

拓尔思水晶球

基于动态本体的知识管理实例



从 数 据 到 智 慧

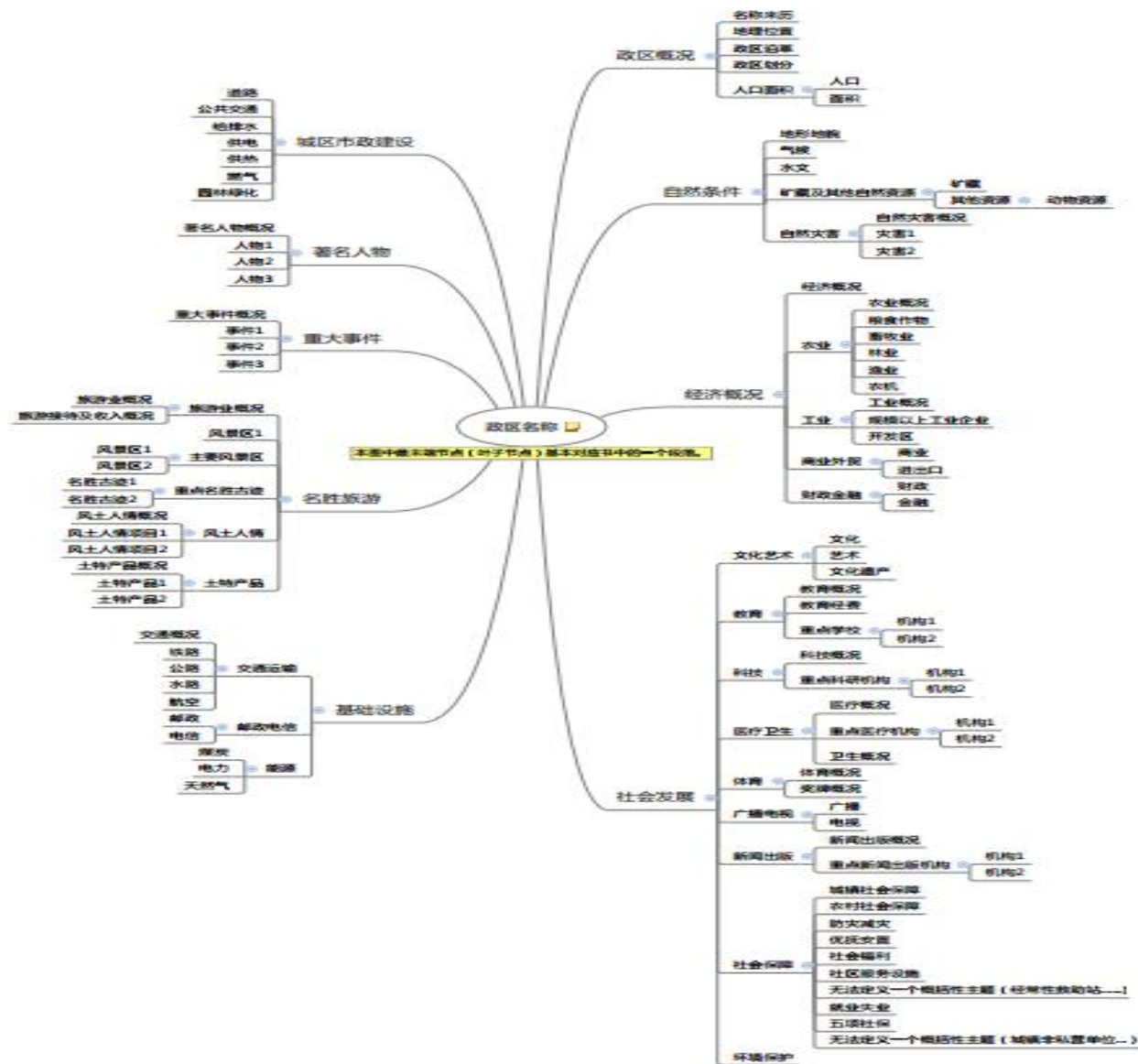
Data • Knowledge • Intelligence

本体是什么

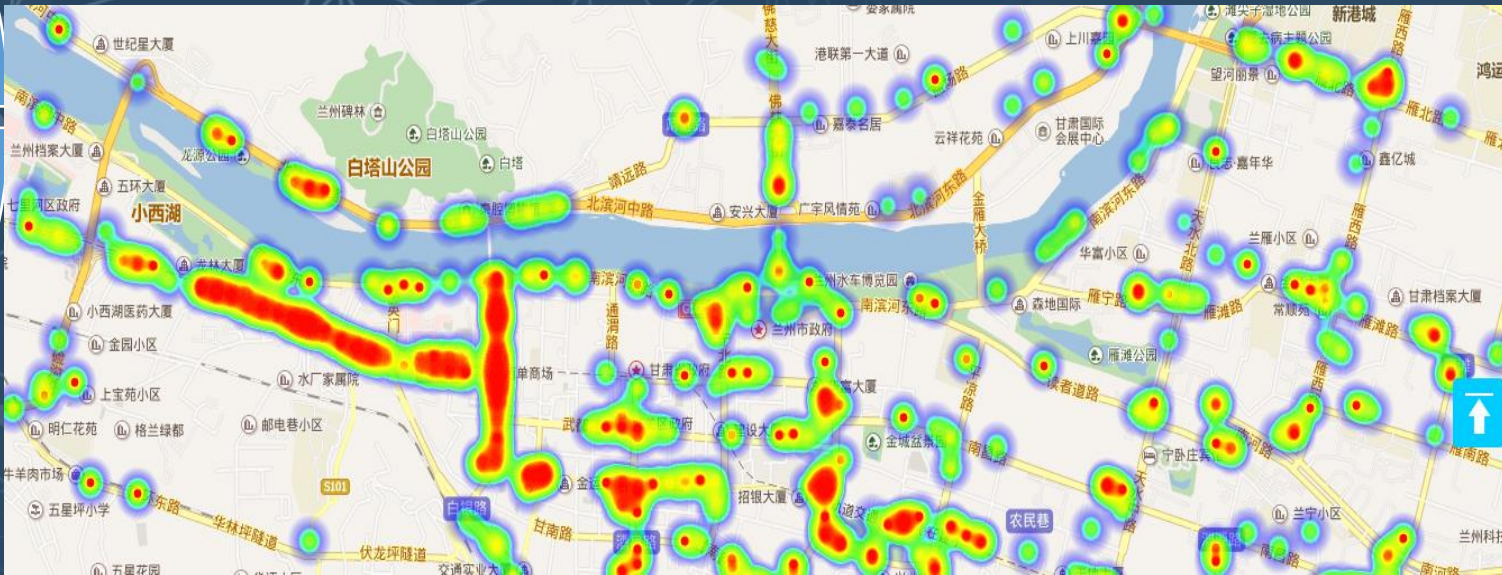
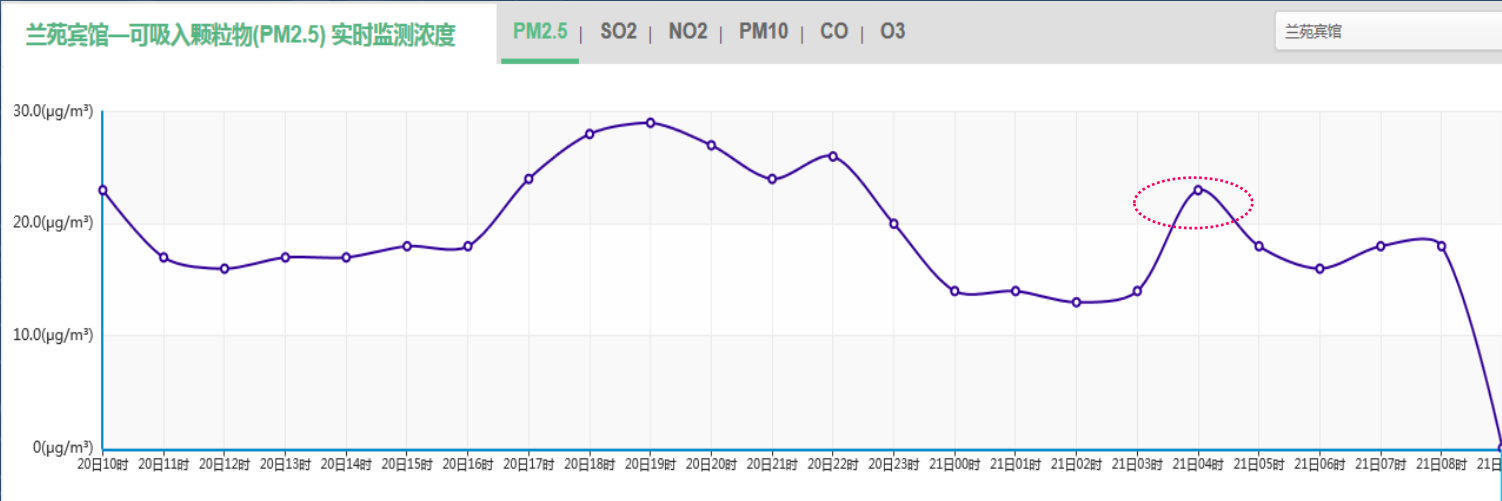
- ❶ 哲学定义形成现象的根本实体，含义指事物的本身；引申为根本的。是事物的主体或自身，事物的来源或根源。
- ❷ 在人工智能领域Neches等人将本体定义为“给出构成相关领域词汇的基本术语和关系，以及利用这些术语和关系构成的规定这些词汇外延的规则的定义”。

知识图谱

- ① 知识聚集与挖掘
- ② 知识抽象与可视化
- ③ 找到最想要的信息
- ④ 提供最全面的摘要
- ⑤ 发现知识深度和广度

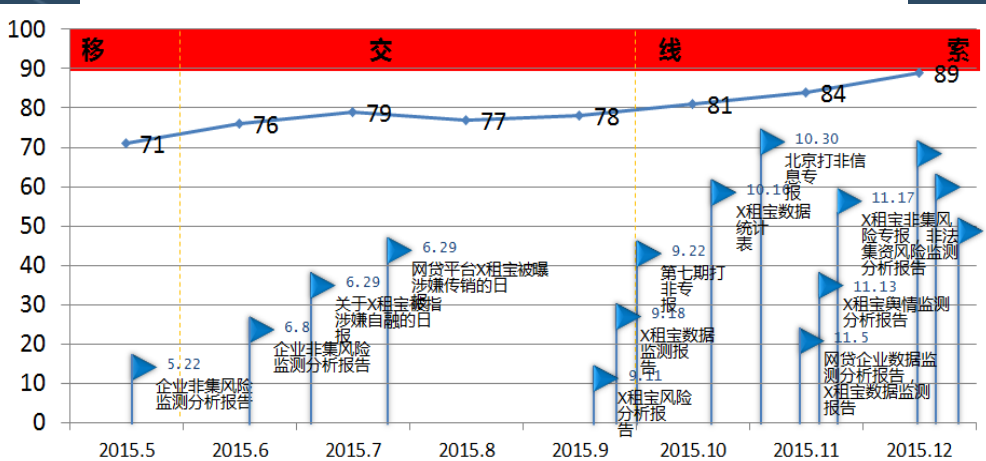
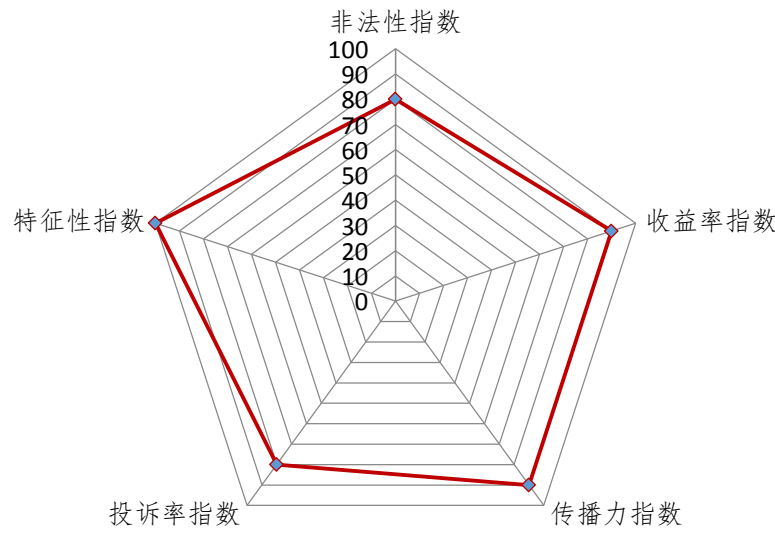
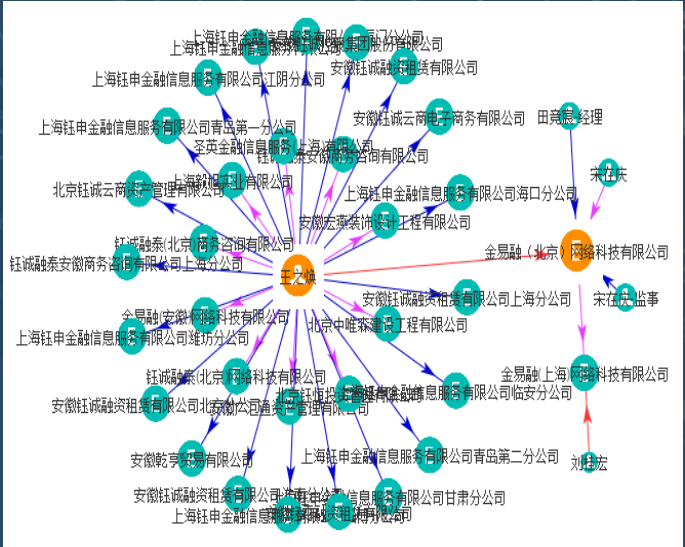
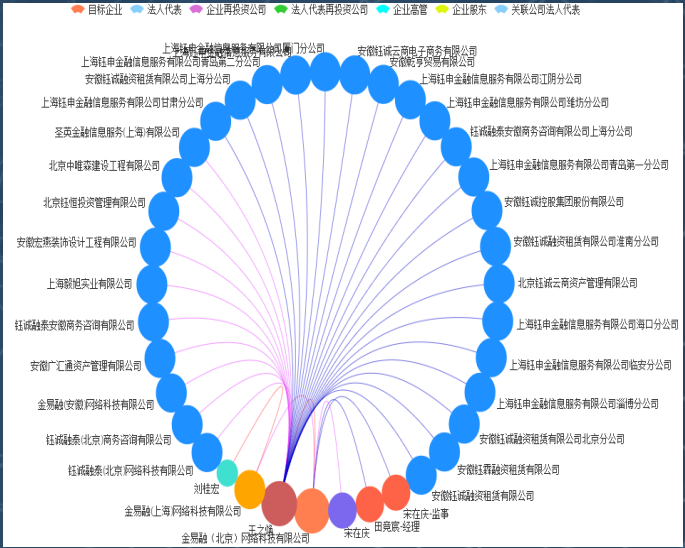


