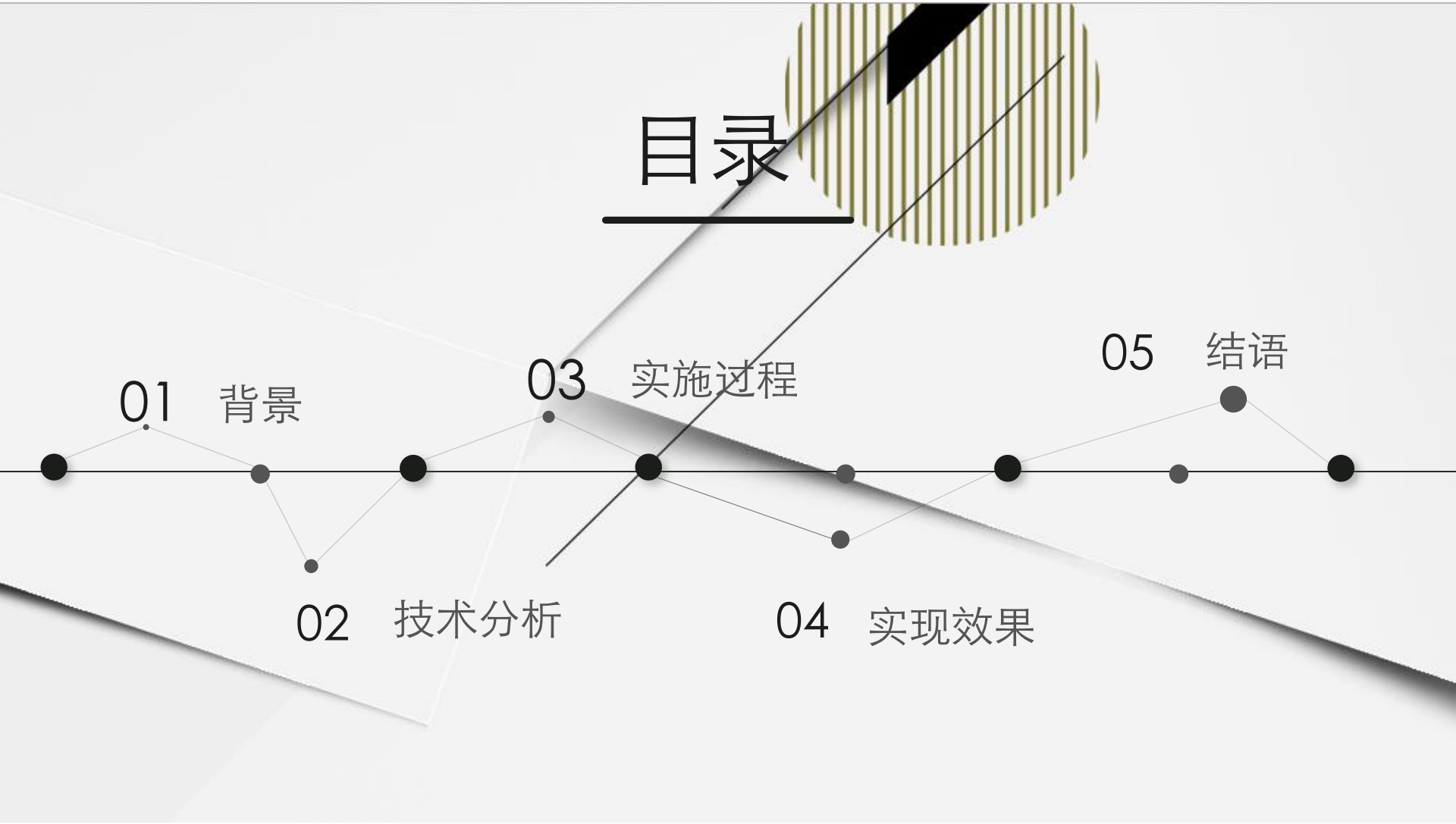
中国矿业大学可视化云图

2018年5月

中国矿业大学　　邓志文　都平平

目录



1

Part 01 背景

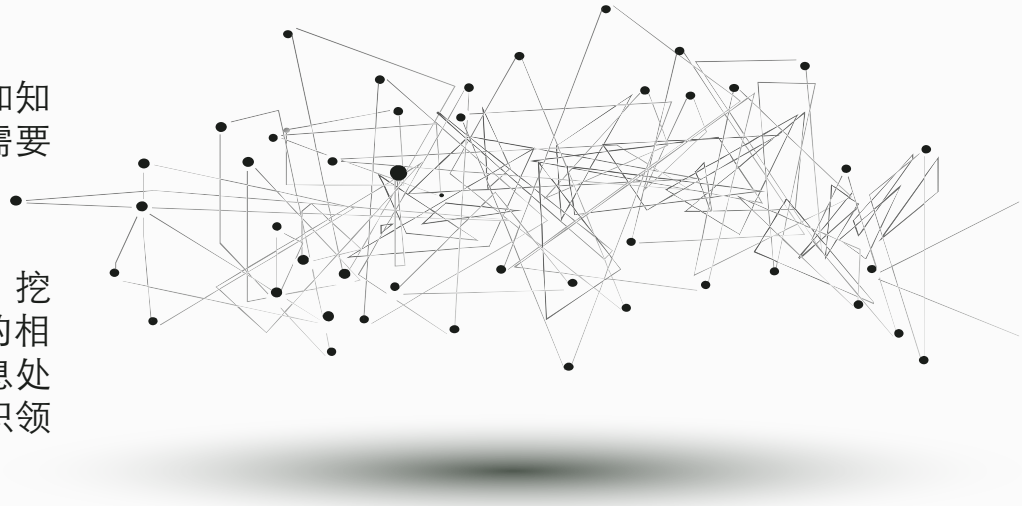
01

背景和意义

»高校图书馆作为机构的信息情报中心，存储和管理着大量的研究文献和机构知识。

»采用文献计量分析方法从中获取有价值的信息和知识是图书馆情报服务的重要内容，而这个过程需要科技分析人员花费大量时间和精力。

