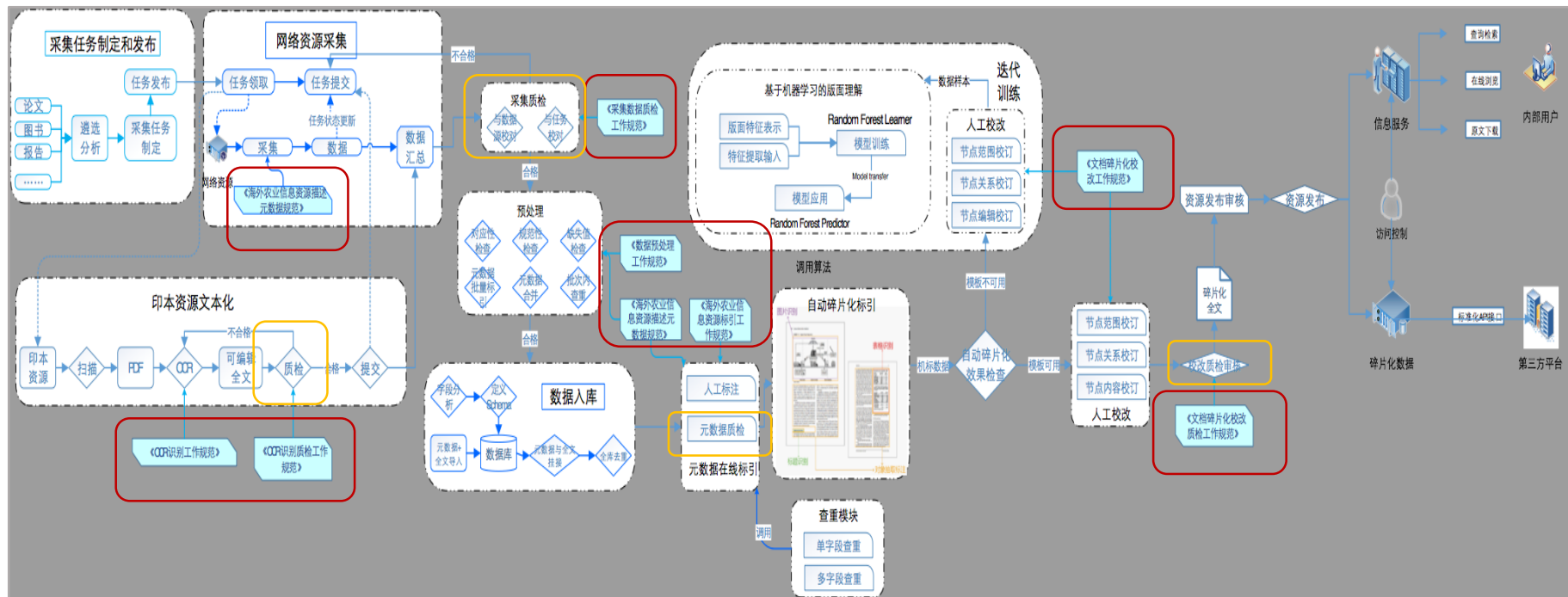
采集获取

数据预处理

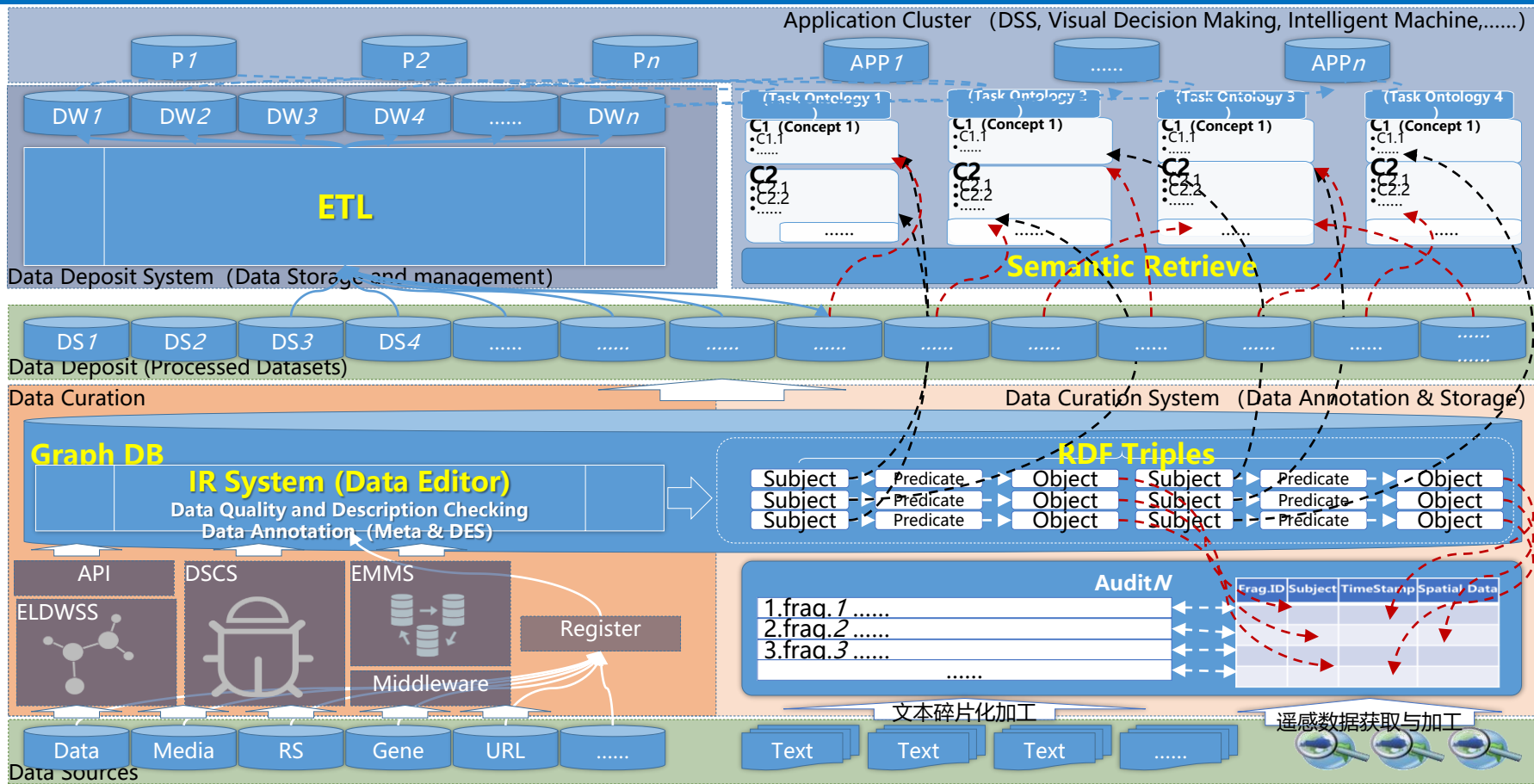
元数据标引

碎片化加工

审核发布



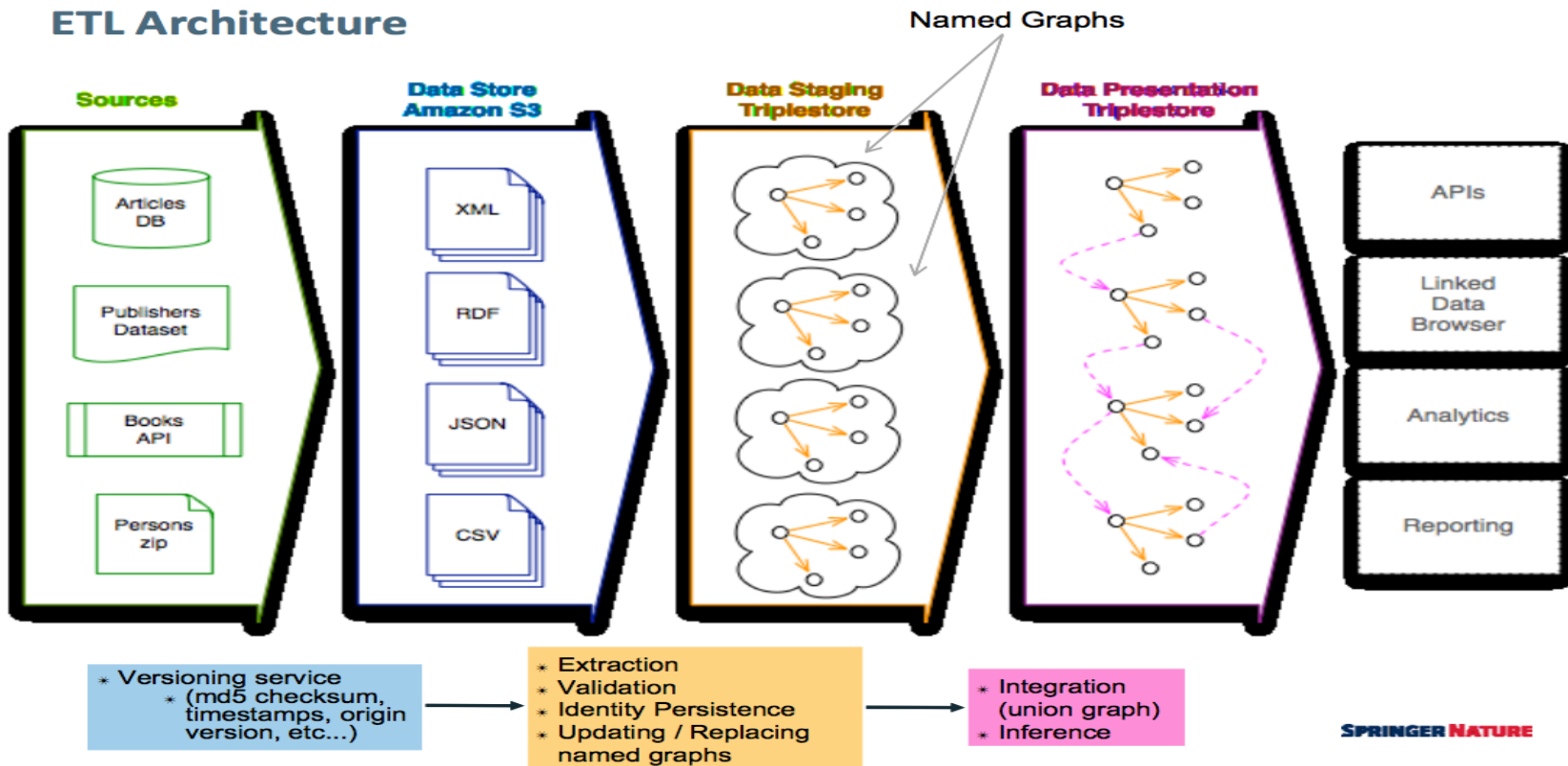
结构化、语义化知识表示



结构化、语义化知识表示



ETL Architecture



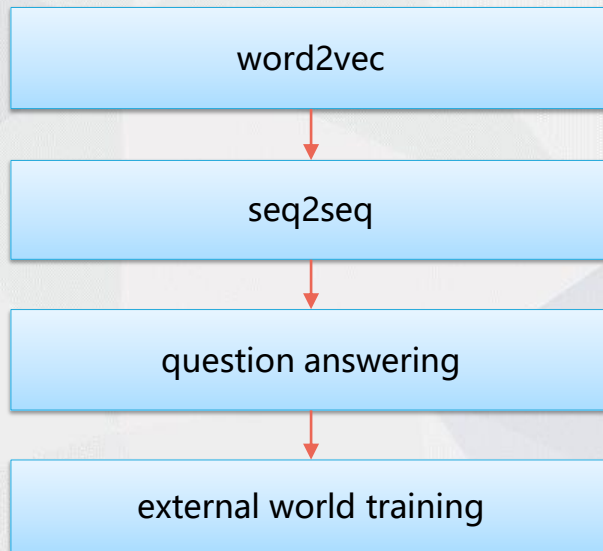
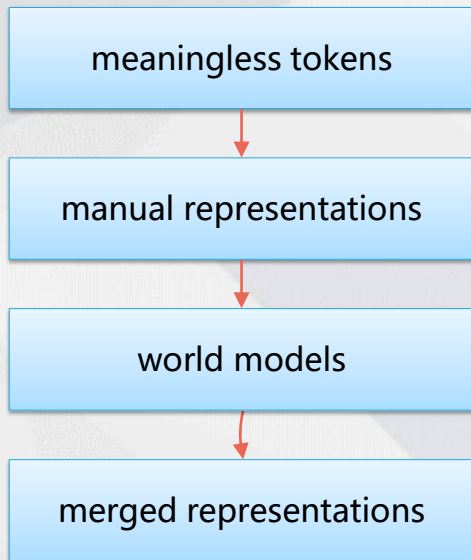
语义化的两种路径

Symbolic path

符号路径:

利用通过本体、元数据、标注、语义推理等技术，通过人为建立知识表达模型，在计算机内构建知识表达体系，通过语义推理实现符合意图的信息的定位与匹配。

在机器学习和人工智能第二次浪潮兴起之前颇为流行。



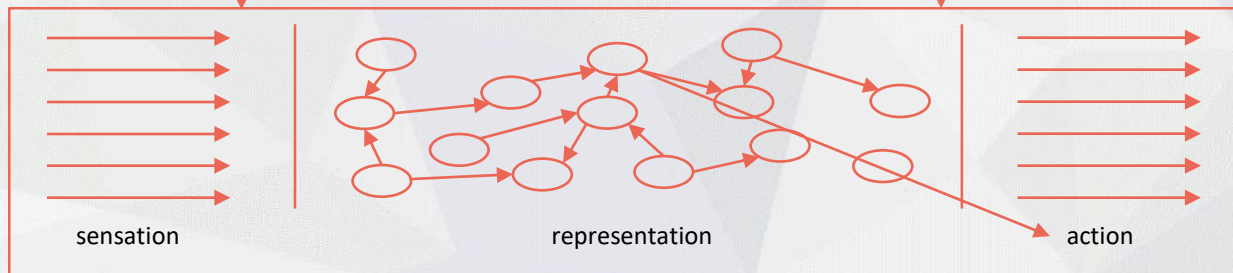
sub-symbolic path

亚符号路径:

利用自然语言处理、机器学习和深度学习等技术，以大规模并行计算为手段，以词、句向量为基础，实现全程无人工干预的信息处理和智能计算，其信息定位与匹配准确率达到或优于符号路径技术手段。

在机器学习、人工智能、深度学习技术兴起之后有成为主流的趋势。

(Google, Facebook、阿里...)

