

拓尔思水晶球

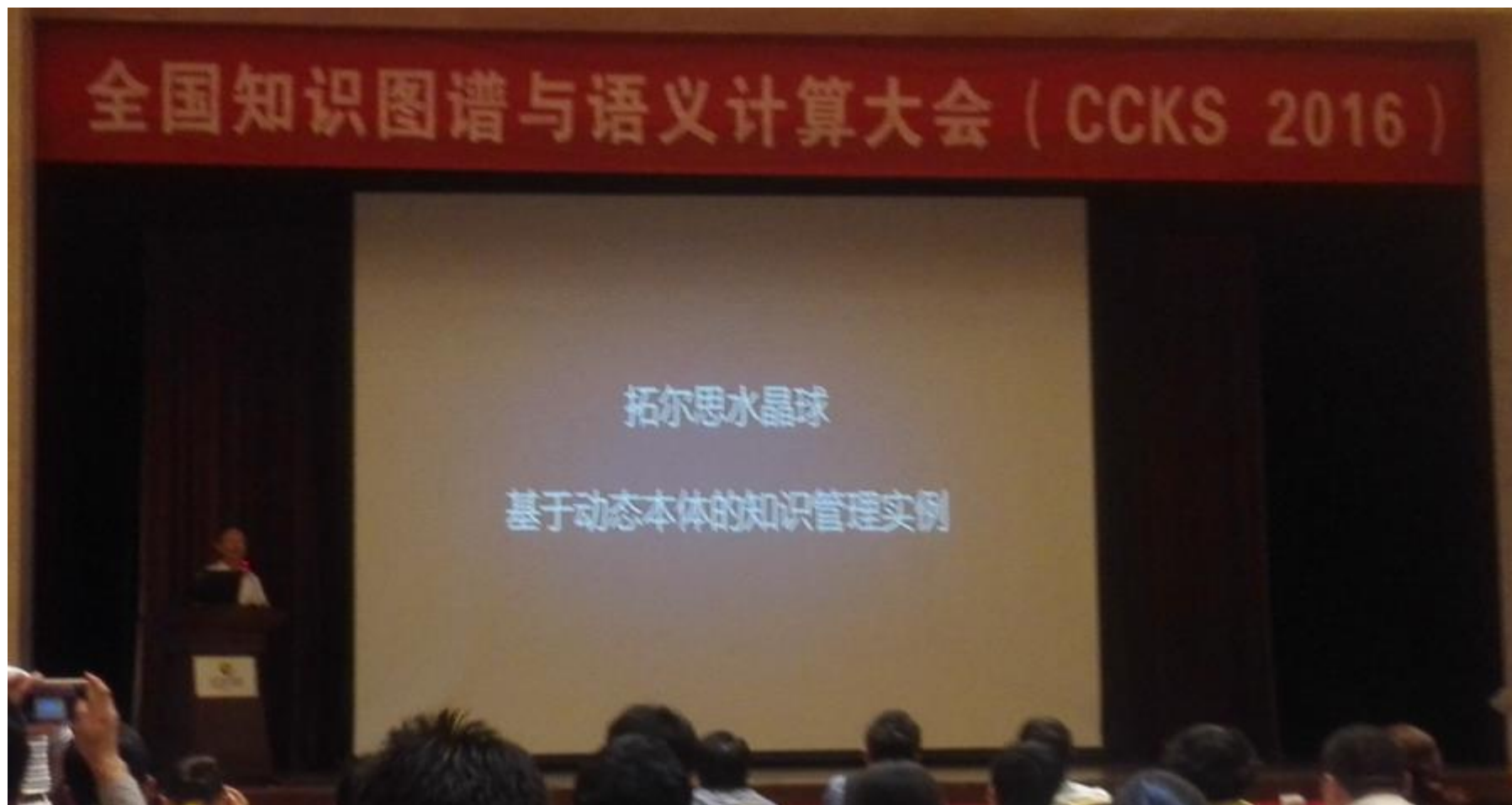
基于动态本体的知识管理实例

刘瑞宝

全国知识图谱与语义计算大会 (CCKS 2016)