»图谱分析用可视化技术描述知识资源及其载体，挖掘、分析、构建、绘制和显示知识及它们之间的相互联系，把复杂的知识领域通过数据挖掘、信息处理、知识计量和图形绘制而显示出来，揭示知识领域的动态发展规律。



01

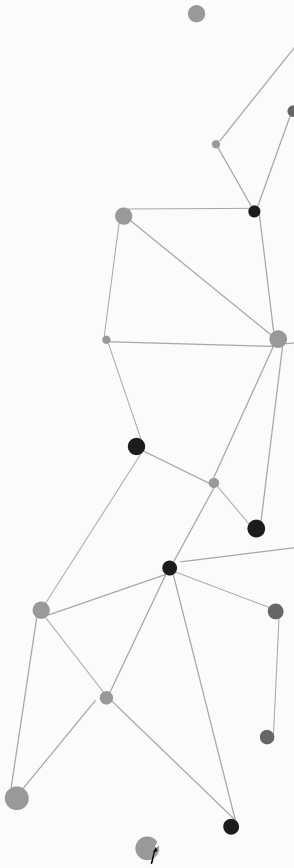
背景和意义

»机构库开源平台Dspace本身没有这种可视化分析模块，很多研究者都是基于第三方工具如Citespace 、 VosViewer等对Dspace进行数据可视化的图谱分析，分析结果都是静态图，并没有将其嵌入到机构库实现其动态可视化。