创建知识图谱的过程



当本体论遇上知识图谱

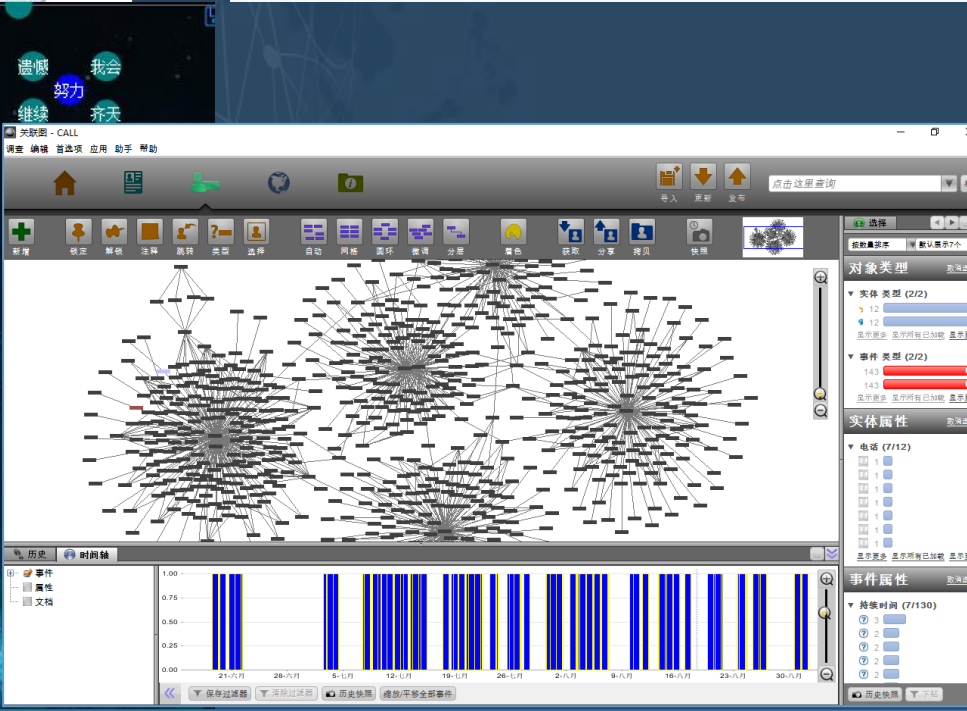
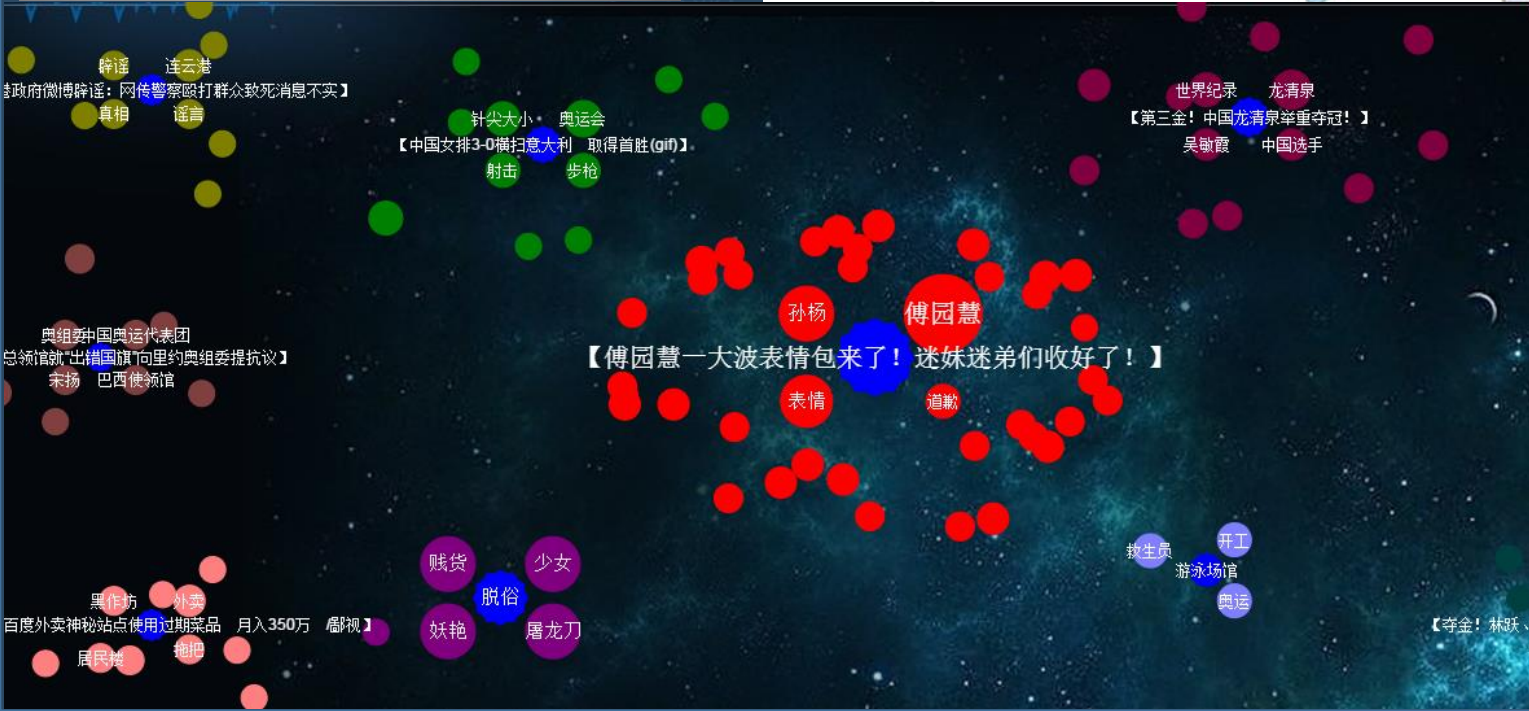
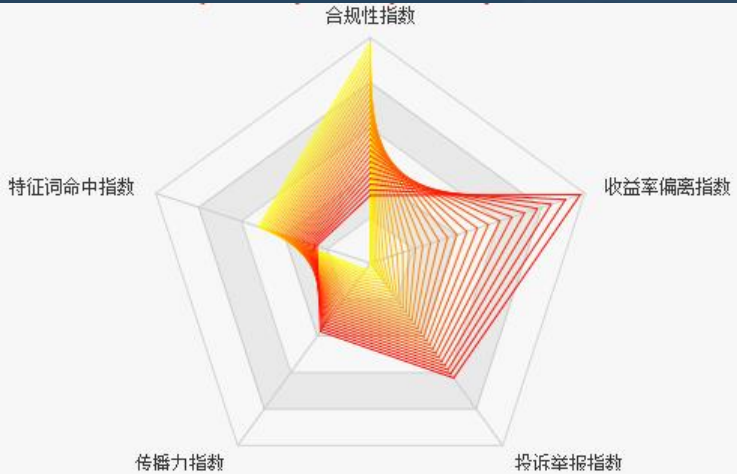
- 1 本体构建
- 2 实体构建
- 3 异构数据融合
- 4 动态发现
- 5 关联挖掘
- 6 预测



知识图谱面临的转变：从“整合”到“融合”



知识图谱面临的转变：从情报计算到计算情报



知识图谱面临的转变：从薄数据到厚数据



【黄萱菁】 【澎湃新闻】
了什么成就？】 <http://www.thepaper.cn/newsDetail.html?categoryId=1&articleId=61111>