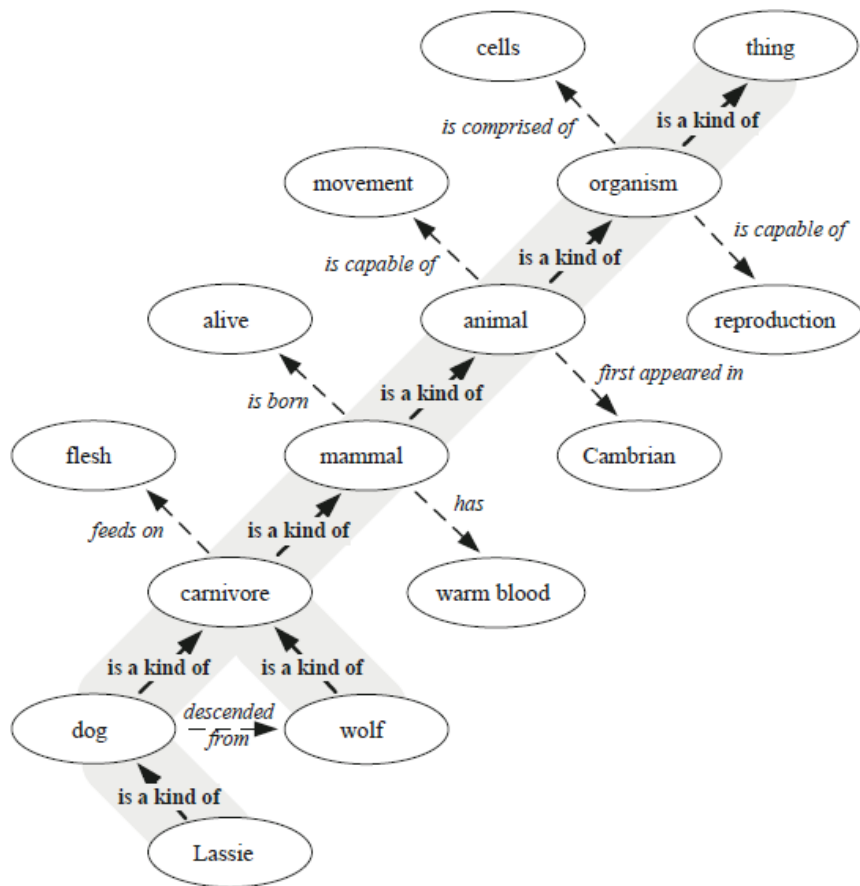


KOS vs Cognitive Computing





KOS模式的困难与挑战

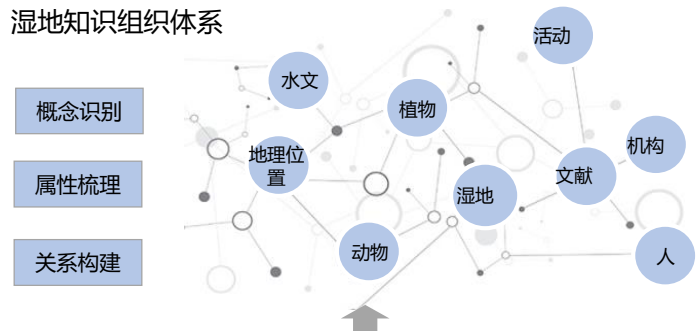


案例分析：湿地知识服务系统



知识组织体系构建

湿地知识组织体系



实体数据集

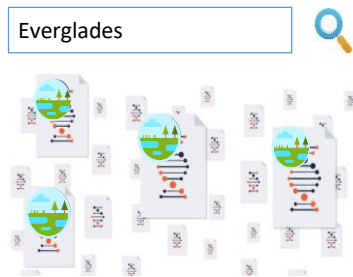


基础数据



支撑科研的知识服务

① 精准检索



② 文献实体识别与标注



③ 新知识关联发现

- 与绿鹭在同一湿地生存的植物、动物



④ 趋势分析与预测

- 领域研究热点及发展趋势



案例分析：湿地知识服务系统



Juncus

拼写建议

检索

研究分面

- ☐ 生物圈 68
- ☐ 水文 13
- ☐ 岩石圈 15
- ☐ 生态

结果筛选

地理位置

- ☐ New Brunswick 16
- ☐ the Florida Everglades 21
- ☐ 15

作者

- ☐ Montemayor, Marilou B.
- ☐ Price, Jonathan S.
- ☐ Rochefort, Line

文献

实体

Temporal variations and spatial patterns in saline and waterlogged peat fields: II. Ion accumulation in transplanted salt marsh graminoids

Montemayor, Marilou B.; Price, Jonathan S.; Rochefort, Line; Boudreau, Stephanie

【摘要】 Our earlier study in New Brunswick, Canada showed that *Spartina pectinata* Link survived very well after transplantation in a barren cutover bog to seawater, in all combinations of salinity and moisture content test

【概念】 [New Brunswick, Canada](#) (地理位置)
[Spartina pectinata](#) (植物) [Juncus balticus](#) (植物)

精检结果

Seed germination of two cattail (*Typha*) species as a function of Everglades nutrient levels

Stewart, H; Miao, SL; Colbert, M; Carraher, CE

【摘要】 Sawgrass (*Cladium jamaicense*) is being replaced by cattails (*Typha domingensis* and *Typha latifolia*) in Water Conservation Area 2A (WCA 2A) of the Florida Everglades. This replacement coincides with changes in nutrients in WCA 2A. Investigating seed germination of *Typha* in response to different nutrient levels, focusing on phosphate, might help in understanding how this replacement occurs.

【概念】 [the Florida Everglades](#) (地理位置)
[Juncus balticus](#) (植物) [Cladium jamaicense](#) (植物)

案例分析：湿地知识服务系统



Juncus

检索

研究分面

- | | |
|-------------------------------|----|
| ☐ 生物圈 | 68 |
| ☐ 水文 | 13 |
| ☐ 岩石圈 | 15 |
| ☐ 生态系统 | 6 |

地理位置

- | | |
|---|----|
| ☐ New Brunswick | 16 |
| ☐ the Florida Everglades | 21 |
| ☐ | 15 |

作者

- | |
|---|
| ☐ Montemayor, Marilou B. |
| ☐ Price, Jonathan S. |
| ☐ Rochefort, Line |

文献

Common Names:

- bog rush (English, United States)
 - Brako, L., A.Y. Rossman & D.F. Farr. 1995. Sci. Comm. Names 1-294.
- deng xin cao shu
 - Flora of China Editorial Committee. 1988-2013. Flora of China (Checklist & Addendum). Unpaginated. In C. Y. Wu, P. H. Raven & D. Y. Hong (eds.) Fl. China. Science Press & Missouri Botanical Garden Press, Beijing & St. Louis.
- deng xin cao shu (Pinyin, China)
 - Flora of China Editorial Committee. 2000. Flora of China (Flagellariaceae through Marantaceae). 24: 1-431. In C. Y. Wu, P. H. Raven & D. Y. Hong (eds.) Fl. China. Science Press & Missouri Botanical Garden Press, Beijing & St. Louis.

Description

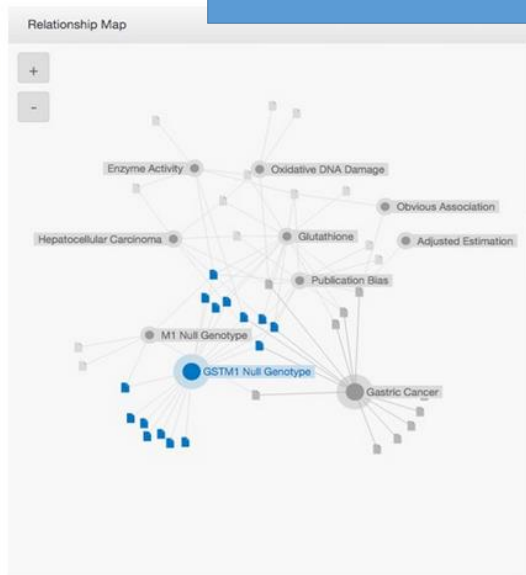
Glabrous herbs, generally occurring in wet places. Leaves flat or cylindric, sometimes septate. Inflorescence compound, paniculate, bracteate; bracteoles sometimes present immediately below and clasping the flower. Perianth segments free, ovate or lanceolate, entire, persistent.