拓尔思水晶球

基于动态本体的知识管理实例



- 本体概念
- 知识图谱的作用
- 知识图谱的创建（同鲍捷的所谓源流）

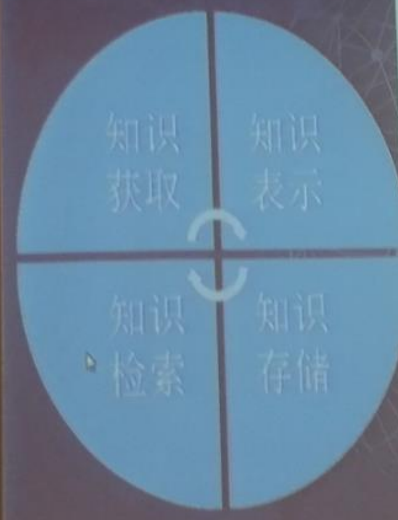
本体是什么

- ① 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。
- ② 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



- 水晶球的六大功能点

当本体论遇上知识图谱

① 本体构建

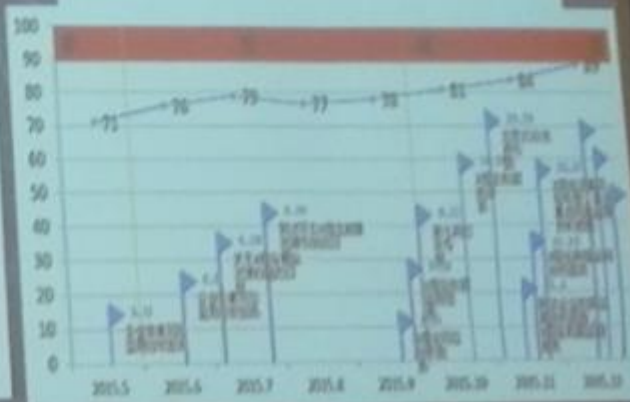
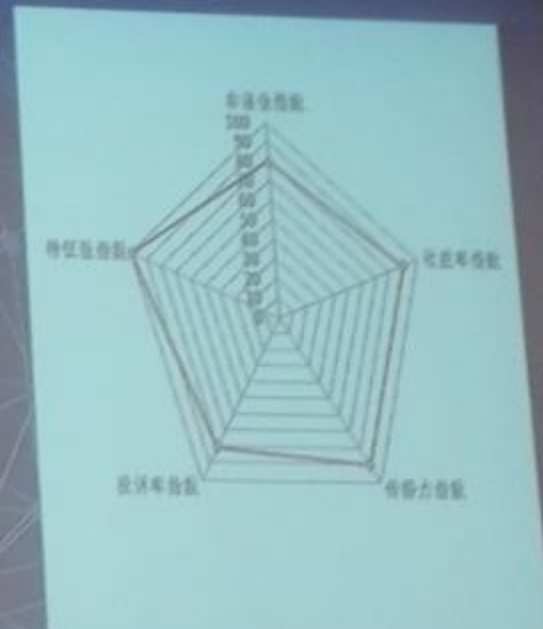
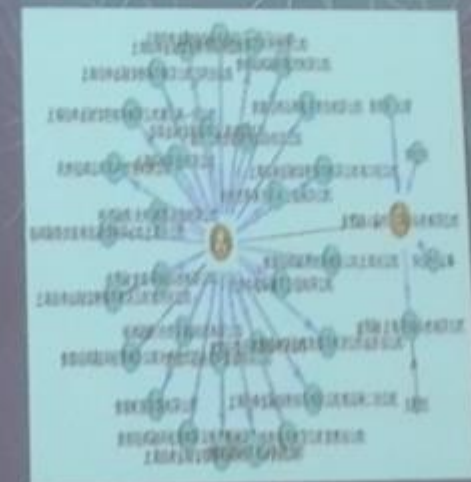
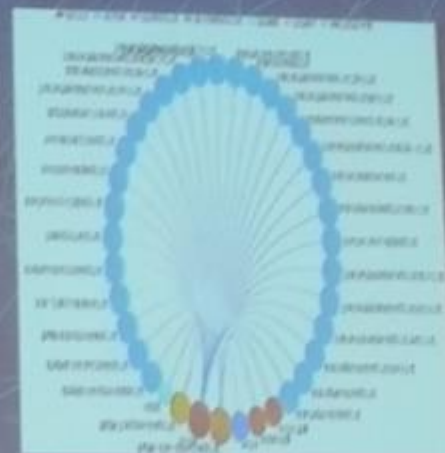
② 实体构建