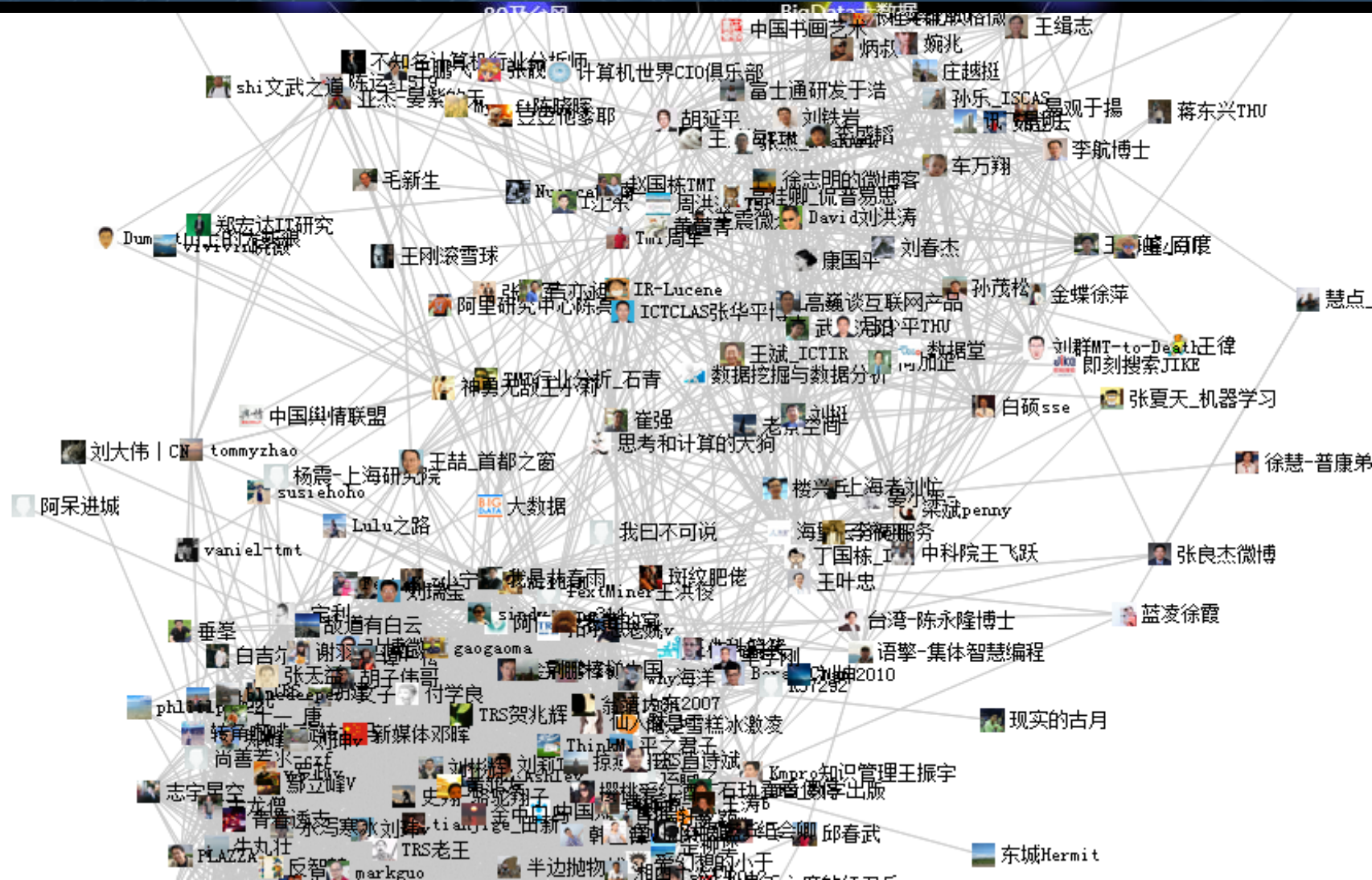
2015-01-29 22:18:00 来自微

原 拓尔思负责实施的iP...
 验访问：0网页链接有网友...
 努力。

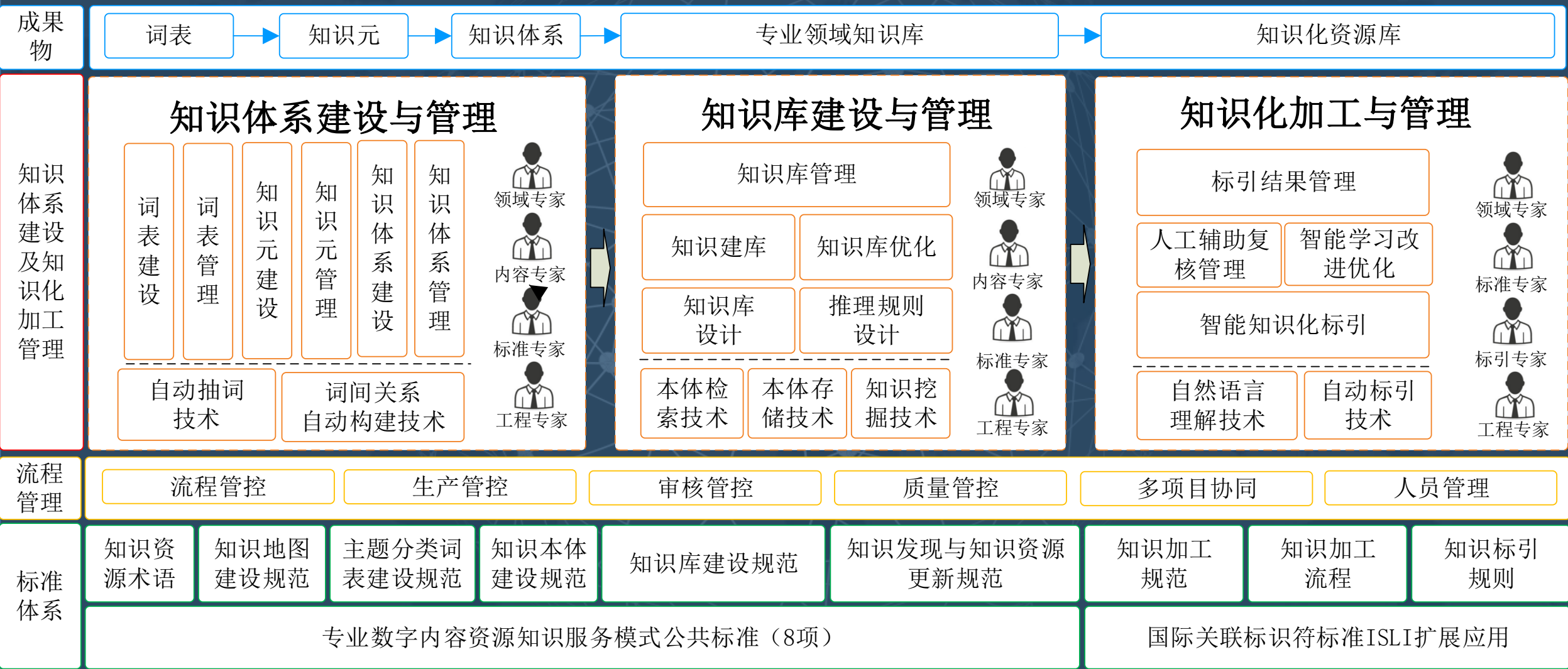
2015-01-28 09:41:00 来自微

原 新年新气象，请朋友
理以及厦门办事处经理，
hi.shuicai@trs.com.cn

2015-01-08 08:41:00 来自微博



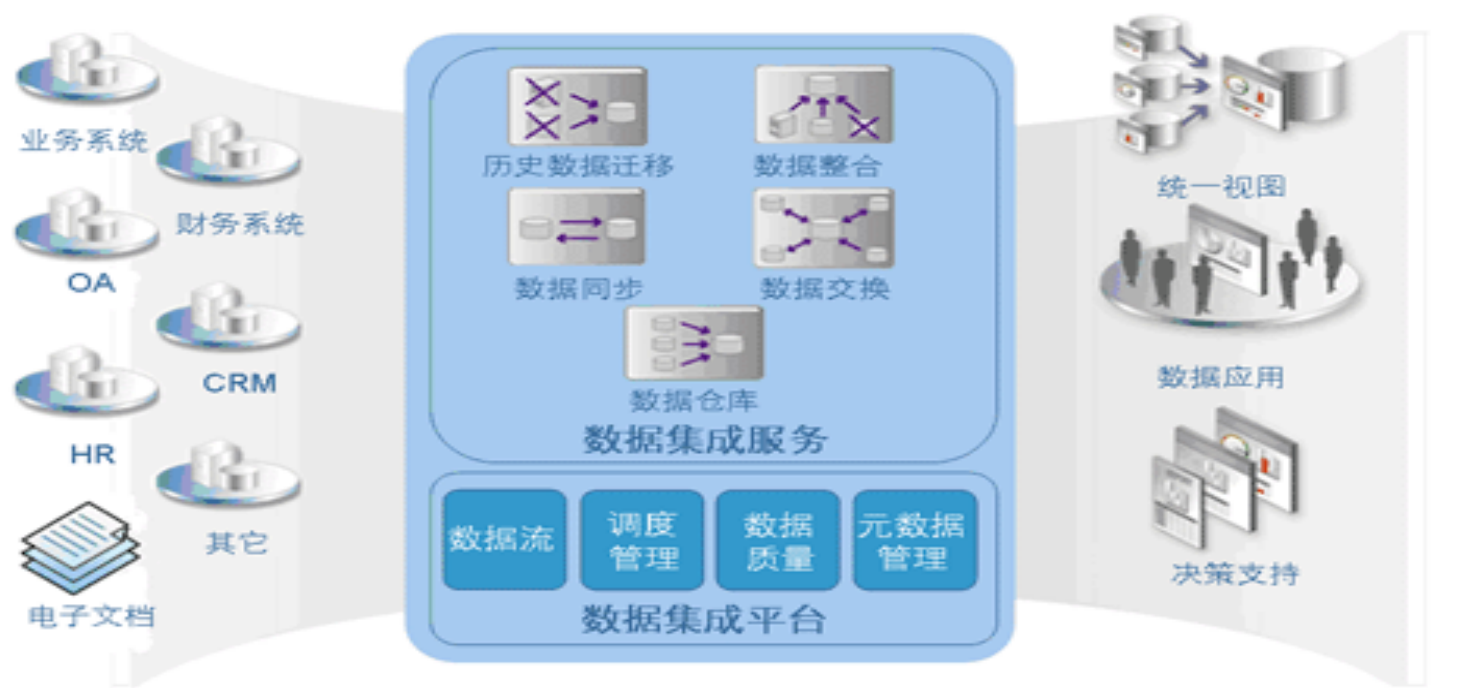
一、建立知识库建设总体框架



二、梳理知识库的构建流程

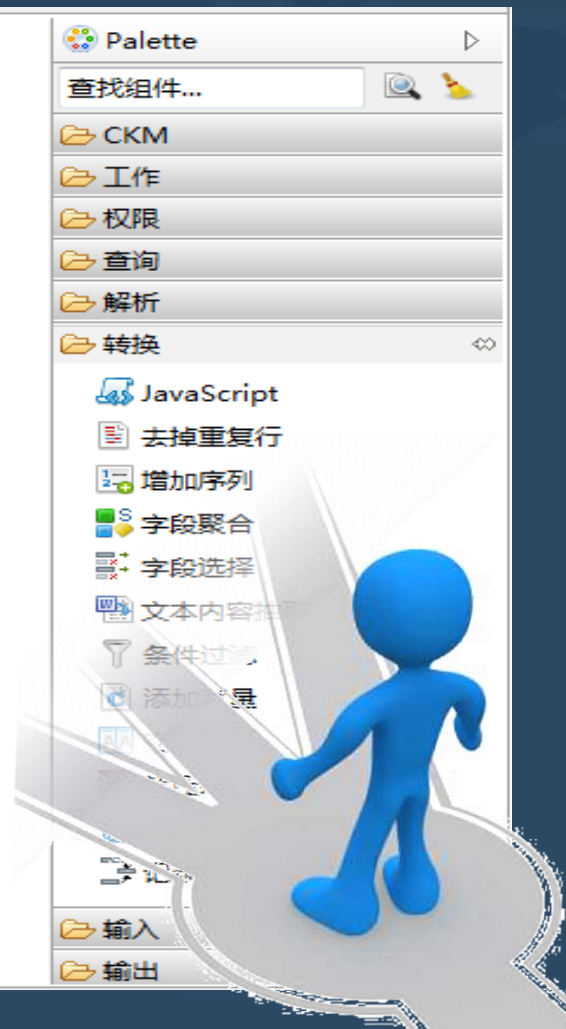
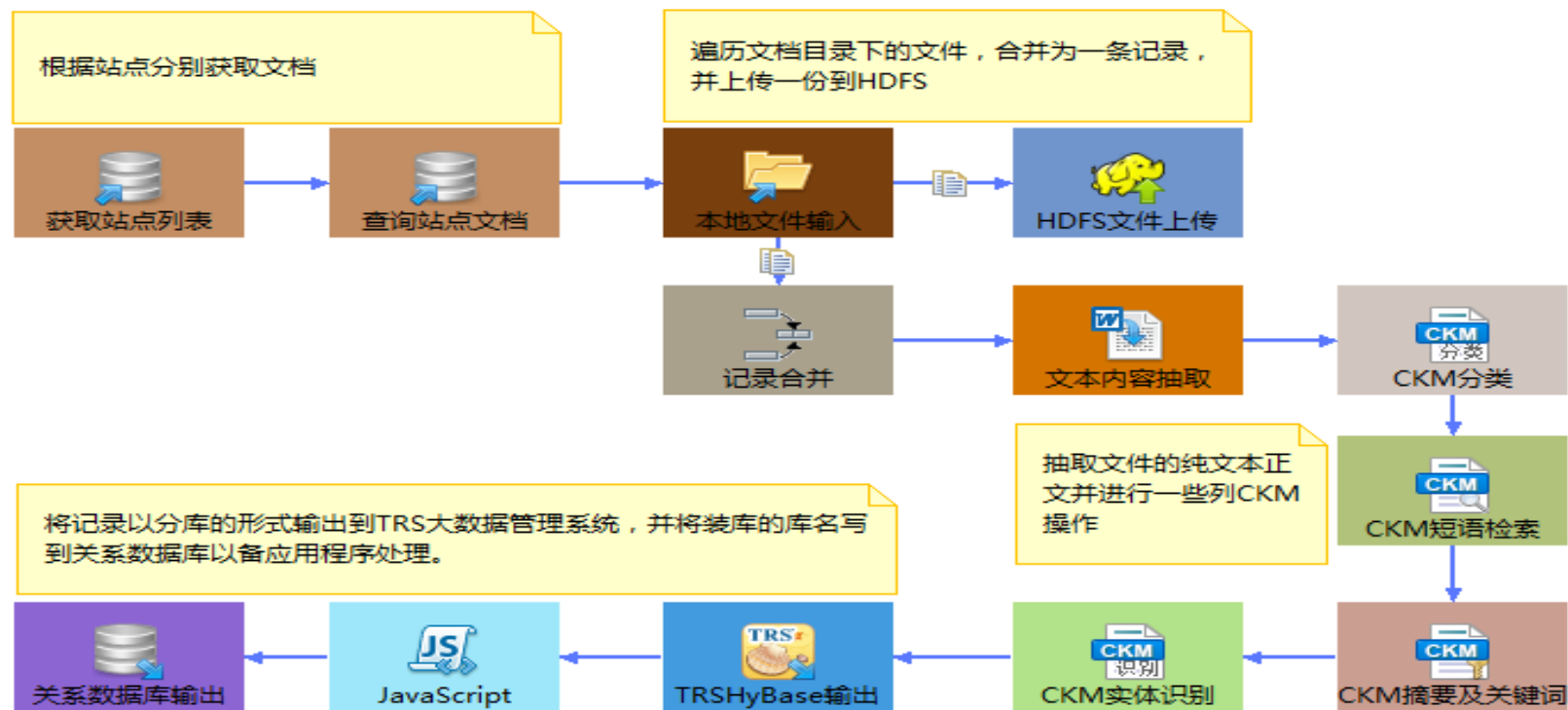