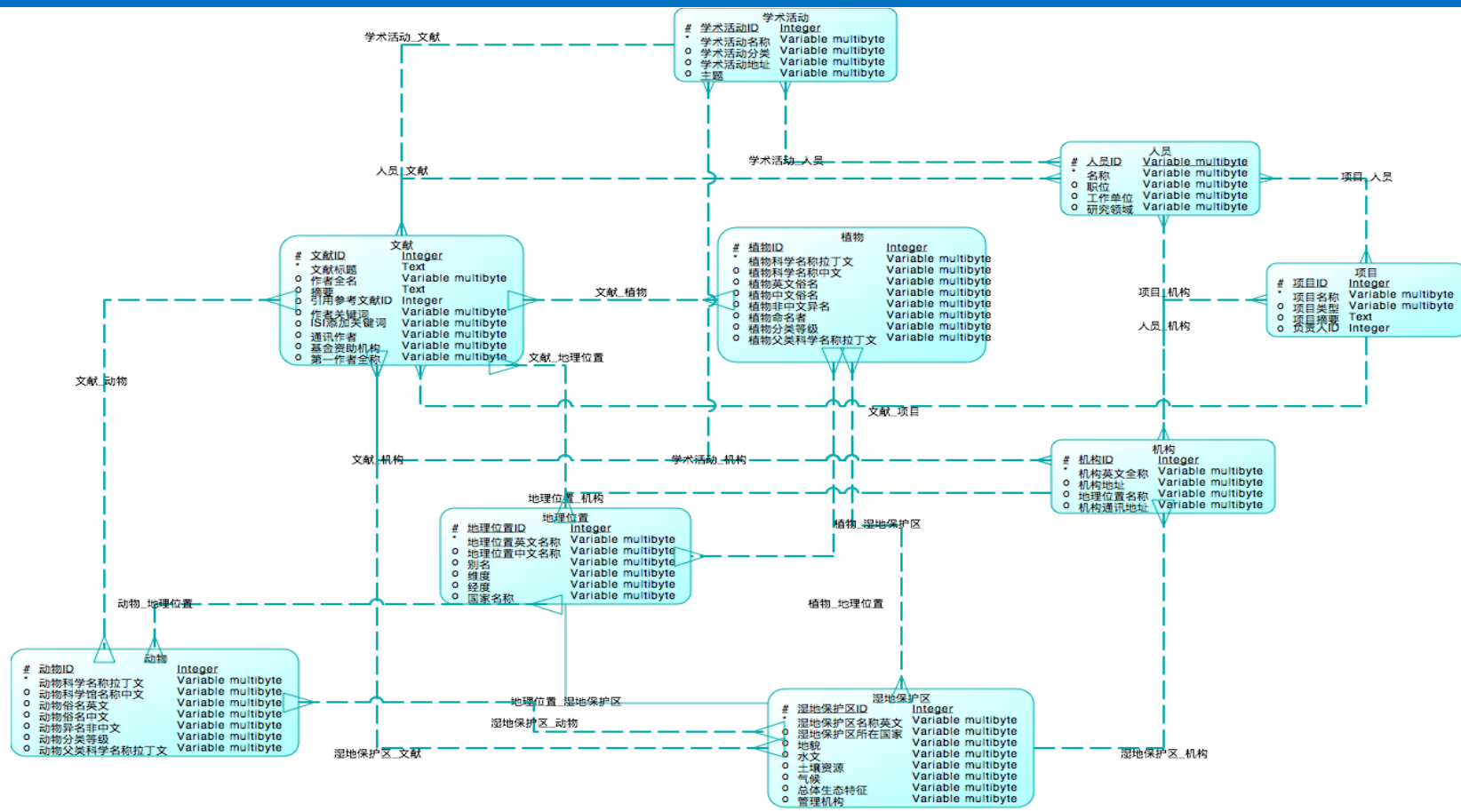
实体属性

实体

实体关系图



湿地知识服务系统语义模型



案例分析：湿地知识服务系统



以湿地的一种类型：关键词“**mire、mires**”，时间“**1900-2014**”，在WOS上检索相关湿地论文，经过机器和人工判读，获得论文**2451**篇文献，提取作者关键词**4709**个（去重）。

- 分析DDC类目表中与湿地相关的类目，从水文学和水生态学等下的**51**个类目获得术语**2642**个；随后进一步扩展，获取环境保护、环境监测等下的DDC类目，共**152**个类目，术语**6910**个。
- 与DDC提取的湿地叙词表（包括概念优选和非优选术语）比对，匹配上了**419**个，仅占**8.9%**。
- 与STKOS所有术语（包括概念优选术语和非优选术语）数据比对，匹配**2151**个术语，比对上的STKOS术语占作者关键词总数的**45.68%**。（备注：**1990**年之前文章没有作者关键词和文摘）。

案例分析：湿地知识服务系统



实验任务：基于本体模型标注的2万余篇湿地文献集合，检索标题中出现“湿地”、“盐碱”、“芦苇”的文献，命中1200篇，但经过科研用户阅读确认只有400篇是正确的。

Teaching Machines to Read and Comprehend
Karl Moritz Hermann (Tom 'a's Ko'chik 'y 'i)
Edward Grefenstette (Lasse Espoholt ' WII
Kay ' Mustafa Suleyman (Phil Blumson)
Google DeepMind (University of Oxford
[arXiv:1306.0153v1, 2013, 1306.0153v1, 1306.0153v1, 1306.0153v1]
why, most of them, phi blumson]
Abstract
Teaching machines to read natural language documents remains an elusive challenge.
Machine reading systems can be tested on their ability to answer questions posed on the contents of documents that they have seen, but until now large scale training and test datasets have been missing for this type of evaluation.

Sets A: 1200

Teaching Machines to Read and Comprehend
Karl Moritz Hermann (Tom 'a's Ko'chik 'y 'i)
Edward Grefenstette (Lasse Espoholt ' WII
Kay ' Mustafa Suleyman (Phil Blumson)
Google DeepMind (University of Oxford
[arXiv:1306.0153v1, 2013, 1306.0153v1, 1306.0153v1, 1306.0153v1]
why, most of them, phi blumson]
Abstract
Teaching machines to read natural language documents remains an elusive challenge.
Machine reading systems can be tested on their ability to answer questions posed on the contents of documents that they have seen, but until now large scale training and test datasets have been missing for this type of evaluation.

Sets B: 400

问题?

Teaching Machines to Read and Comprehend
Karl Moritz Hermann (Tom 'a's Ko'chik 'y 'i)
Edward Grefenstette (Lasse Espoholt ' WII
Kay ' Mustafa Suleyman (Phil Blumson)
Google DeepMind (University of Oxford
[arXiv:1306.0153v1, 2013, 1306.0153v1, 1306.0153v1, 1306.0153v1]
why, most of them, phi blumson]
Abstract
Teaching machines to read natural language documents remains an elusive challenge.
Machine reading systems can be tested on their ability to answer questions posed on the contents of documents that they have seen, but until now large scale training and test datasets have been missing for this type of evaluation.