»北京大学、清华大学以及中科院等采用可视化图谱分析技术对Dspace进行二次开发，实现了可视化设计。

»不同的图谱分析技术都有其特点，可视化的价值实现有赖于它的视觉表现形式。

»本案例采用第三方可视化工具进行二次开发，并对Dspace数据的动态实时可视化分析。



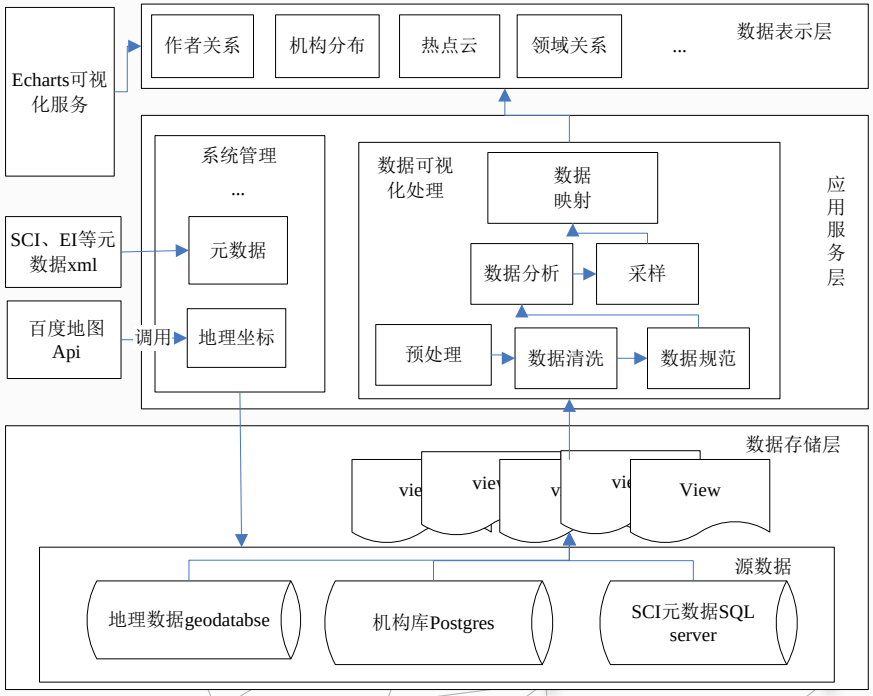
2

Part 02 技术分析

02 技术分析

» 平台体系结构。

平台系统架构如图所示，平台主要包括：数据存储层、应用服务层、数据表示层。

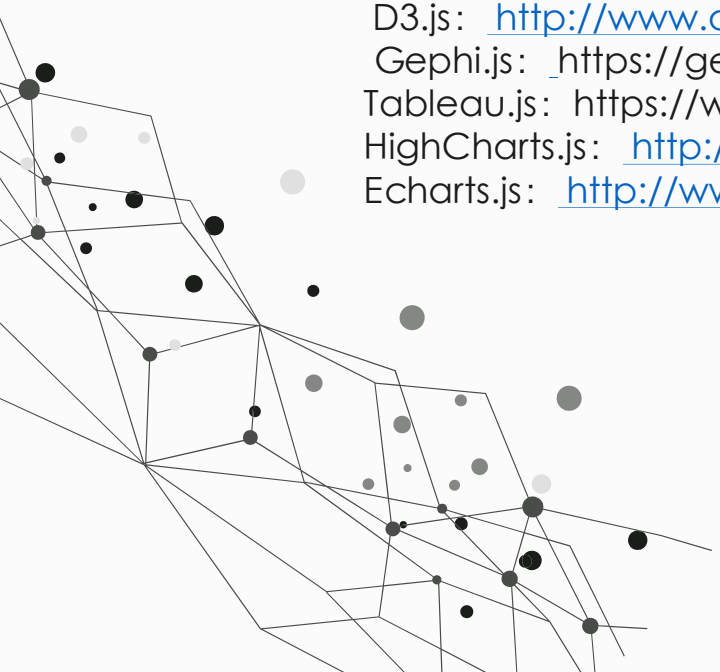


02 技术分析

» 图谱可视化开发包。

互联网中也出现了很多数据可视化开发包，如：Arbor.js、D3.js、Gephi、Tableau、HighCharts、Echarts。在充分查阅资料后，我们选取了Echarts作为平台可视化开发包。ECharts是一款由百度前端技术部开发的，基于Javascript的数据可视化图表库，提供直观，生动，可交互，可个性化定制的数据可视化图表。Echarts体积小，开源，而且是国内人员开发，文档丰富，学习相对容易。