③ 异构数据融合

④ 动态发现

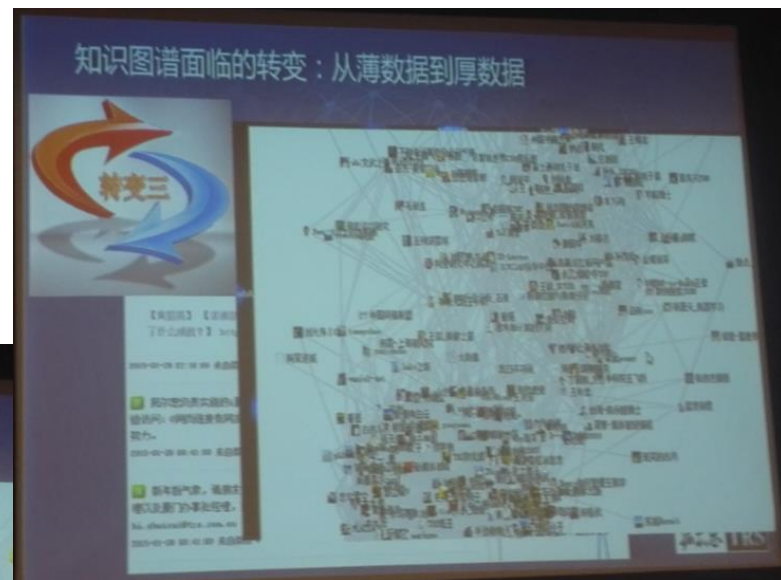
⑤ 关联挖掘

⑥ 预测



- 知识图谱的转变
(一般知识图谱和水晶球的区别?)