Sets C: 800

类和属性有缺失

按Tbox进行完善

词表与Sets A、B、C
分别共现聚类

语义关系有缺失

按Abox进行完善

分别对Sets A、B、C
进行基于词表的抽取及关系识别

基于认知计算的路径



现代知识发现与服务中的科学问题



- **知识获取过程忽略用户意图生成和凝练，导致知识获取与人类认知行为脱节，知识获取过程发散，需用户对结果集进行大量人工筛选和加工；**
- **缺乏基于内容发现的工具，无法得到与知识获取意图直接相关的各类语料信息，噪音过多，对知识获取与决策构成干扰和障碍；**
- **缺乏对知识载体进行有效宏观和微观分析的工具，导致隐藏在只是载体中的大量有用知识难以发现，无法直接服务于决策；**
- **结果呈现缺乏互动性和良好的可视化技术引导，获取知识过程鼓噪，易疲劳，效率低。**

需求分析-以植绥螨性别决定相关研究为例



- **目标：植绥螨生殖决定可能相关的基因及其功能研究，背景研究**

- 其他物种中已经明确有相关功能的基因及其研究历程；Keywords：sex determination; gene；难点：如何缩小第一步搜索范围；

- **关注对象**

- 节肢动物（Arthropoda）、昆虫纲（Insecta）、蛛形纲（Arachnoidea）、蜱螨目（Acari）
- 难点：分类单元较大，文章摘要中未必会出现，出现更多的是更具体的物种名称，检索工作量大难以穷举，如何缩小关注对象的范围；

- **所需知识**

- 哪些基因可能是比较重要的
- 那些物种上相关研究比较透彻
- 相关基因研究发展历程如何

需求分析



- **知识搜寻意图形成**

- 无法明确说出自己想要找什么
- 能够说出大致方向，但不知具体有哪些需要查找的方向（大多数人）
- 不能明确给出完整的检索所涉及到的知识体系（遗漏、片面或偏离）
- 需要不断地搜索、完善，迭代多次完善搜索意图
- 需要在脑子里面构建一个相对完整、精确的知识框架
- 信息自动抽取与迭代筛选

- **知识搜索**

- 需要找出所有满足要求的知识载体（文献、文章、片段、图像、数据、.....），且不仅局限于关键字匹配层面
- 需要在检索结果基础上进一步实现一些互动分析功能，如研究热点分析、核心观点抽取、核心知识抽取、给定关键关系及其实体抽取、机构与学者分析、基于共现和共引的合作关系分析、核心研究者分析、相关资源导航、.....



- **帮助用户完善并优化知识获取意图**

- 新涉足一个领域不能完整、精确地描述知识获取意图；

- **精确获取各种载体形式的知识**

- 通过与用户的交互形成能够精确反映用户检索意图的，知识库能够识别的检索式，提交给各类知识库，获取所需知识（文本无需全文，摘要亦可支撑分析）；

- **探索式分析帮助用户快速获取所需知识**

- 宏观
 - 跨时空范围、学科级别、主题级别、演化
- 微观
 - 文档、段落、句子甚至术语级别

项目目标



- 基于NLP、NLU和深度学习技术，研发面向科研人员的一站式语料搜索与知识发现工具；
- 系统定位
 - 帮助科研人员以尽可能少人工干预的方式通过完成从信息采集、意图精炼、问题完善、语料获取、深度洞察到探索式分析的全线Online学科知识发现、处理与分析的工具。
- 结果展示（互动式分析）
 - 热词分析及相应语料摘要定位
 - 实体关系展现
 - 关键事件及关系展示
 - 研究热点时间轴分析
 - 研究者集群及机构集群分析
 - 交叉学科分析
 - 知识点与文献互动分析
 - 研究趋势预测
 - 学科关键关系挖掘与展示
 -

相关技术

意图形成

🔗 主题分布挖掘、词嵌入向量、短语、聚类、相似度计算

语料获取

🔗 意图树→检索式；检索式构建(文献服务部合作)、文档自动分类；

结果展示

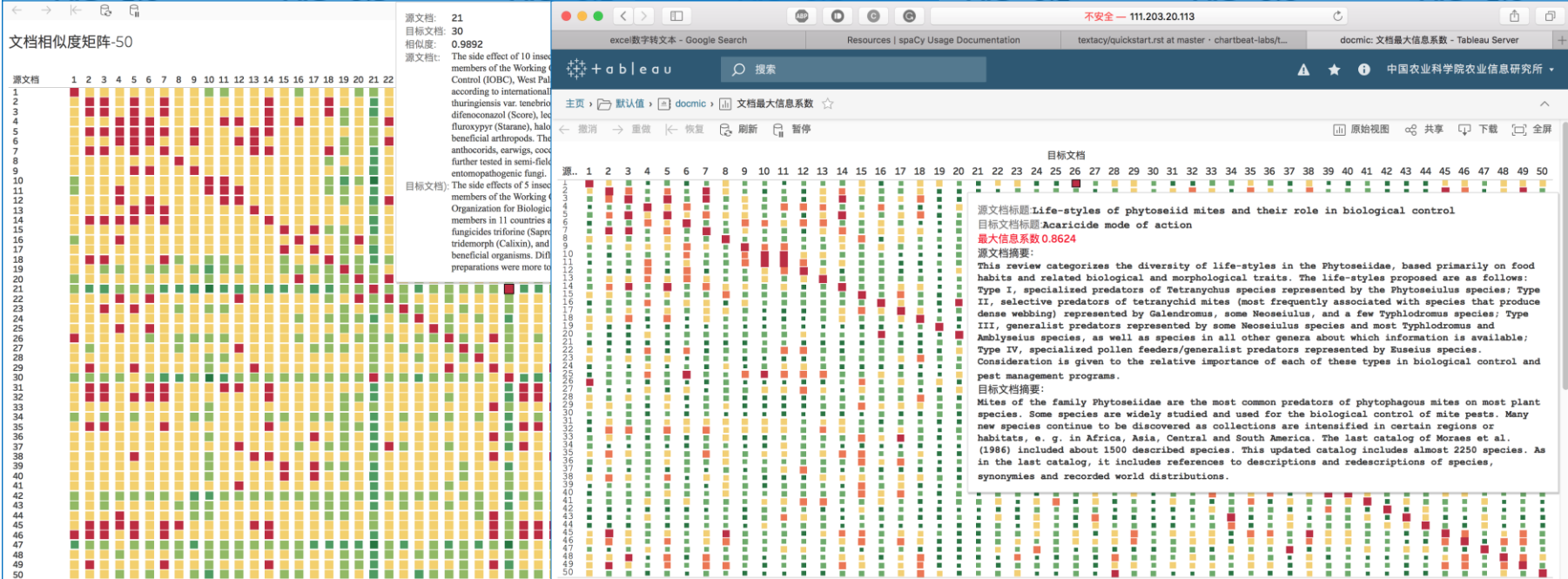
🔗 文档/段落相似度计算、文献计量可视化、标签云、语法依赖解析、NER（专业领域模型训练，主动学习工具）与可视化展现，自动摘要、社交网络分析、事实抽取、主题Hyper Cycle、通用知识图谱

已开展工作



与婧姮蝶安例佳团队 (婧保所与我所文献服务中心) 建立

Pearson $r=1.0$ MIC=1.0 Pearson $r=0.8$ MIC=0.5 Pearson $r=0.4$ MIC=0.2 Pearson $r=0.0$ MIC=0.1 Pearson $r=-0.4$ MIC=0.2 Pearson $r=-0.8$ MIC=0.6 Pearson $r=-1.0$ MIC=1.0