三、获取数据——TRS Adapter数据集成平台



- 泛数据类型支持能力
- 强大的ETL批量调度能力
- 可视化任务配置管理能力
- 企业级运行监控能力

TRS Adapter数据配置管理流程



四、智能挖掘——TRS CKM

从非结构化文本中发掘有价值的信息

- 自动分类组件
- 自动聚会类组件
- 信息抽取组件
- 关联关系挖掘组件
- 相似检索
- 自动摘要



项目	性能指标
文本分类	- 自动分类正确率：在几十个类别之内，经过分类体系精心组织和语料细心训练后，召回率达到 85% 以上，分类准确率达到 86% 以上。 - 单用户下能完成自动分类 31 篇/秒，50 并发用户能完成 215 篇/秒。
相似性检索	- 细排重：10 万排重库，单用户时每秒执行细排重 25 次；100 万排重库，单用户时每秒执行细排重 5-6 次。 - 粗排重：1000 万排重库，单用户时每秒执行检索 40 次，16 并发时，每秒执行检索 280 次左右；
自动摘要	- 单用户能完成摘要 31 篇/秒，50 个并发完成摘要能达 266 篇/秒。 - 支持大文件的摘要，轻松处理几十 M 字节文本的摘要；
文本信息抽取	- 实体识别（人名地名机构名）准确率为 80% 左右。 - 单模块抽取速度： 实体识别 500KB/秒；案件名识别 400KB/秒； 房屋信息识别 2.5MB/秒；数字识别 10MKB/秒； 简历抽取 200KB/秒；政府信息公开目录抽取 350KB/秒； 基金信息抽取 500KB/秒； - 设置抽取所有实体串，包括：人名地名机构名、案件名、房屋信息、数字抽取。单用户时完成信息抽取 15 篇/秒，16 个并发时为 95 篇/秒；
拼音检索	单用户时每秒执行拼音检索 162.9 次，到达 50 个并发时，每秒执行拼音检索次数为 330.8 次
相关短语检索	单用户时每秒执行相关短语检索 89.1 次，到达 50 个并发时，每秒执行相关短语检索次数为 336.7 次；
文本聚类	- 对一般的网络新闻文本，可以达到 5000 篇/分钟 - 单用户时每分钟执行聚类 1 次，到达 8 个并发时，每分钟执行聚类次数为 4 次；（每次参与聚类的文本数量为 5000 篇，每篇文本的平均长度为 10KB。 - 自动聚类准确率达到 75% 以上，满足大多数应用的实用要求。
（政治）常识校对	速度平均为支持每秒 100K 文本以上。（环境：P4 2.4G，512M 内存）
文本信息过滤	平均速度为每秒 40 篇以上。（环境：p4 1.5G，512M 内存）
高速串匹配	200 个常规字符串模式的匹配速度为每秒处理文本 200M 左右，环境 IBM XSERIES_3650(CPU 主频 2.49G,内存 4G)，操作系统 Windows Server2003)

TRS CKM智能信息处理应用

文本挖掘应用

可视化分析

语义检索

敏感信息
监测

文本信息
结构化

内容推荐

... ..

文本挖掘功能组件

智能检索

关联关系
展示

信息过滤

常识校对

热点发现

词典挖掘

语义关系
抽取

褒贬分析

关键词标引

信息排重

分词/词性
标注

信息抽取

自动分类

文本摘要

自动聚类

HADOOP/MapReduce

机器学习

自然语言处理

信息检索

数据挖掘

理论基础

分词词典

语义词典

语料库

语法词典

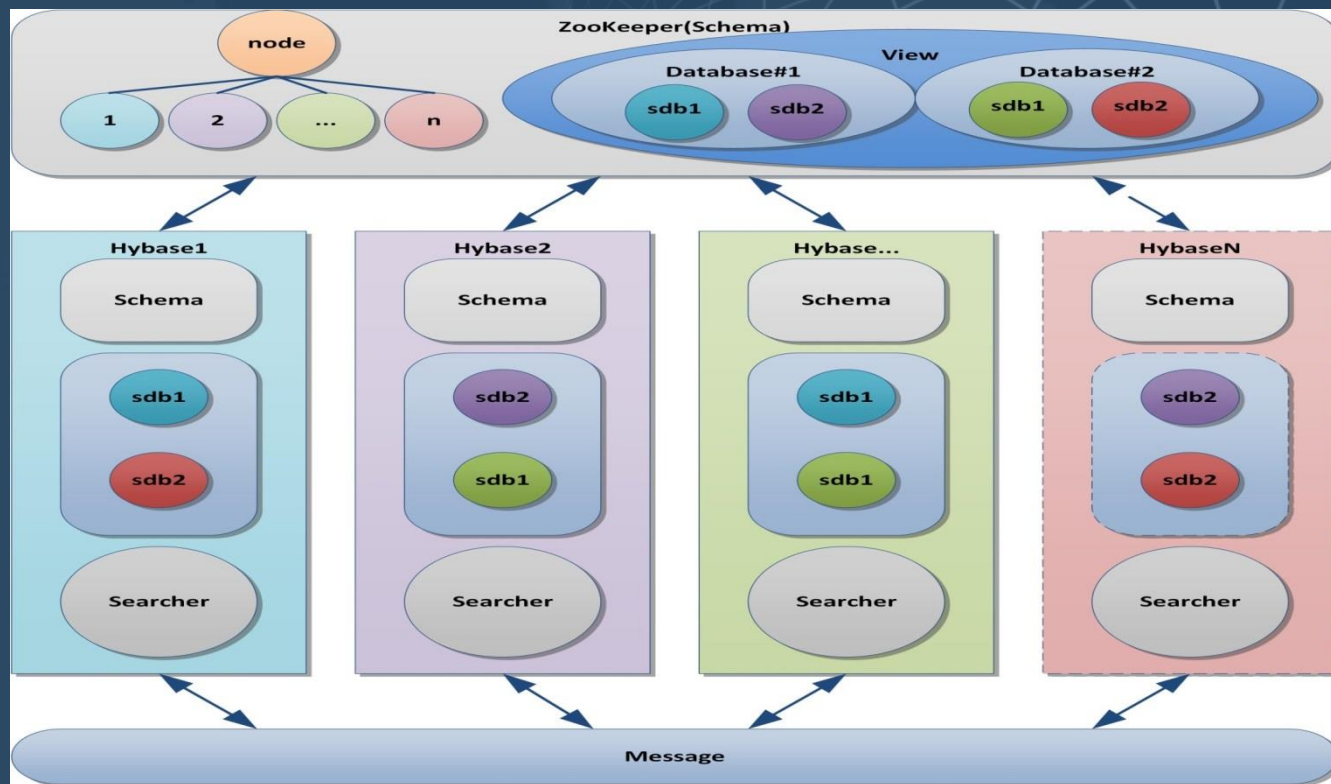
语言学资源

实时热点分析

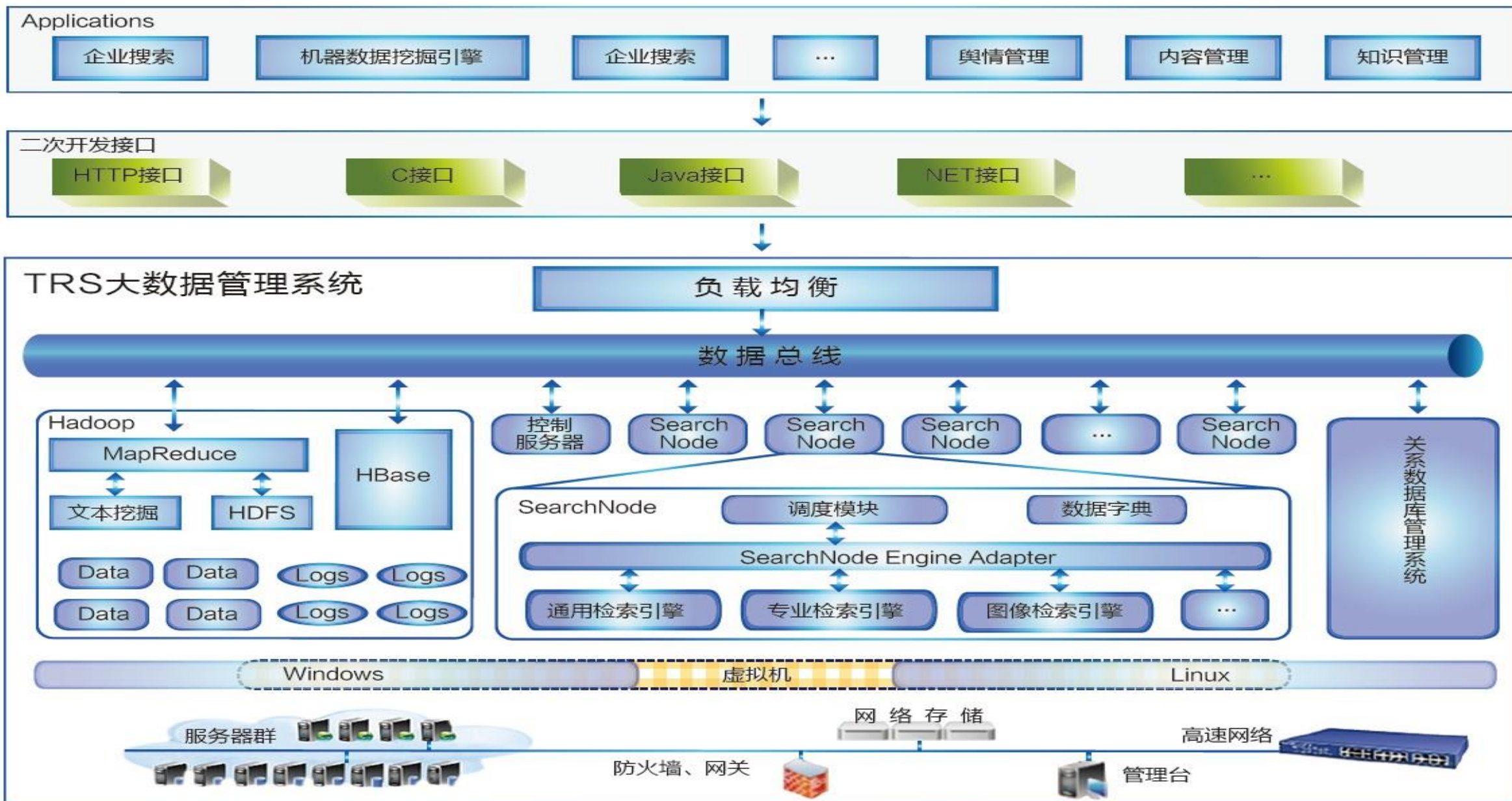


五、海量数据管理——TRS海贝

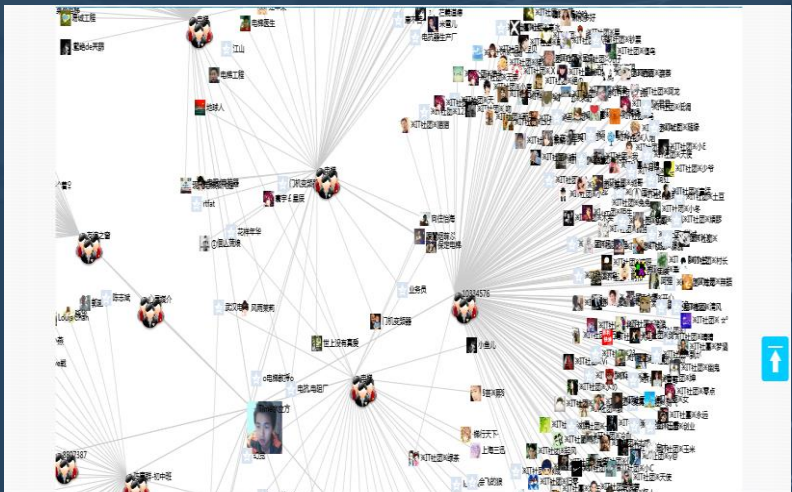
- 融合文本检索、分布式并行计算，通过多引擎、索引分片、多副本、对等节点（去中心化）、列数据库存储机制，借助自然语言处理、Hadoop HDFS等技术，无缝支持Spark，而研制的非结构化数据管理系统（NoSQL），可以为各类分析应用提供非结构化数据高效管理和智能检索的平台支撑。