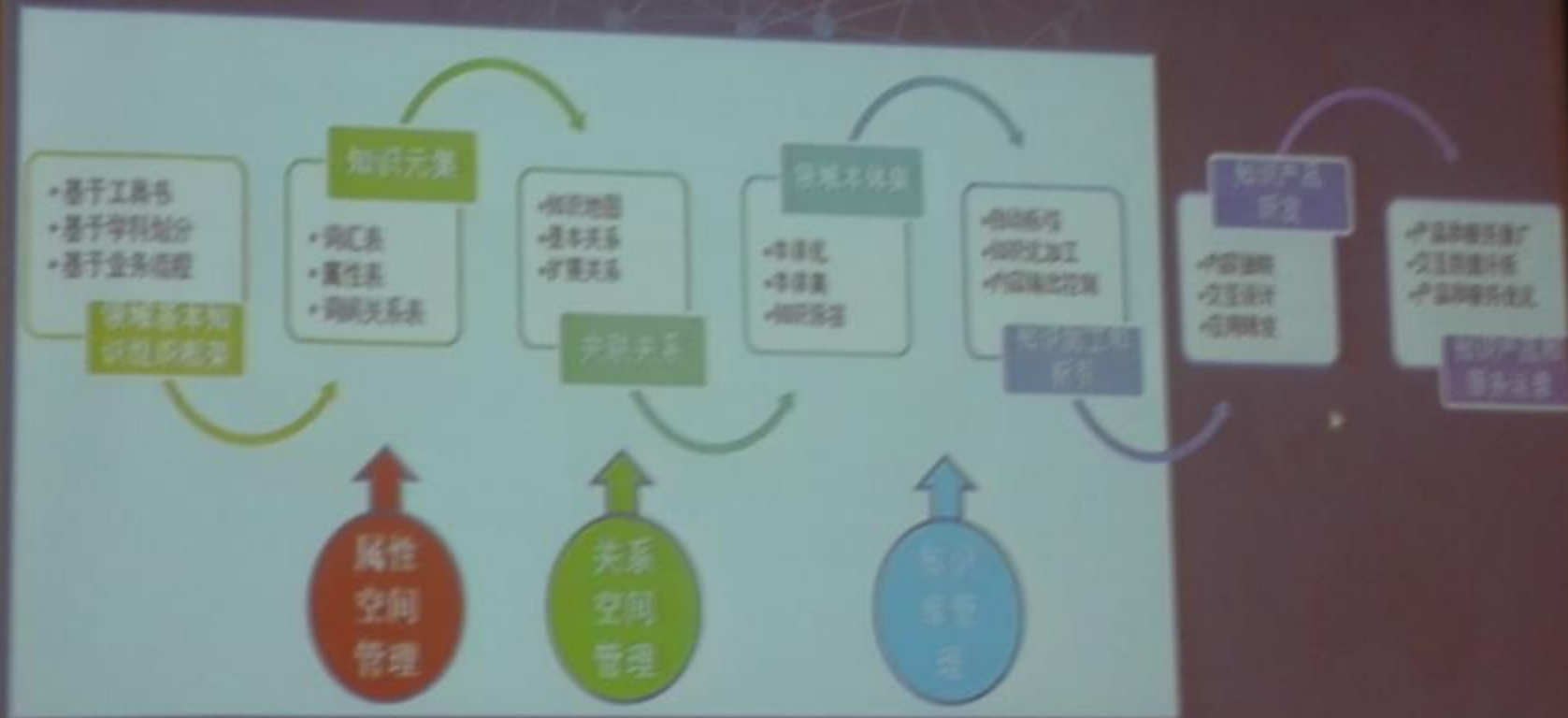
- 从“整合”到“融合”：信息的有机重构
- 从情报计算到计算情报：挖掘的信息由知道自己不知道到不知道自己不知道
- 从薄数据到厚数据