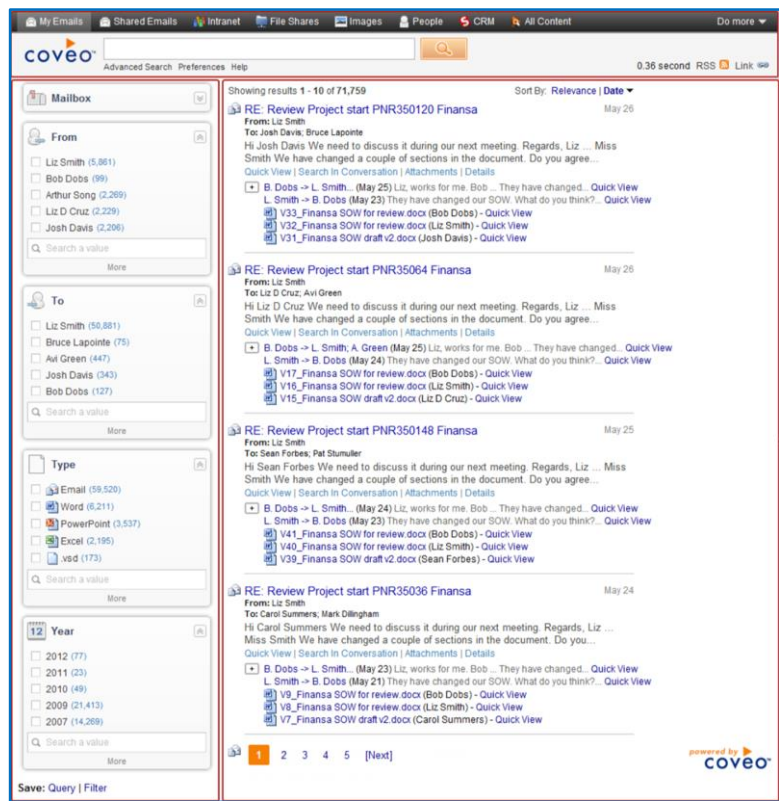


词嵌入向量在语义文献分析领域的应用



• 文献检索词推荐及动态分面抽取

- 对全部（或分领域）文献摘要计算计算词向量
- 根据用户检索词进行相似度检索推荐相关词
- 通过基于词向量的词聚类推荐可能的分面内容与用户交互参与检索



水稻文献词向量计算结果



Enter word or sentence (EXIT to break): rice rust

Word: rice Position in vocabulary: 15

Word: rust Position in vocabulary: 10275

Word	Cosine distance
akte	0.609442
yield_losses	0.604716
fungus_disease	0.585722
varieties	0.582189
rusts	0.575174
ipuccinia	0.569658
wheat	0.561889
puccinia	0.561224
bacterial_leaf	0.559551
infests	0.558015
sorgho	0.556355
crops	0.555031
smut	0.554524
foliar_blight	0.554129
crop	0.547200
cultivars	0.545741
rust_puccinia	0.543391
plant	0.536509
disease	0.536113
rot_caused	0.531086
maydis	0.528316
commonly_grown	0.518207
tan_spot	0.518179
serious_pest	0.517019
major_insect_pest	0.512582
leaf_rust	0.511574
stem_rust	0.511564
stem_rot	0.511301
southern_united_states	0.511249
pyricularia_grisea_cooke_sacc	0.510720
yellow_rust	0.510668
maize	0.509211

rice-classes.sorted.txt

rust

词嵌入向量在文献语义检索方面的应用



$$V_{doc} = \frac{1}{n} \sum_{i=0}^n V_{w_i} (w_i \in doc)$$

$$\cos(\theta) = \frac{\sum_{i=1}^n (x_i \times y_i)}{\sqrt{\sum_{i=1}^n (x_i)^2} \times \sqrt{\sum_{i=1}^n (y_i)^2}}$$

- 结合词向量训练文档向量/段落向量/句向量用于对文档进行相似度计算，可用于选定一篇文献（摘要）后过滤其相似度高的文献并按照相似度排序。

文献①题名: Evaluation of the GeneSequenceReg. DNA hybridization method in conjunction with 24-hour enrichment protocols for detection of Salmonella spp. in select foods: collaborative study.

文献②题名: Validation of a Modification to Performance-Tested Method(SM) 010403: Microwell DNA Hybridization Assay for Detection of Listeria spp. in Selected Foods and Selected Environmental Surfaces

进行多实验室研究以比较GeneSequenceReg DNA杂交DNAH方法的性能。所述DNAH方法包含新的24小时富集方案和参考培养程序以用于检测选定食物中的沙门氏菌属物种六种食物类型生鸡肉干全蛋牛奶巧克力核桃和干宠物食品通过DNAH方法和美国农业部 - 食品安全和检验局USDA-FSIS或美国食品和药物管理局的细菌学分析手册FDA/BAM的15种实验室的培养方法进行验证参与研究四通过卡方分析确定的DNAH方法和参考程序之间在性能上没有统计学显著性差异。DNAH方法的灵敏度范围为92%至100%DNAH方法的灵敏度范围为92%至100%方法与特定浓缩协议评估发现b e对核桃沙门氏菌检测无效对于一项试验的生牛肉结果显示性能有统计学显著性差异，参考方法获得的阳性结果更多但是有证据表明，阳性数量的差异可能是由于缺乏同质性的测试样品而不是DNAH方法的性能

2 对性能测试方法SM 010403 GeneSequence R李斯特菌测试DNAH方法的修改描述修改后的方法使用新的培养基配方LESS富集培养液用于食品和环境卫生拭子样品的单步浓缩方案食品样品富集27-30小时30摄氏度。环境样品30摄氏度24-48小时实施这些缩短浓缩程序可以在第二天获得测试结果在用接种样品进行的内部比较研究中测试14种食物类型时，DNAH方法和参考培养方法在27小时富集时间点上只有2种食品巴氏杀菌蟹肉和生菜的统计学显著性差异，以及在30小时富集时间点的一次试验中仅有单一食品巴氏杀菌蟹肉肉上使用3种食物进行实验室检测显示所有食物和结果的方法之间的统计学等同性它支持内部试验的结果整体考虑内部和独立的实验室试验DNAH方法相对于参考培养程序的灵敏度为90%测试5个环境表面接种各种李斯特菌菌株的结果表明DNAH方法更多生产效率高与参考美国农业部 - 食品安全和检验局USDA-FSIS培养程序适用于3面不锈钢塑料和铸铁。而结果与其他2面瓷砖和密封混凝土的参考方法在统计上相当独立实验室用接种单核细胞增生李斯特菌的菌株进行试验证实了在24小时时间点DNAH方法的有效性相对于USDA-FSIS方法24小时的DNAH方法的总体灵敏度为152%DNAH方法表现出极高的特异性整体上只有1%的假阳性反应

A multilaboratory study was conducted to compare performance of the GeneSequenceReg DNA hybridization DNAH method incorporating new 24 h enrichment protocols and reference culture procedures for detection of Salmonella spp in select foods Six food types raw ground turkey raw ground beef dried whole egg milk chocolate walnuts and dry pet food were tested by the DNAH method and by the culture methods of either the US Department of Agriculture-Food Safety and Inspection Service USDA-FSIS or the US Food and Drug Administration's Bacteriological Analytical Manual FDA/BAM Fifteen laboratories participated in the study Four of the foods tested raw ground turkey dried whole egg milk chocolate and dry pet food showed no statistically significant differences in performance between the DNAH method and the reference procedure as determined by Chi square analysis Sensitivity rates for the DNAH method ranged from 92 to 100% The DNAH method with the specific enrichment protocol evaluated was found to be ineffective for detection of Salmonella spp in walnuts For raw ground beef results from one trial showed a statistically significant difference in performance with more positives obtained by the reference method However evidence suggests that the difference in the number of positives was likely due to lack of homogeneity of the test samples rather than to DNAH method performance