Arbor.js：http://arborjs.org/
D3.js：http://www.oschina.net/p/d3
Gephi.js：https://gephi.org/
Tableau.js：https://www.tableau.com/products/desktop
HighCharts.js：http://www.oschina.net/p/highcharts
Echarts.js：http://www.oschina.net/p/echarts

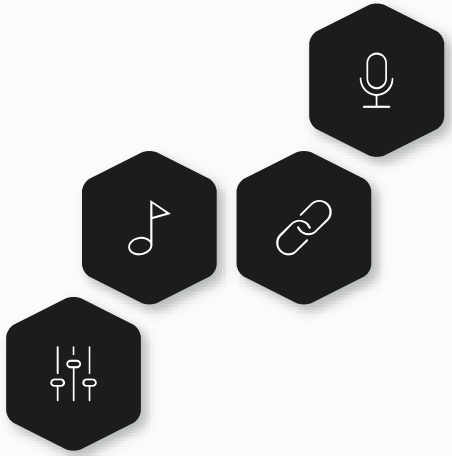


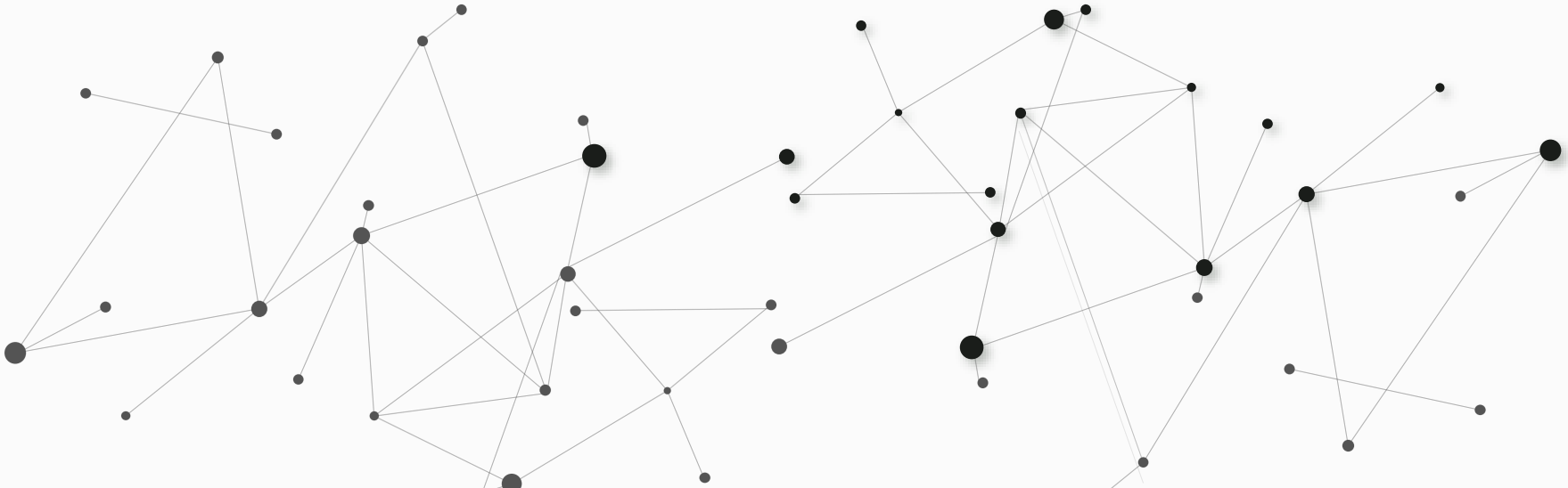
02 技术分析

» Dspace数据库结构分析。

本文以第三方可视化工具，研究一种可二次开发的可视化分析技术并与Dspace等进行整合。

数据可视化需要有数据的支持，Dspace平台的后台数据库为Postgres，表有40多上，为了便于可视化设计，需要分析Dspace数据库各表功能，尤其是条目表item、元数据表metadatavalue，确定各表之间的关系，建立用于数据可视化的视图，本文在Dspace基础数据库的基础上自定义了多张视图结构。





02 技术分析

»元数据的清洗。