一、建立知识库建设总体框架



二、梳理知识库的构建流程



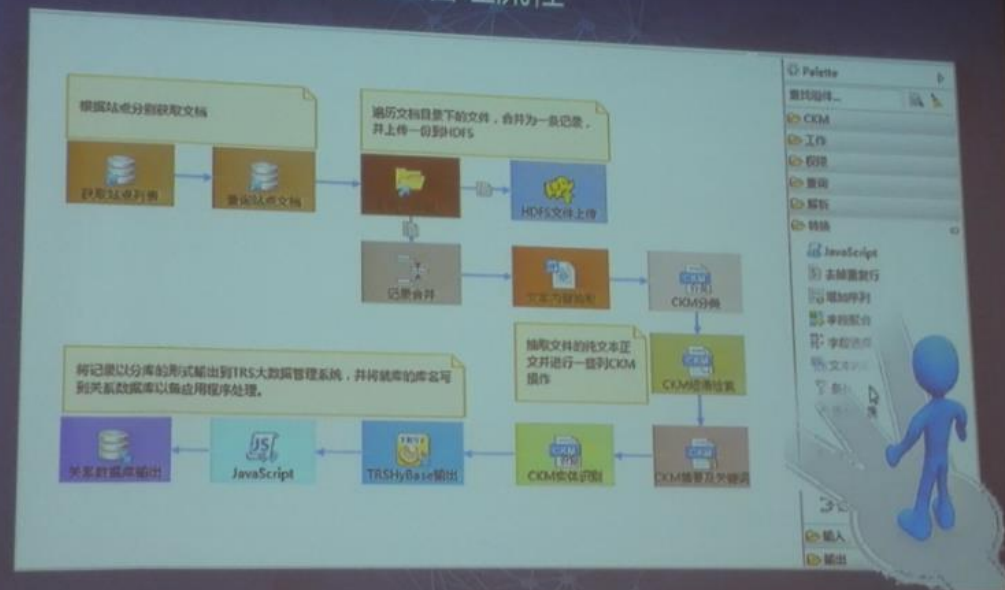
三、获取数据——TRS Adapter数据集成平台



- 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

拓尔思 TRS

TRS Adapter数据配置管理流程



拓尔思 TRS

产品介绍：TRS Adapter：数据适配集成平台


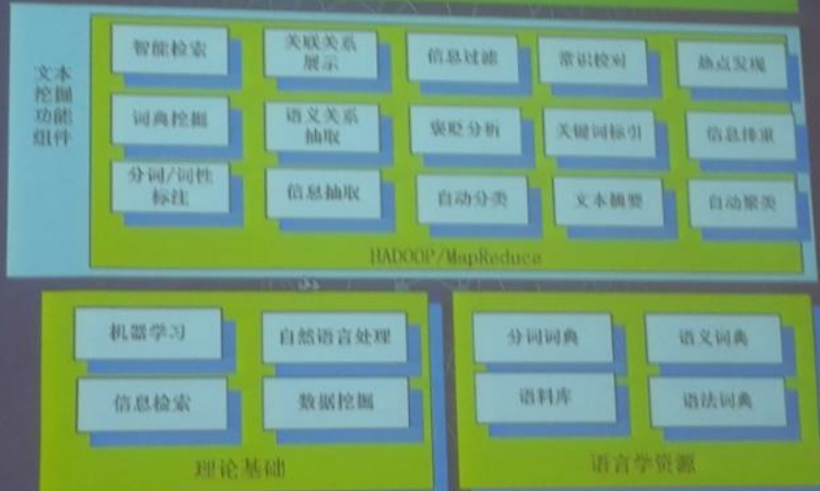
从非结构化文本中发掘有价值的信息

项目	性能指标
文本分类	<ul style="list-style-type: none"> 1 自动分类准确率: 在几十个类别之内, 经过分类体系精心训练和语料库训练达到 85% 以上, 分类准确率可以达到 90% 以上。 2 单用户 10 万级自动分类准确率: 30 级用户可产生 200 万级分类。
相似性检索	<ul style="list-style-type: none"> 1 检索量: 10 万级文档, 单用户/每小时每秒执行检索量 25 次, 300 万级文档, 单用户/每小时每秒执行检索量 5-6 次。 2 检索量: 1000 万级文档, 单用户/每小时每秒执行检索量 40 次, 16 万级时, 每小时执行检索量 280 次左右。
自动摘要	<ul style="list-style-type: none"> 1 单用户生成摘要量: 31 篇/秒, 50 个非自动摘要生成量: 266 篇/秒。 2 支持大文档的摘要: 每篇长度大于 1M 字节的文本摘要。
文本信息抽取	<ul style="list-style-type: none"> 1 支持新闻 (人名地名机构名) 抽取率为 80% 以上。 2 单机抽取速度: <ul style="list-style-type: none"> 1 实体抽取 500K/秒, 关系抽取 400K/秒; 2 新闻信息抽取 2.5M/秒, 标题抽取 100K/秒; 3 新闻抽取 200K/秒, 政治财经科技抽取 350K/秒; 4 基金信息抽取 500K/秒。 3 提取抽取新闻摘要: 包括: 人名地名机构名, 事件名, 标题信息, 数字信息, 单用户/每小时信息抽取量 15 万条, 16 个非实时/每小时。 4 单用户/每小时每秒执行摘要量 162.5 万条, 10 个非实时/每小时, 每小时执行摘要抽取量为 320.8 万。
排布检索	<ul style="list-style-type: none"> 1 单用户/每小时每秒执行全文检索量 909.1 次, 10 级 500 个非实时, 每小时执行全文检索吞吐量达 236.7 万。
交叉知识检索	<ul style="list-style-type: none"> 1 对一般知识问答和文本, 可以达到 5000 条/秒。 2 单用户/每小时每秒执行流量 1 万条, 10 级 8 个非实时, 每小时抽取流量达 8 万条, (得获来自美国文本数据库为 5000 篇, 每篇文章平均长度为 1000)。
文本聚类	<ul style="list-style-type: none"> 1 信息聚类检索率达到 75% 以上, 满足大多数场合的应用需求。
(政治) 关键词识别	<ul style="list-style-type: none"> 1 通识型 20 万级非实时 5000 文本以上, (训练: 94.2, 94.2, 523M 内存)。
文本相似计算	<ul style="list-style-type: none"> 1 平均速度为每秒 40 条以上, (训练: 94.15, 523M 内存)。
高维矩阵分解	<ul style="list-style-type: none"> 1 200 个非实时/每小时每秒抽取速度为每小时每秒抽取文本 20000 万条, 环境: IBM X3100 服务器, 360GB/秒 CPU 主频 2.40G, 内存 4G, 操作系统 Windows Server2003。