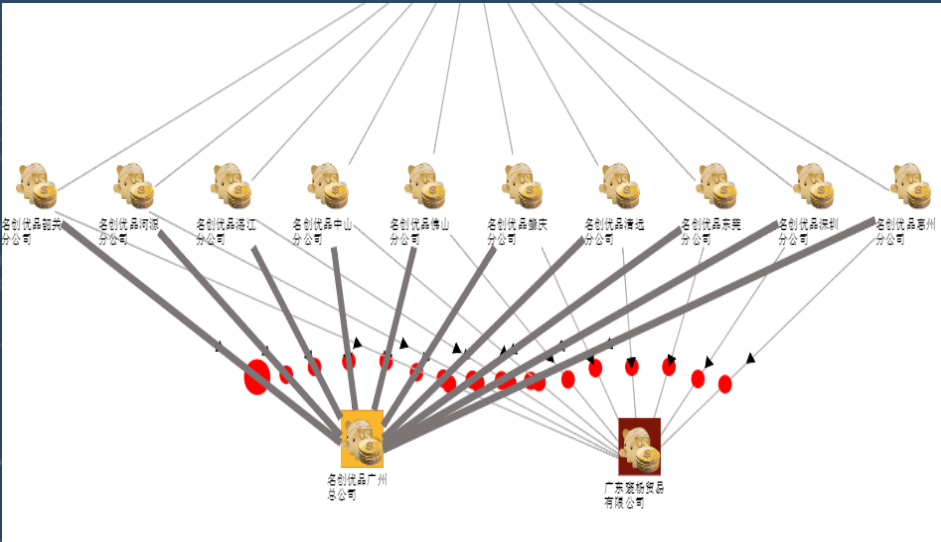
TRS海贝大数据管理系统架构



可视化让结果更直观



社交媒体可视化



数据流可视化

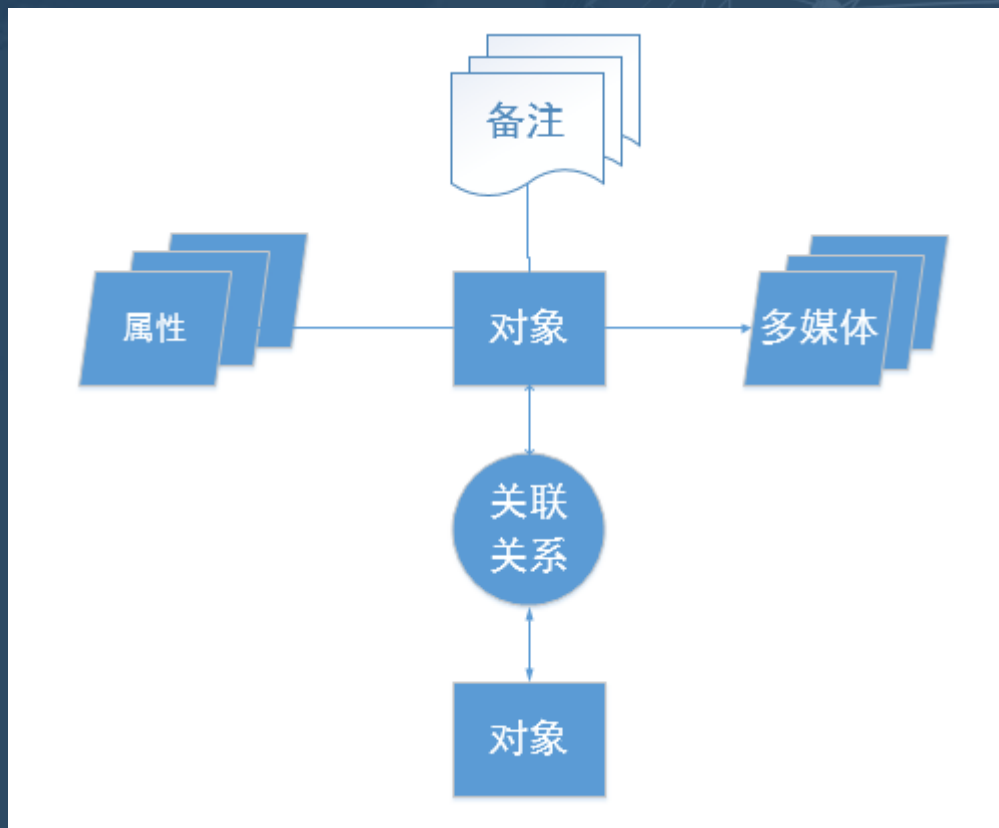


访问行为可视化



可视化大屏指挥

TRS水晶球动态本体的知识管理



- 对象—代表世界上的事物。
 - 对象包括：实体、事件、文档、多媒体
- 属性—对象的属性
- 关系—对象之间的联系
- 备注—对象相关的文字备注
- 多媒体—对象的多媒体附件
- 通过“实体 - 关系 - 实体”三元组描述物理世界中的概念及其相互关系

面向领域的分析师工具：本体+数据+扩展



从数据融合到知识图谱

数据融合

结构化数据、文档、多媒体、网页、邮件等，通过工具由系统或个人导入到平台中。

检索发现

通过检索（对象、文本、地图等）、直方图下钻、关系图、事件轴等方式，交互式分析数据、发现知识。

知识管理

自主添加更新各种对象、事件和关系，通过平台能够很好的管理起来。

协同工作

工作成果分享、传递、互动，工作组内协同工作，传递分析过程或者结果。

通过实体、事件、属性、关联抽象知识本体

实体可以是人、物品、文档等任何内容的抽象



↓ 0

↑ 18

安雅

相关

2

18

0

对象信息

类型: 个人

创建人: trs test

创建时间: 2016/07/25 16:39:53 +08:00

谁在关注这个?

没有关注者

关注此对象

摘要

属性

相关的

注释和多媒体

历史

属性

类型	值
名称	安雅
国籍民族	汉族
地址	广州市白云区尖山路朝晖花园 贵州省石阡县河坝场乡金坪村安家寨组
所属大学	贵州大学
证件号码	5222241989122***29

添加到关联图

注释

相关事件

类型	关系	标签	文本
人员火车出行 (18)			

添加到关联图 添加到地图

相关实体

关系	相关	起始日期	截止日期	文本
拥有 (1)	贵gg6958			
母亲 (1)	文微			

实体有各维度属性

实体与实体、实体与事件存在相应关联

数据融合

基本信息



性别: 男
毕业院校: 西安电子科技大学
学历: 硕士
学位: 硕士
专业: 计算机科学与技术
婚姻状况: 已婚
现任职务: 发表100多篇论文, 1993年易主, 长等职务, 创业板上市

现任职务:
(股票简称)

施水才

相关: 6
1
2

对象信息
类型: 个人
创建人: trs test
创建时间: 2016/09/20 13:56:12 +08:00

谁在关注这个?
没有关注者
关注此对象

摘要 属性 相关的

属性	值
Email	shi.shucai@trs.com.cn
云搜出生日期	1966年4月3日
名称	施水才
性别	男
所在地	江苏省溧阳市
所属大学	西安电子科技大学
电话	13901188243

相关事件

类型	关系
事件 (1)	简单

相关实体

关系	相关
同事 (1)	安淮
简单 (5)	北京信 北京信 南京大 清华大 西安电

添加到关联图

注释

没有相关注释。[点击此处添加。](#)

多媒体

image2.jpeg 水晶球...

TRS水晶球数据导入

步骤 1 · 选择数据源

选择任何结构化或非结构化数据源, 导入到TRS水晶球工作空间。

来源导入

类型: MySQL

主机: 192.168.1.1

DB 名: dbtest

端口: 3306

添加 取消

来源

- sample1.docx
- 测试数据.eml
- address_book.xls
- calllist_3.xml
- sample.webm



数据

Excel

文档

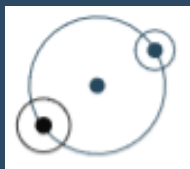
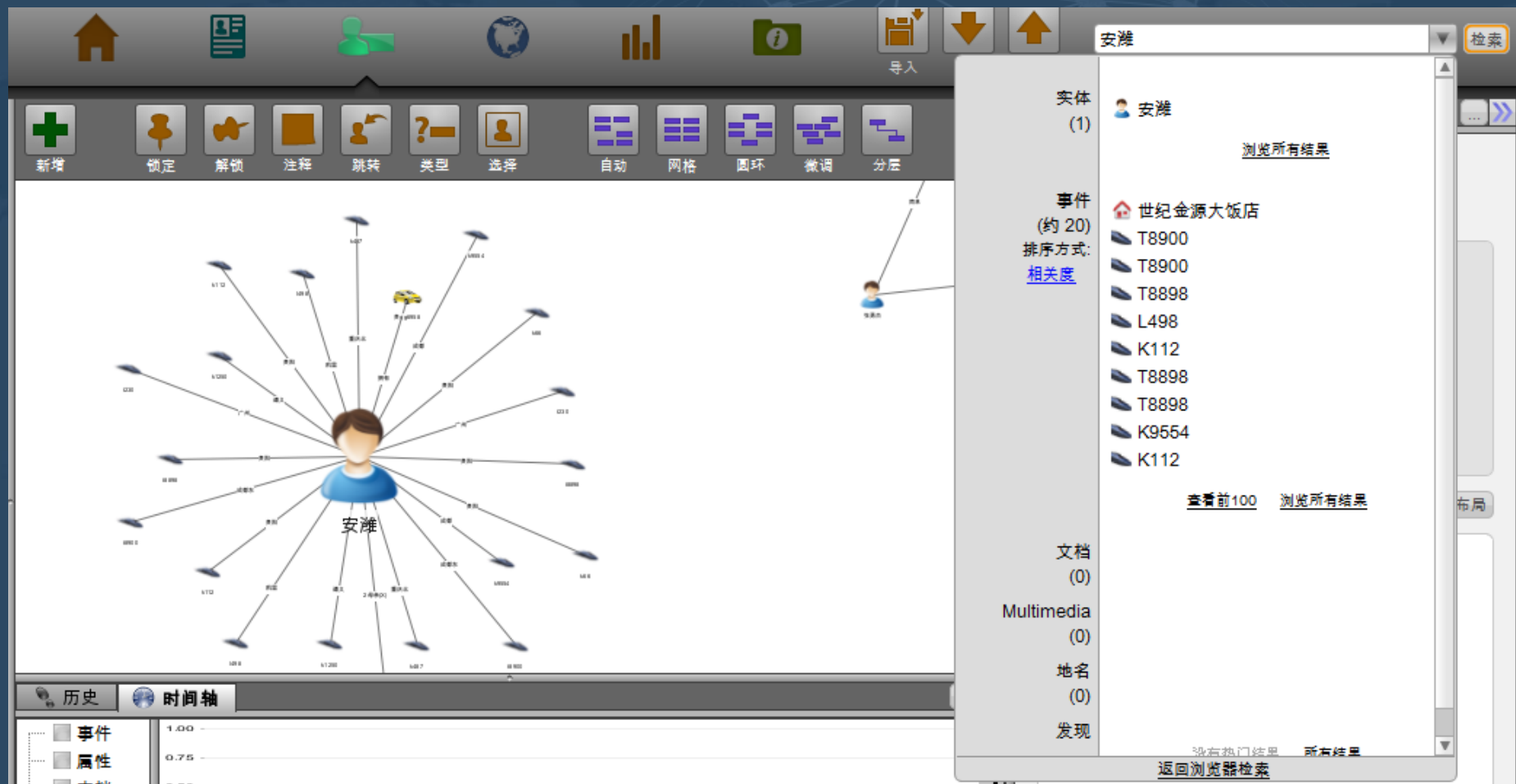
邮件