由人工向机构库录入数据的过程中，由于操作不规范，在格式和内容上存在一些问题，如关键词间用逗号、汉字的分号，内容中有不该存在的字符等，这就造成元数据中存在异常的数据，需要对这些引起异常的数据进行清洗去除。因此需要分析所有异常数据可能性，并采用一定的方法最大限度的降低这些异常数据。

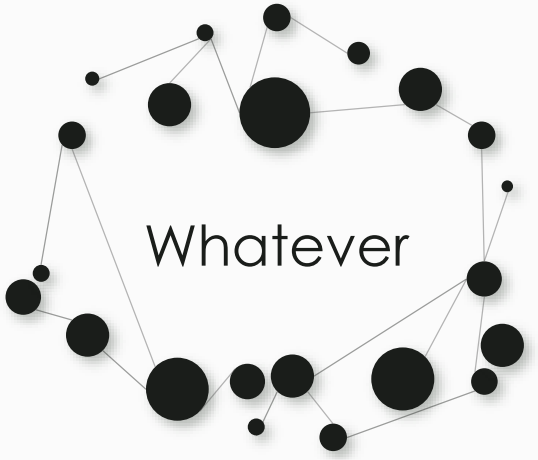
»优化的数据生成算法。

我馆机构库中已经有元数据接近100万级别，而数据可视化有需要对这些元数据进行多次的查询，也就是说一个可视化过程需要处理的累计数据量可能是千万级。因此需要对数据操作算法进行充分优化，能快速的响应可视化数据生成需求，本文主要从三方面去对数据处理过程进行优化：1采用内存数据操作；2 建立内存数据索引；3多分辨下的数据请求，对不同的请求分析不同的数据量。

3

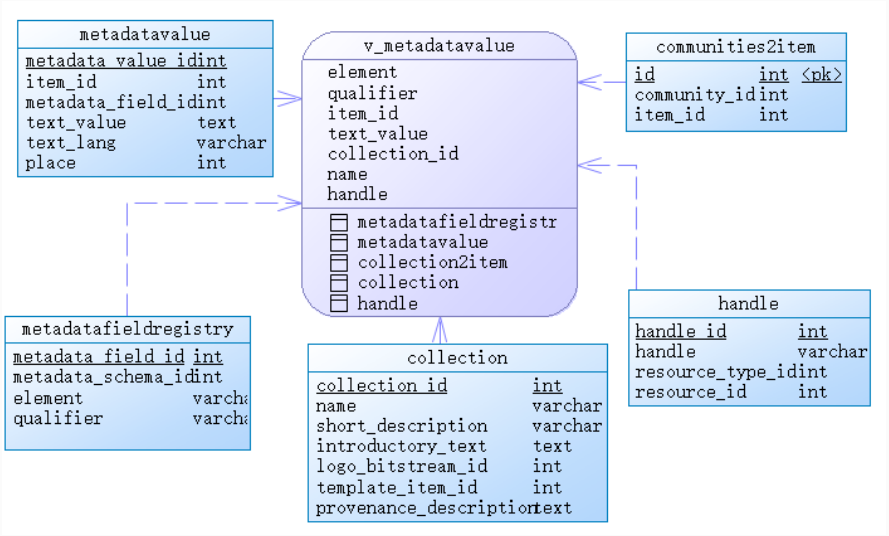
Part 03 实施过程

03 实施过程



- **多数据源整合**

为了向应用层提供统一规范的元数据，需要对各种源数据进行整合，从这些异构或异源的数据中自动抽取信息得到备用知识单元。对于Dspace平台通过重新定义各视图组织各元数据，本文定义的几个关键视图主要包括：v_metadatavalue, v_commun ity2collection。