2

3

4 A modification to Performance-Tested MethodSM 010403 GeneSequence R Listeria Test DNAH method is described The modified method uses a new media formulation LESS enrichment broth in single-step enrichment protocols for both foods and environmental sponge and swab samples Food samples are enriched for 27-30 h at 30 degrees C and environmental samples for 24-48 h at 30 degrees C Implementation of these abbreviated enrichment procedures allows test results to be obtained on a next-day basis In testing of 14 food types in internal comparative studies with inoculated samples there were statistically significant differences in method performance between the DNAH method and reference culture procedures for only 2 foods pasteurized crab meat and lettuce at the 27 h enrichment time point and for only a single food pasteurized crab meat in one trial at the 30 h enrichment time point Independent laboratory testing with 3 foods showed statistical equivalence between the methods for all foods and results support the findings of the internal trials Overall considering both internal and independent laboratory trials sensitivity of the DNAH method relative to the reference culture procedures was 90% Results of testing 5 environmental surfaces inoculated with various strains of Listeria spp showed that the DNAH method was more productive than the reference US Department of Agriculture-Food Safety and Inspection Service USDA-FSIS culture procedure for 3 surfaces stainless steel plastic and cast iron whereas results were statistically equivalent to the reference method for the other 2 surfaces ceramic tile and sealed concrete An independent laboratory trial with ceramic tile inoculated with L monocytogenes confirmed the effectiveness of the DNAH method at the 24 h time point Overall sensitivity of the DNAH method at 24 h relative to that of the USDA-FSIS method was 152% The DNAH method exhibited extremely high specificity with only 1% false-positive reactions overall