媒体

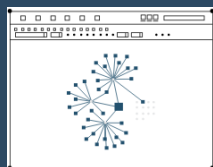
网页

拓尔思 TRS

检索发现



全文检索



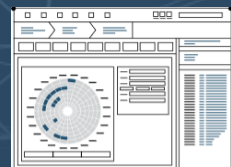
线索发现



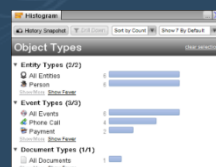
地图分析



关系分析



对象浏览

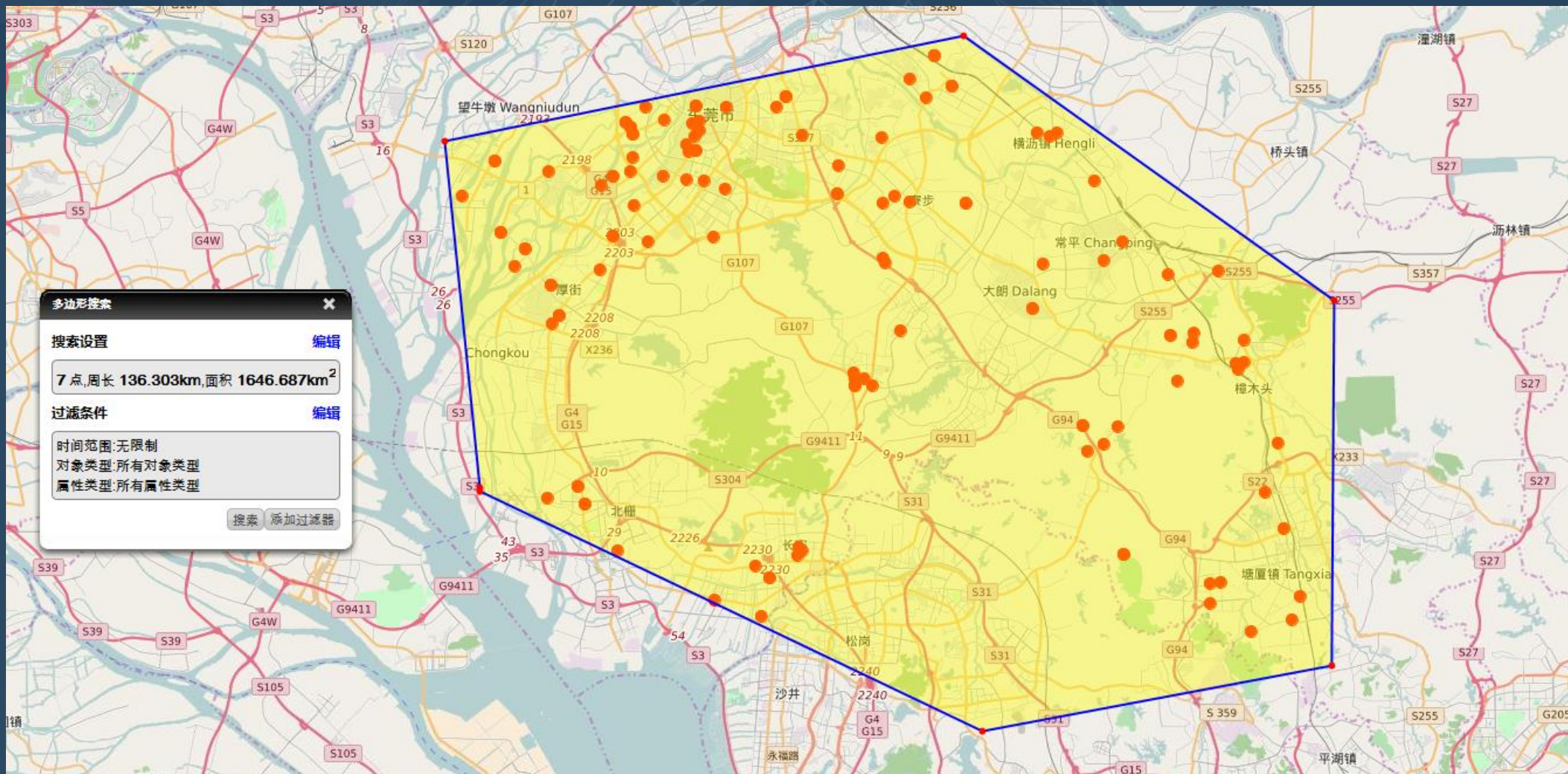


直方图

检索的交互让用户逐步发现未知和隐藏的关联



基于LBS检索，实现信息与位置的统一



检索结果多屏互动实现结果多种展现

关联图 - 重点人员

调查 编辑 首选项 应用 帮助

导入 更新 发布 点这里查询 检索

新增 锁定 解锁 注释 跳转 类型 选择 自动 网格 圆环 微调 分层

历史 时间轴

调查 编辑 首选项 应用 助手 帮助

导入 更新 发布 点这里查询 检索

后退 前进 框选 放大 雷达 多边 画线 画框 箭头 路径 形状 文本 图像

按下ctrl+shift键拖动鼠标测量两点之间的距离

时间轴

浏览器 - 重点人员

调查 编辑 首选项 应用 助手 帮助

导入 更新 发布 点这里查询 检索

李静

后退 向前 文件夹 文档检索 图形 地图 分享 +字体 -字体 射击

李静

更多...

相关 2 0 0

对象信息

类型: 医生
创建人: trs test
创建时间: 2016/07/10 21:20:59 +08:00

谁

摘要 属性 相关的 注释和多媒体 历史

属性

类型 值

调查 编辑 首选项 应用 帮助

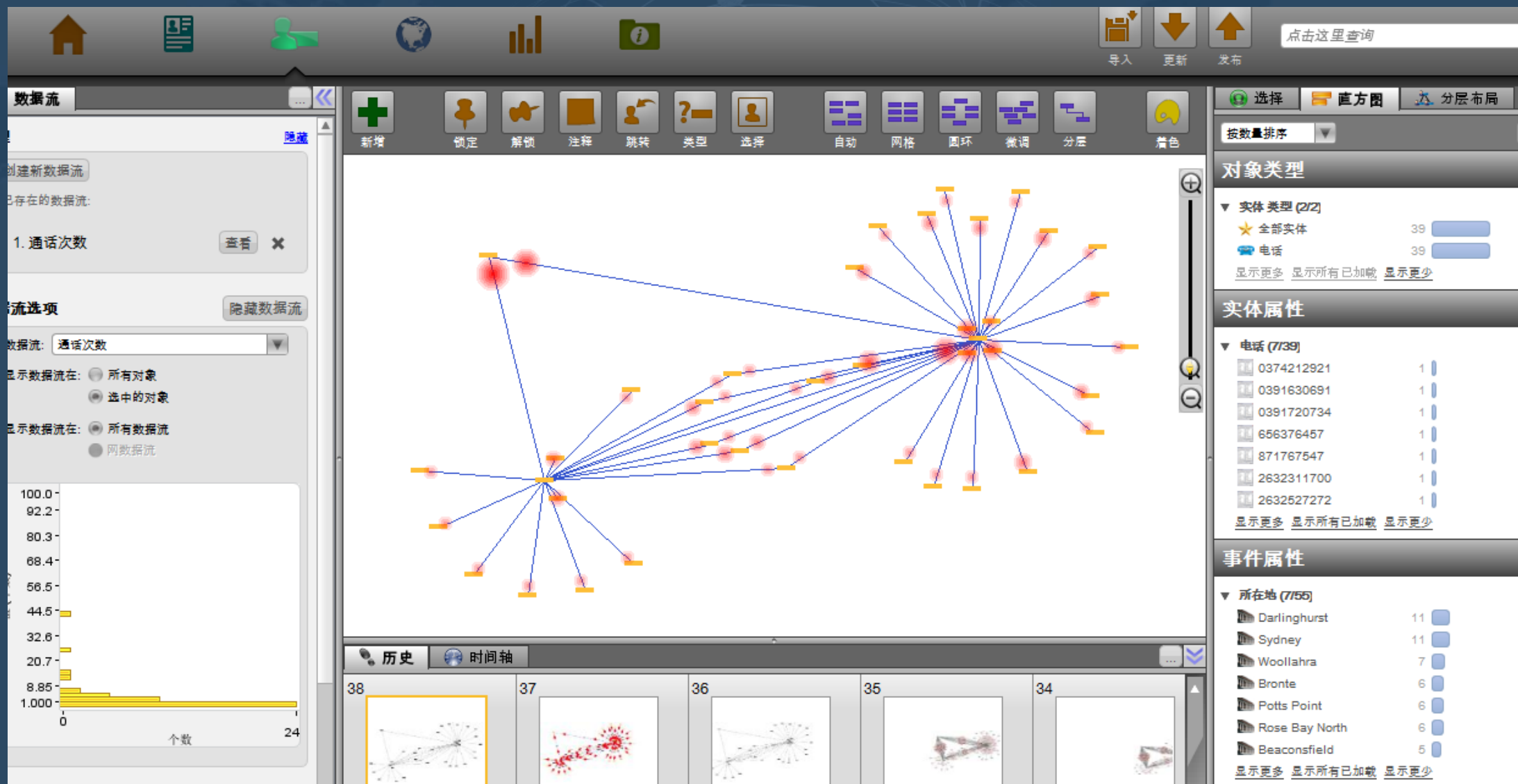
导入 更新 发布 点这里查询 检索

TRS水晶球

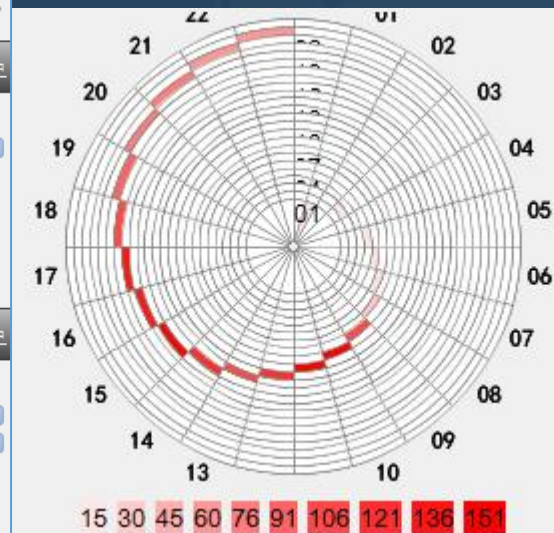
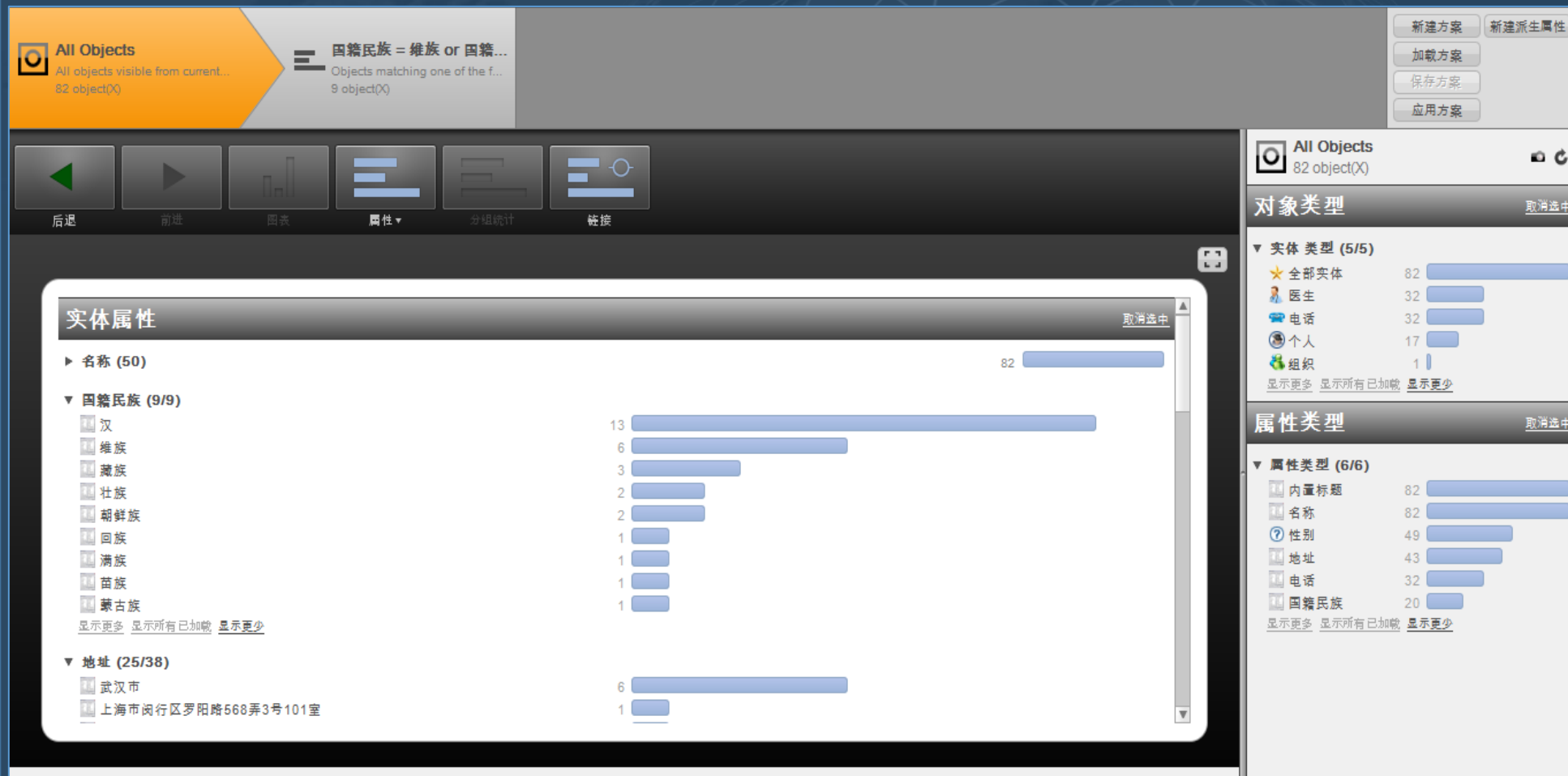
输入检索词 检索