03 ➢ 实施过程

- **数据清洗**

数据清洗的主要原理：利用有关技术如数理统计、数据挖掘或预定义的清理规则将脏数据转化为满足数据质量要求的数据。

- （1）数据输入造成的单词空格、符号中英文输入不统一等问题。
- （2）同一内容用不同关键词表达的问题。
- （3）数据不一致性等问题。

03 实施过程

信息抽取

知识图谱最适合处理关联密集型的数据，因此首先需要存放的是图谱中的节点和边的数据；本文采用自底向上的方式从各种数据源中提取出实体（概念）、属性以及实体间的相互关系，在此基础上形成本体化的三元组知识表达G=(E,R,S)。

$E = \{e_1, e_2, \ldots, e_n\}$是实体的集合，n 为实体的数量；

$R = \{r_1, r_2, \ldots, r_m\}$是实体间关系的集合，其关系 R 结构定义为：$r = (E_s, E_T, V_{weight})$，属性$V_{weight}$是表示两实体间关系的密切程度，它反应到图谱中可以是节点间的距离。

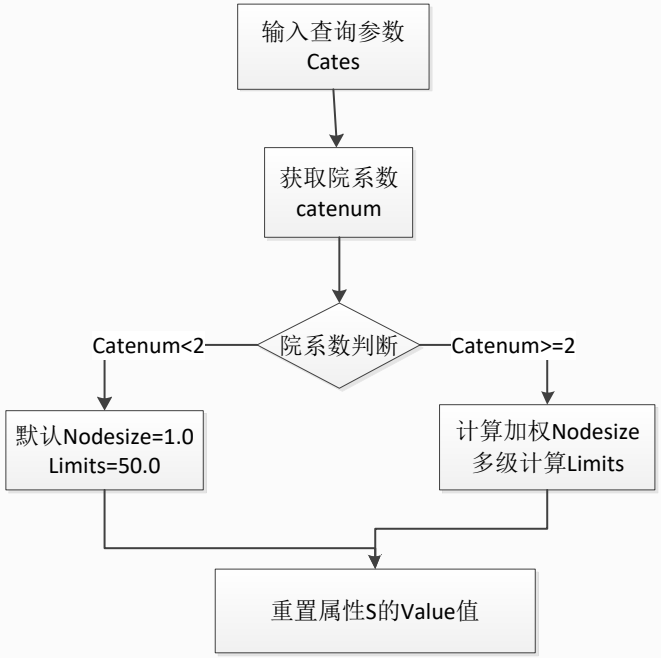
$S = \{s_1, s_2, \ldots, s_t\}$是实体属性值集合，属性 s 结构定义为：$s = (\text{Category}, e_i, \text{Value})$，其中 Value 表示实体$e_i$的重要程度，它反应到图谱中可以是节点的大小。

03

实施过程

- **加权取样**

由于各单位以及个人发文量不一样，差别很大，属性集S的Value值也差别非常大，如不加处理会导致图谱中展现的节点大小差距也非常大，显示效果会很不美观和协调，因此需要对数据源进行预处理，本文设计出一种不同粒度下多级采样和数据加权处理方法。处理流程：