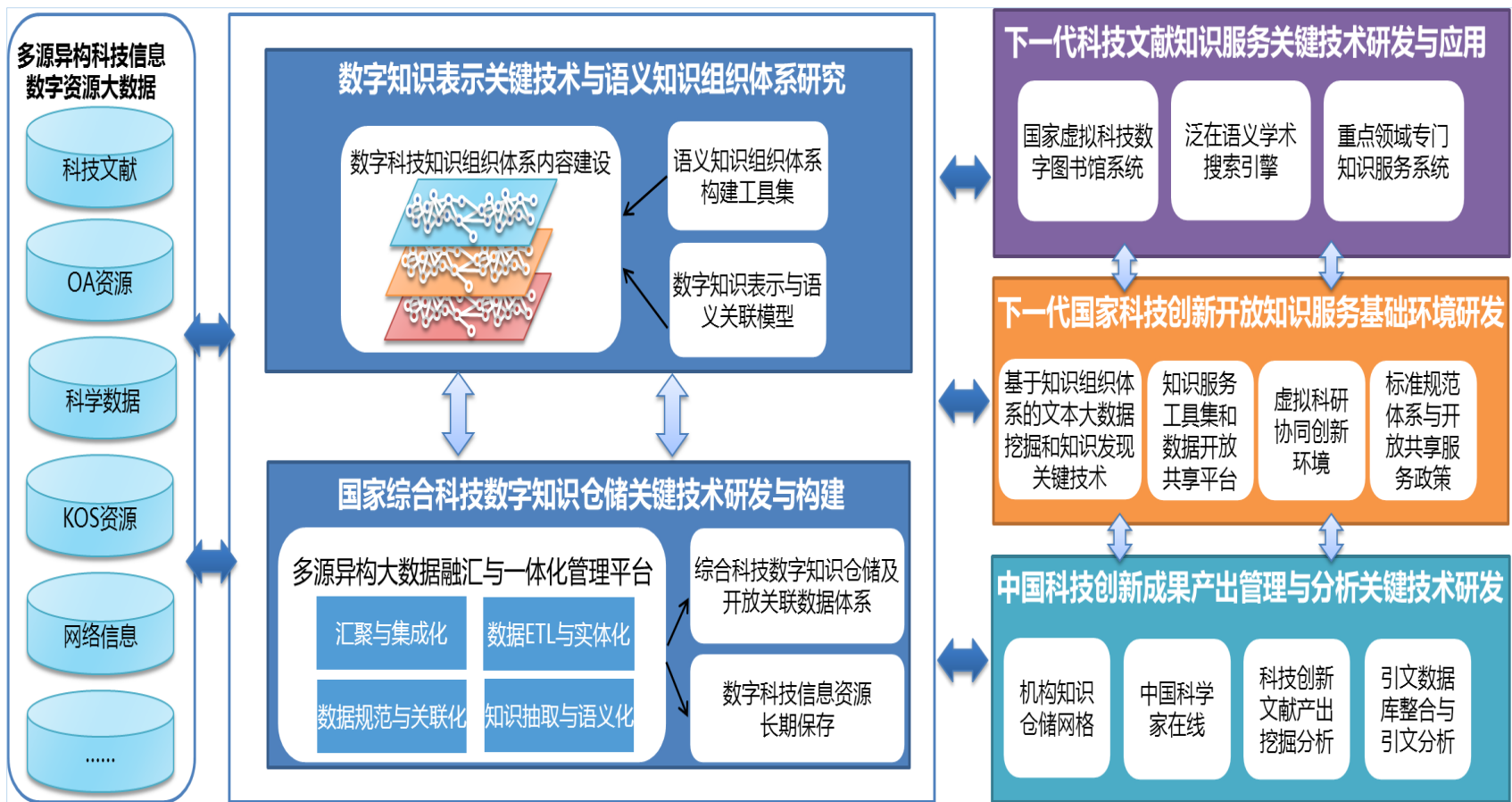
文摘①与文摘②相似度 = 0.9800212642599903

存在的困难与挑战

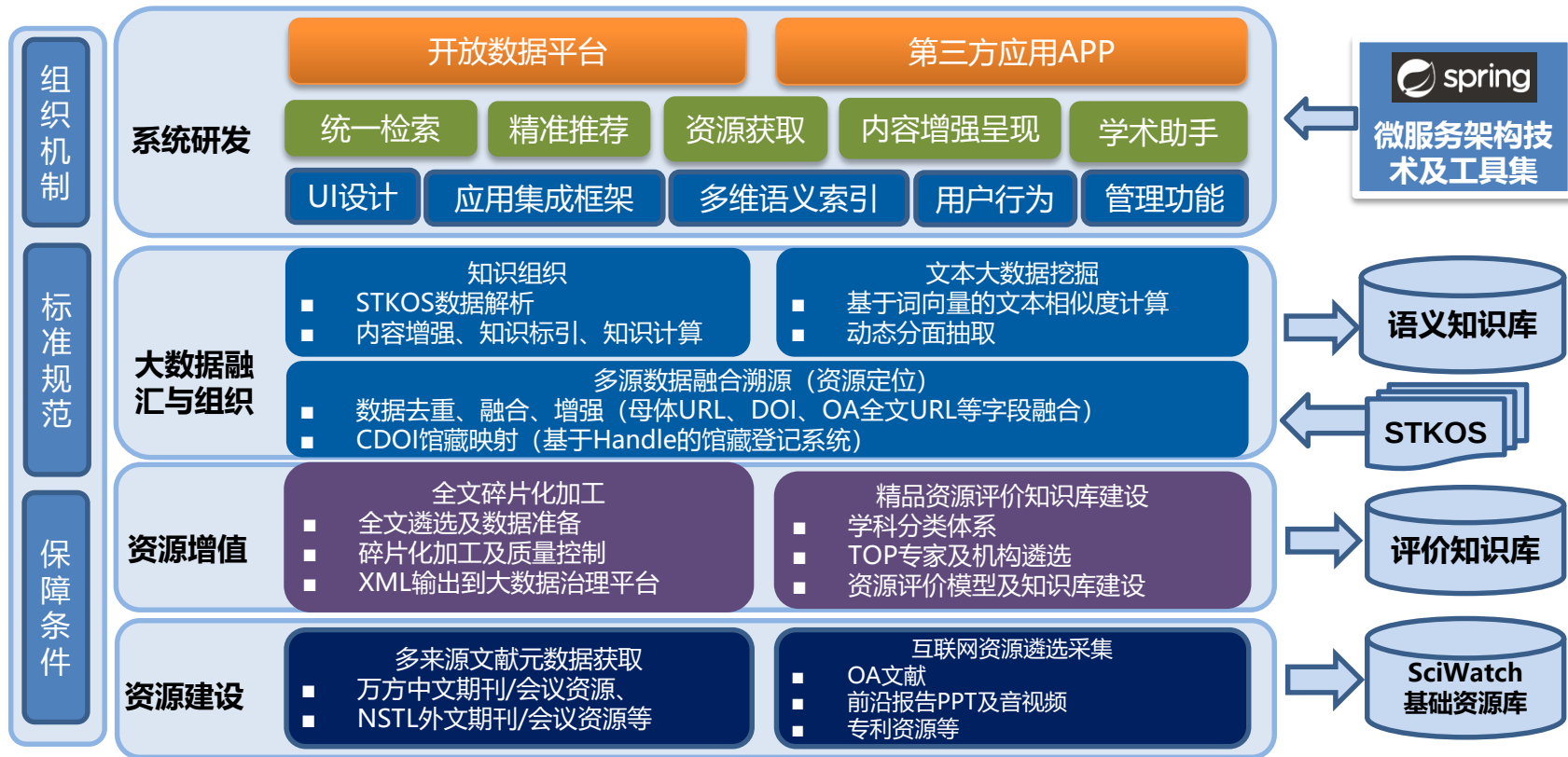


- 词向量计算与语料选择紧密相关，是整体计算所有语料还是只取一部分参与计算？
- 全集语料大小关系着计算时间的长短；
- 精确的词向量计算时空复杂度较高，计算量很大，即使目前的GPU网络也无法承受太大量语料的计算；
- 文档相似度的计算需尝试不同的方法（侧重语法和语义的词向量计算）并评估其结果；
- 文档相似度对于大量文档计算较慢，且目前无GPU支持，无法实时出结果，需先做一定筛选降低计算文档的数量，或者事先计算好把结果存起来。

融合两种路径的设想与尝试



语义文献知识发现系统——SciWatch

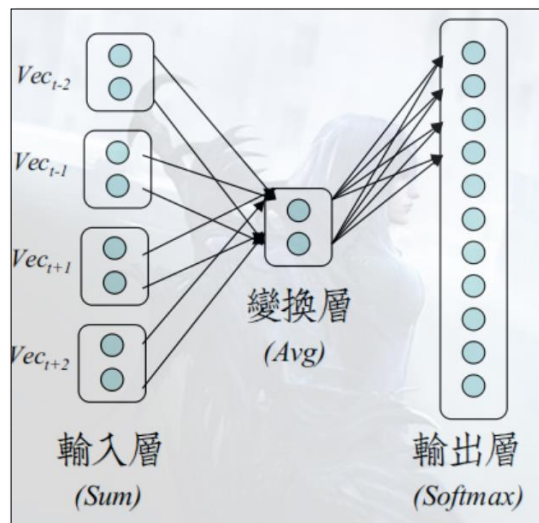


云基础设施

思路1：词嵌入向量



- 利用词向量训练方法，把相关元数据关联，使用词嵌入向量加以训练，解决作者、机构标注不规范问题；
- 训练内容：所有能起到互相标识作用的元素，如个人名称、机构名称、题名、基金项目名称、会议名称、学科分类、.....
- 需从底层改写词向量计算算法。



基于上下文窗口的词嵌入

	I	Love	NLP	And	Like	Dogs
I	0	1	0	1	1	0
Love	1	0	1	0	0	0
NLP	0	1	0	1	0	0
And	1	0	1	0	0	0
Like	1	0	0	0	0	1
Dogs	0	0	0	0	1	0

基于词共生矩阵的词嵌入

我来到达观数据参观

“<ap”, “app”, “ppl”, “ple”, “le>”

相应的bigram特征为：我来 不到 到达 达观 达观数 观数据 数据参 据参观

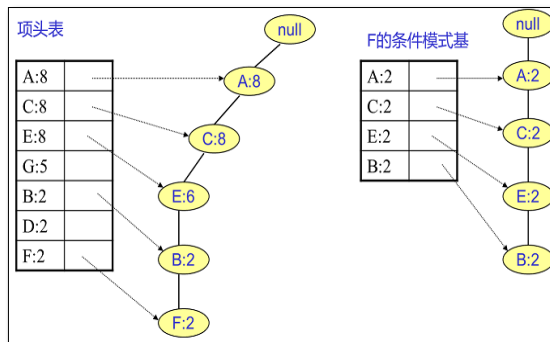
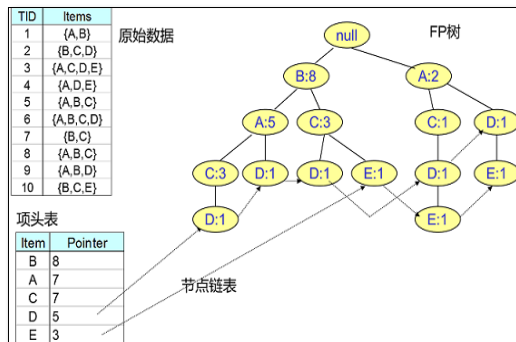
相应的trigram特征为：我来到 来到达 到达观 达观数 观数据 数据参 据参观

基于n-gram和子词的词嵌入

思路2：关联规则发现



- 假设的提出：在一起出现最多的项集（最频繁项集）代表规范参照项。
- 规范库的情况与啤酒纸尿裤的例子有相似之处
- 基于频繁项集的数据挖掘算法
 - 传统Association rules发现算法效率低，时空复杂度高，中间计算结果数量巨大，无法适用于大规模数据集计算
 - 新的关联规则算法：基于树的算法，时空复杂度低，用于大规模数据集效果好
 - 通过自然语言处理对人名、机构名等进行编码，转换为可计算的数值型数据



F的频繁2项集：
{A:2,F:2}, {C:2,F:2}, {E:2,F:2},
{B:2,F:2}

频繁3项集：
{A:2,C:2,F:2}, {A:2,E:2,F:2},...

.....

频繁5项集：
{A:2,C:2,E:2,B:2,F:2}



- **选择大量不同领域、不同年度文献元数据，用不同模型进行试验，评估、改进模型算法**
- **可尝试思路1 + 思路2的结合，关联发现用于元数据层面，词嵌入向量用于知识层面，互相参考验证**
 - 关联规则发现用于发现规范参照项，不同时间规范参照项的变化反应名称变化
 - 词向量方法用于发现相似项，与规范参照项进行映射
- **与现有技术路径实现效果进行比对，验证思路可行性**
- **工程化开发与应用方面还需解决一些问题**

数字图书馆未来还会存在吗？



- **科学共同体的门槛和边界不会随着技术创新而消失**
- **出版机构的语义知识服务和ICT界的AI/MI/QA等系统也很难成为科学共同体数据密集型科学发现的主体应用系统**
- **专门知识服务系统将仍然主要由政府公益机构开发运营，会呈现鲜明的学科化、专题化色彩**
- **NLM-Mesh-NCBI-Entrez-PubMed.....更代表着未来学术图书馆在数字图书馆领域的职能和定位方向**
- **基于微服务架构的数据共享、工具众创和服务协同将成为学术图书馆协同创新、联合服务的必然路径**

感谢聆听！



致谢：

崔运鹏、刘峥、赵瑞雪、寇远涛、鲜国健、黄永文、颜蕴、刘敏娟