以下选项可通过TRS检索

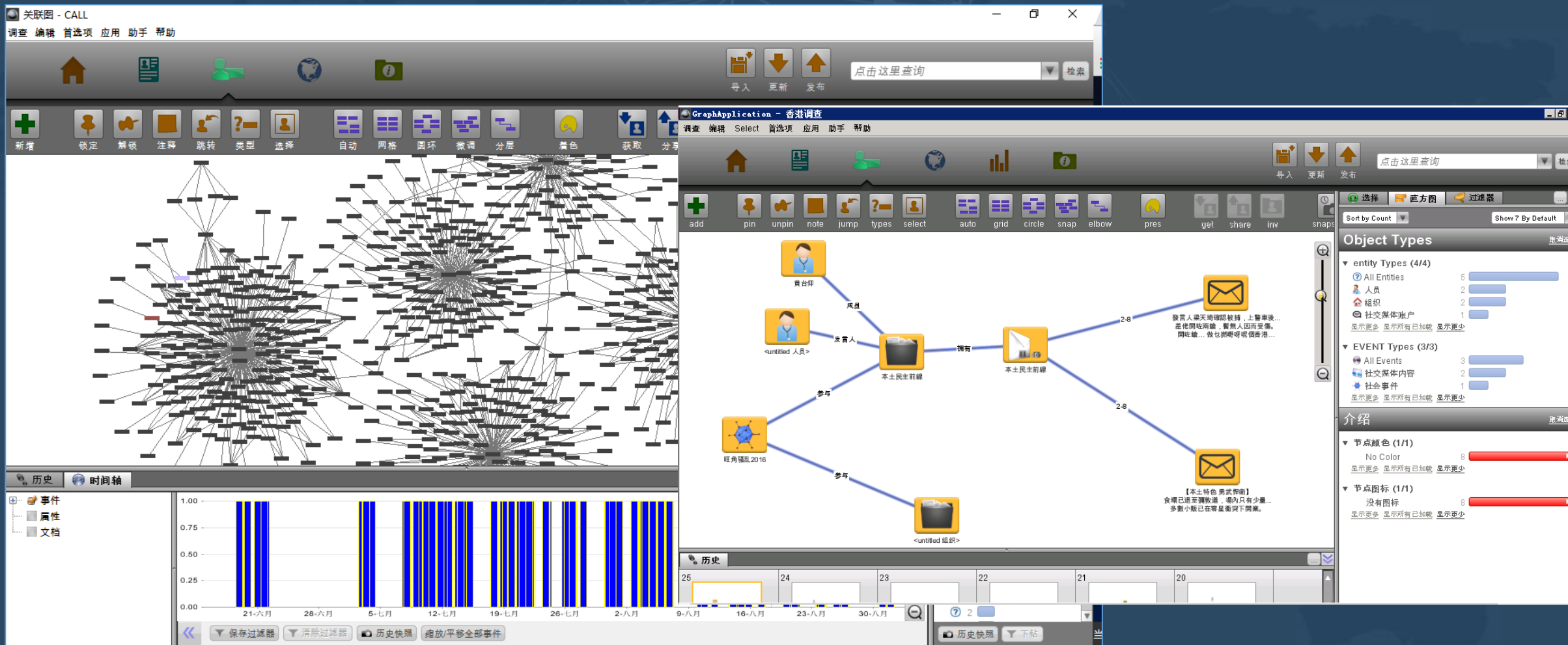
动态多维分析



从宏观到微观



知识图谱



动态模型

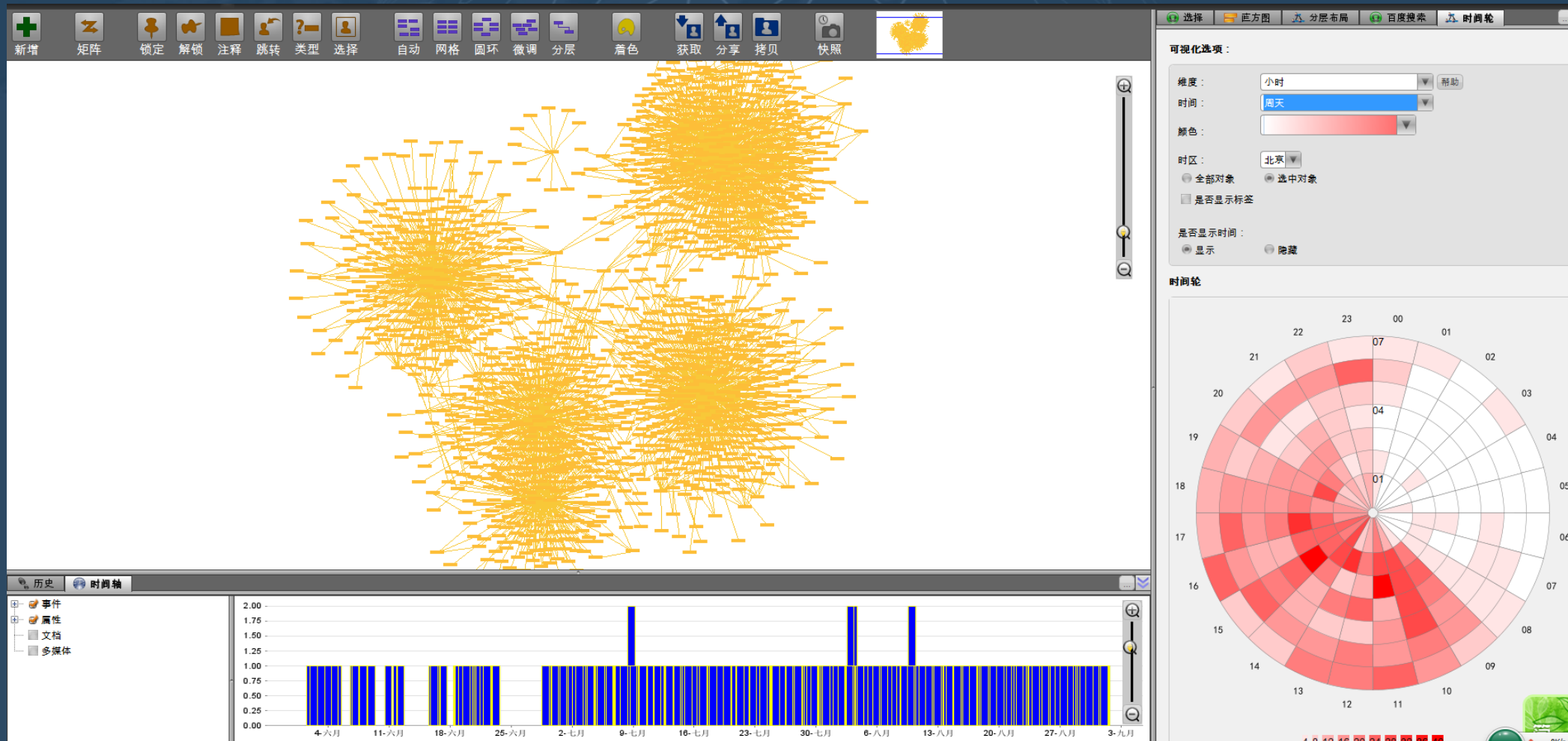


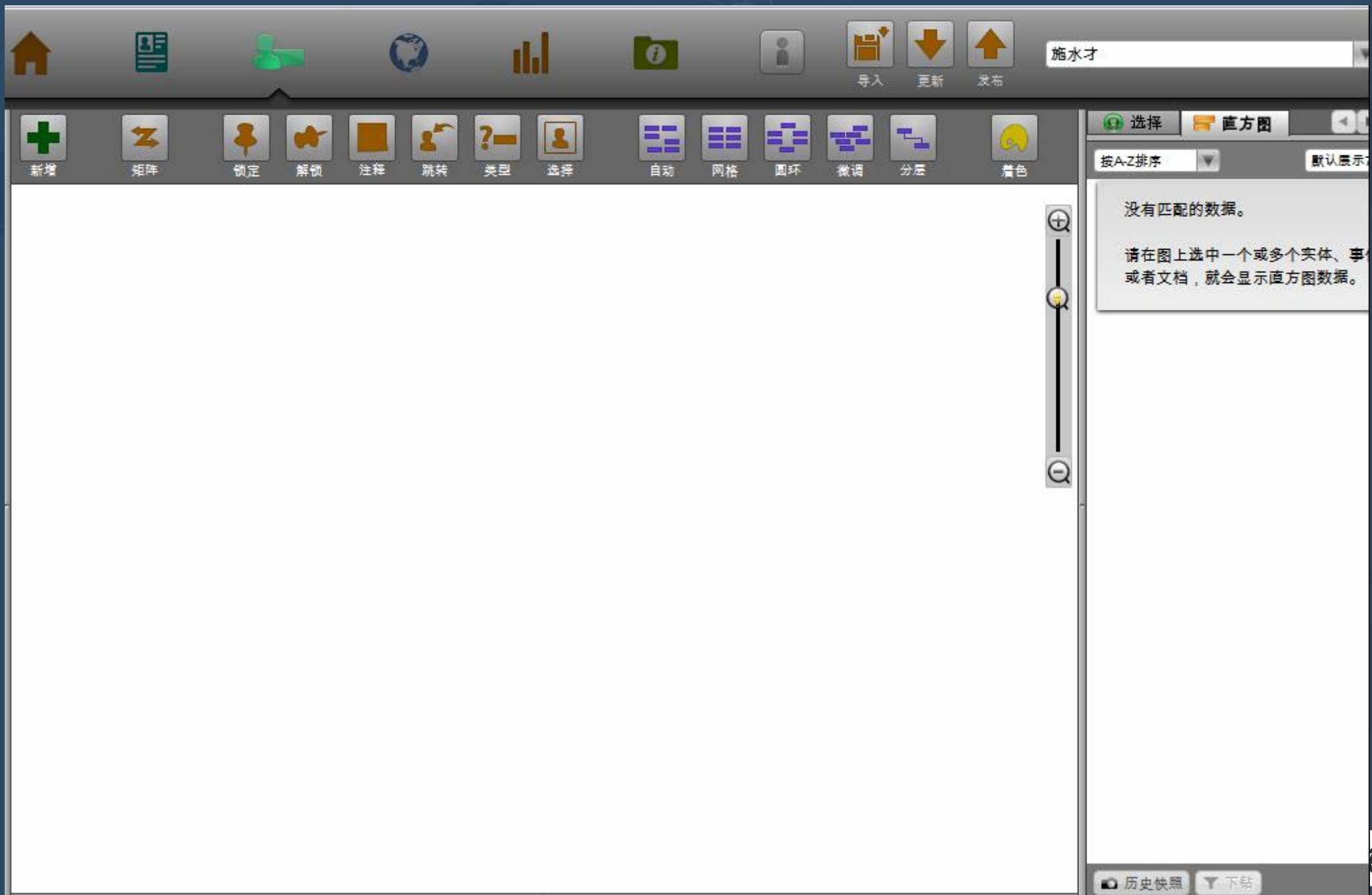
知识编辑



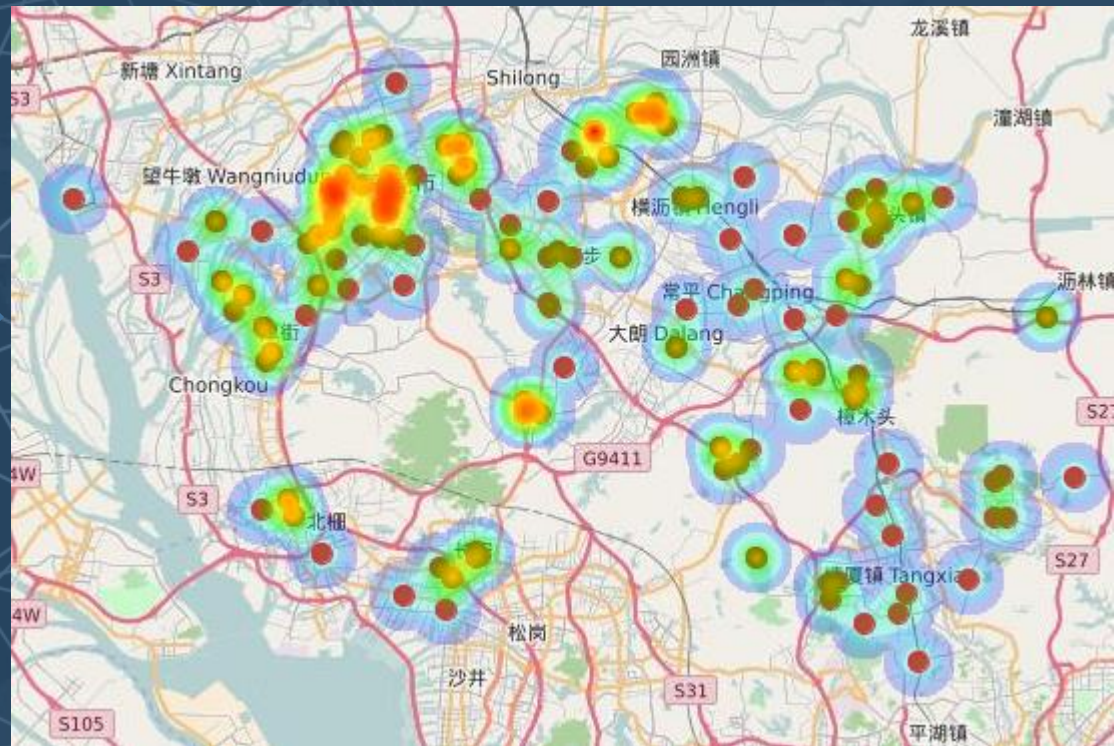
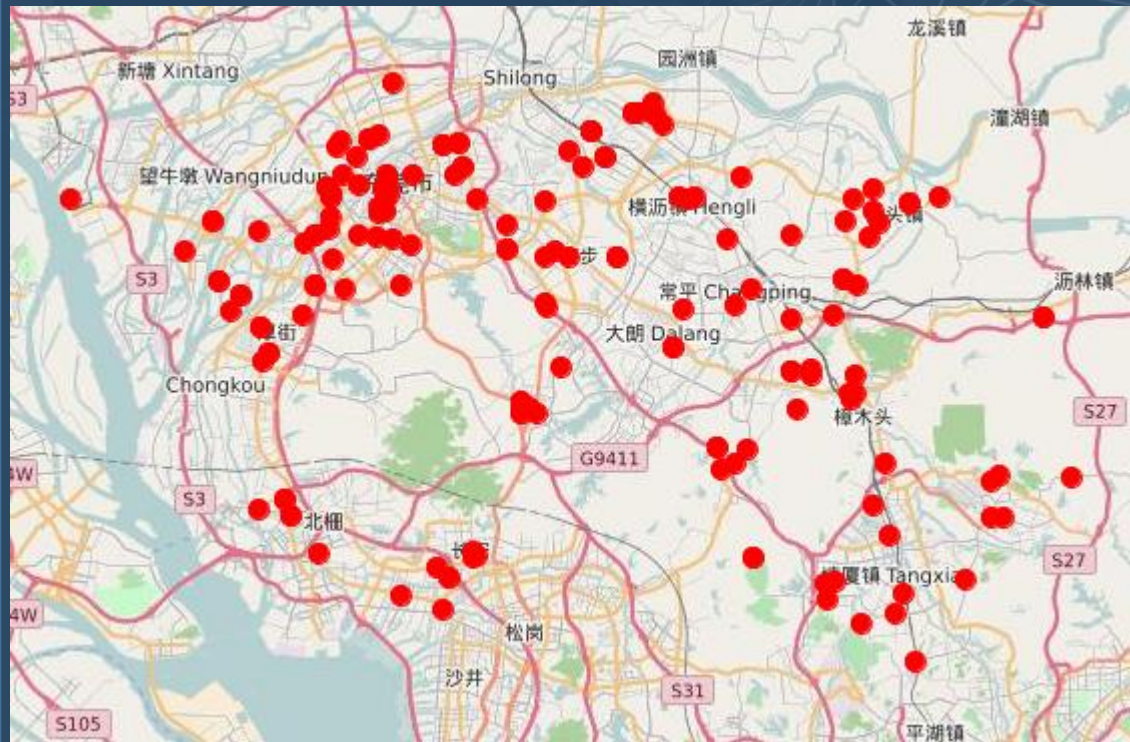
文本标记

大规模知识图谱操作能力

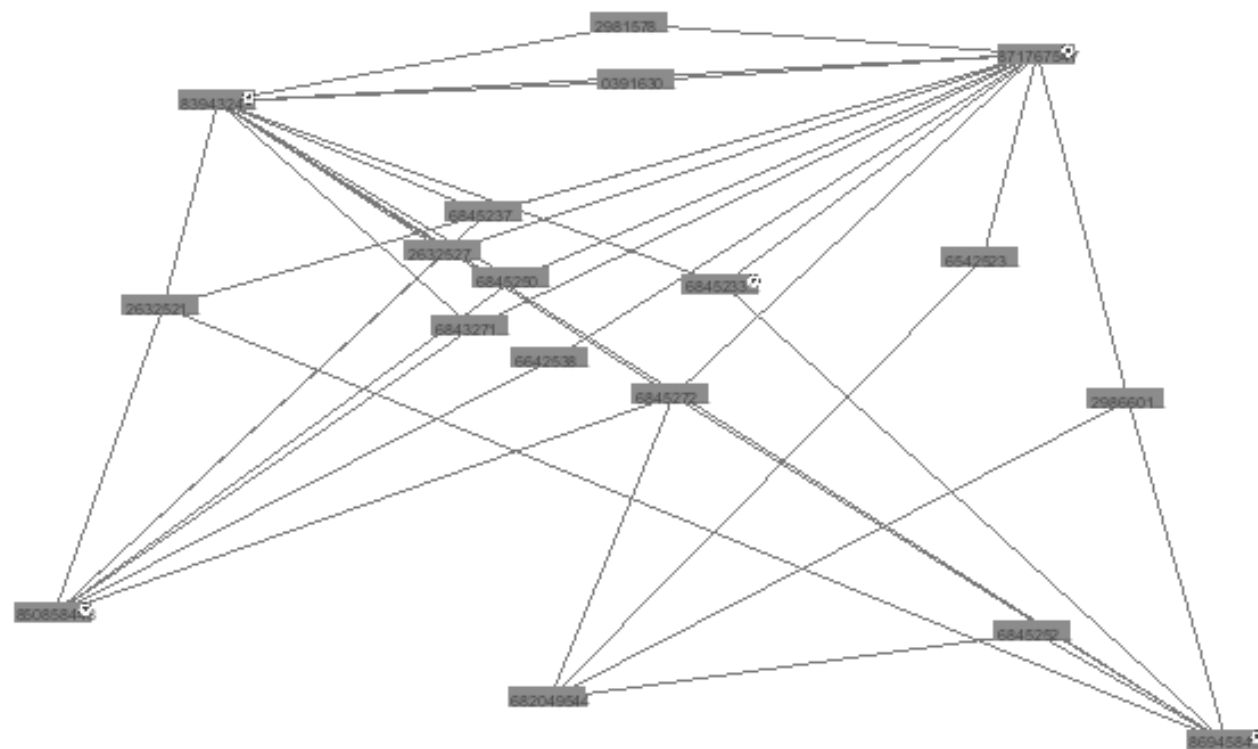




知识图谱的意义在于分析预测



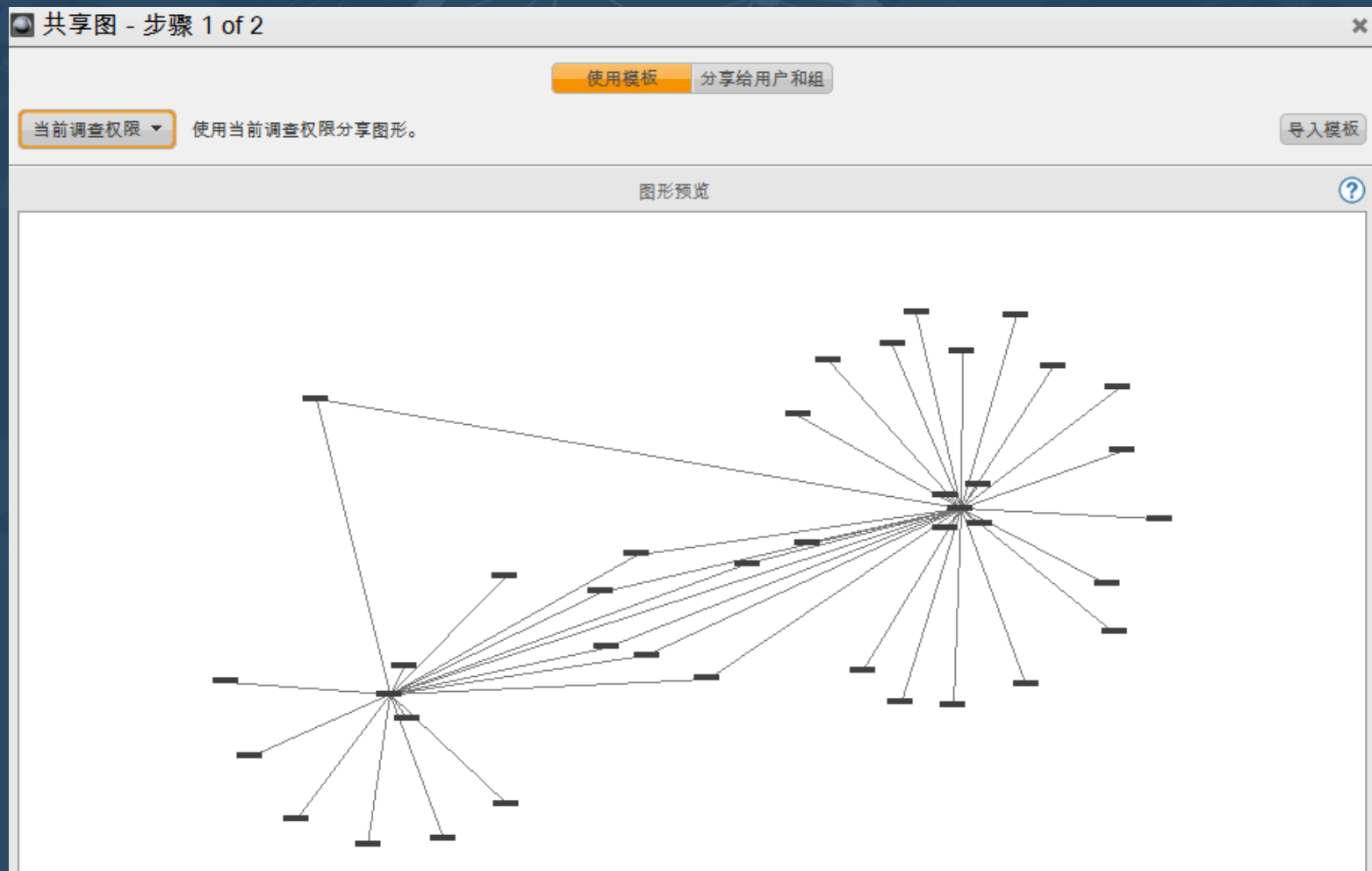
一个使用TRS水晶球知识图谱的实例



▼ 简单 (18/18)

000000001	3
000000004	32
000000048	21
000000047	116
000000020	9
000000022	4
000000070	6
000000030	9
000000021	9
000000034	127
000000034	8
000000016	14
000000036	13
000000021	17
000000032	123
000000030	26
000000034	14
000000007	3

协同工作



知识链接



协同标注



成果分享

TRS 水晶球面向领域



- TRS水晶球是大数据+的落地
- 结合各种数据抽象本体分析
- 融入分析模型洞悉数据内涵
- 通过数据可视化发现隐藏信息
- 结合行业知识面向行业分析师



刘瑞宝

13910671923

北京拓尔思信息技术股份有限公司

Beijing TRS Information Technology Co., Ltd.

股票代码：300229