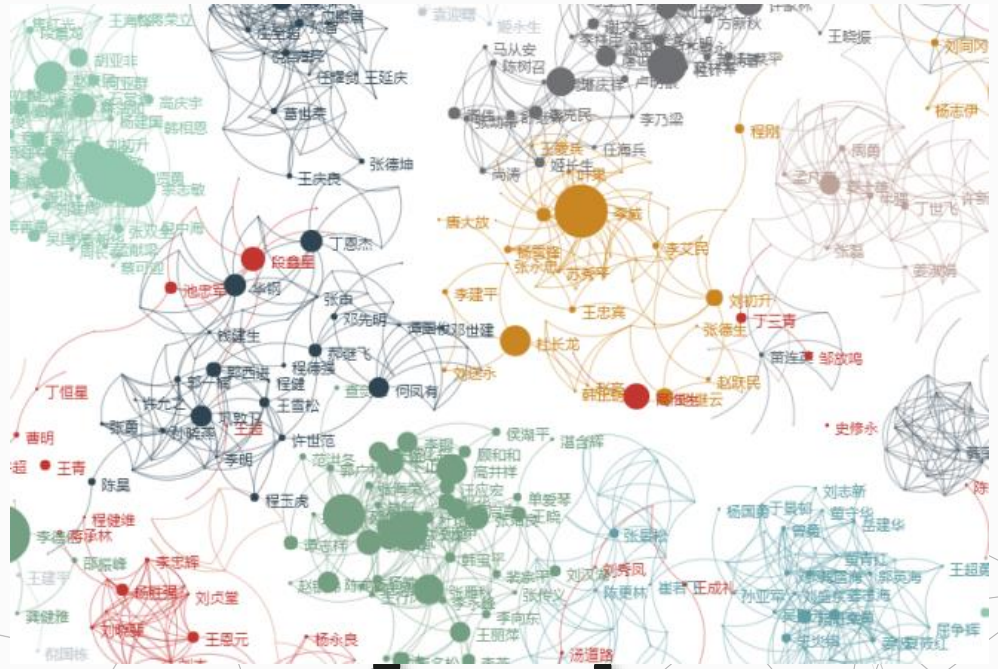


4

Part 04

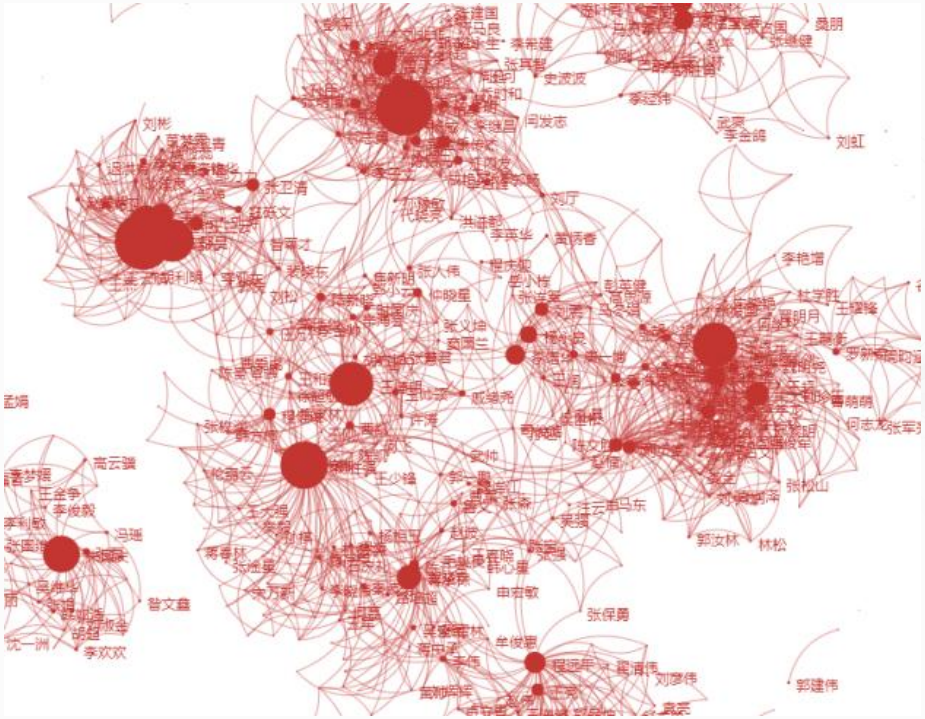
平台实现

04 ➢ 平台实现


王庆良
张德坤
丁恩杰
殷鑫星
池忠军
华钢
张宙
邓先明
钱建生
蒋建飞
程伟强
程健
何凤有
丁恒星
许允之
王雪松
张勇
王岩
许世范
李明
程玉虎
曹明
王青
陈昊
程健雄
李宏辉
刘贞堂
王恩元
杨永良
王丽萍
季向东
张传义
陈更林
刘双凤
张宏宇
单爱琴
顾和和
高井祥
侯湖平
谭合辉
李晓
范洪冬
张昌松
刘志新
王超
史修永
刘向阳
杨志伊
程刚
周勇
丁世飞
张磊
姜洪涛
李乃坚
王爱兵
任海兵
尚涛
姬长生
唐大放
李新
李文民
杨雪锋
张尔东
苏秀平
李建平
王忠宾
杜长龙
刘初升
张德生
丁三青
邹放鸣
赵跃民
韩世云
王晓振
方新秋
马从安
陈树召
程庆祥
卢明振
王延庆
晋世荣
高庆宇
孙亚宁
王成礼
崔君
于贵和
曾子华
岳建华
王超强
吴青红
邹克海
吴强
李炎锦
吴晓红
尹争辉
杨国勇

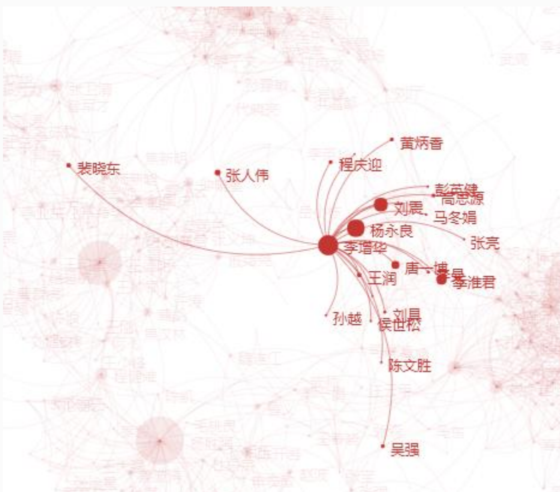
04

➢ 平台实现



04

平台实现



黄炳香
裴晓东
张人伟
程庆迎
马冬娟
杨永良
李增华
张亮
王润
孙越
侯世松
陈文胜
吴强

无线传感器网络
遗传算法
监控系统
煤矿
PLC
仿真
Web服务
CAN总线
支持向量机
数据挖掘
云计算
神经网络
软件测试
数据采集
模糊控制
远程监控
嵌入式系统
矢量控制
微粒群优化
特征提取
计算机仿真
迁移学习
信息融合
小波分析
交互式遗传算法
关联规则
过程控制
开关磁阻电机
无刷双馈电机
时间序列
开关磁阻电动机
多目标优化
路径覆盖
双馈电机
通信协议
半监督学习
计算机网络
提升机
单片机
聚类
矿井
视频
语义Web
ANSYS
电流控制
进化优化
领域本体
可靠性
选煤厂

04

平台实现

5

Part 05 结语

05

结语

本案例研究了基于echarts组件对机构知识库进行图谱可视的关键技术，并依据中国矿业大学图书馆dspace机构库进行了知识图谱分析的实现。

其主要贡献：

1.给出了图书馆构建自己的机构知识图谱平台架构。

2.采用开源可视化开发包实现信息动态可视化。

3.采用加权取样算法实现信息的抽取和计算。

4.结合地图实现位置可视化分析。

Thank You