華爾思 IRS

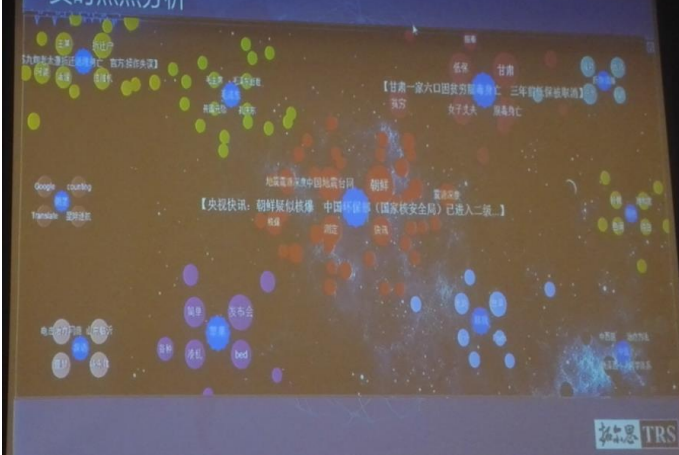
文本挖掘应用

- 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
- ...

 TRS

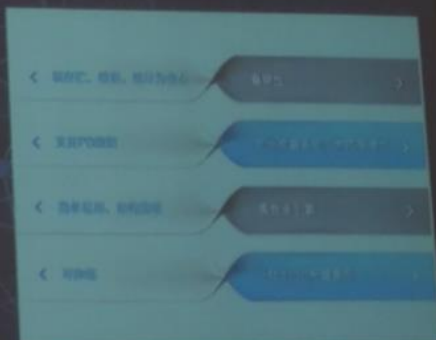
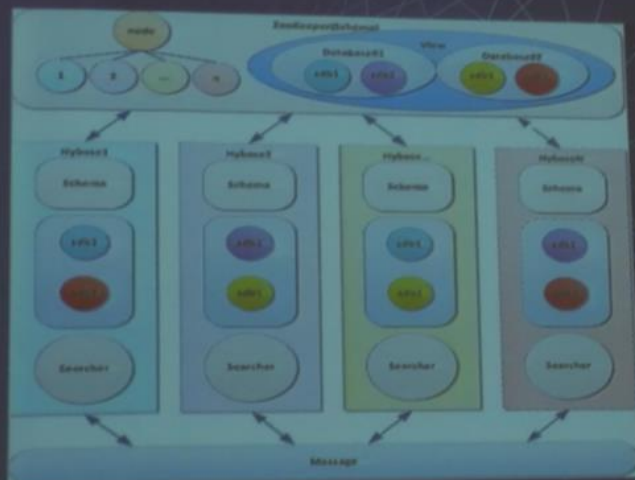
产品介绍：TRS CKM：智能信息处理，从非结构化文本挖掘信息

实时热点分析

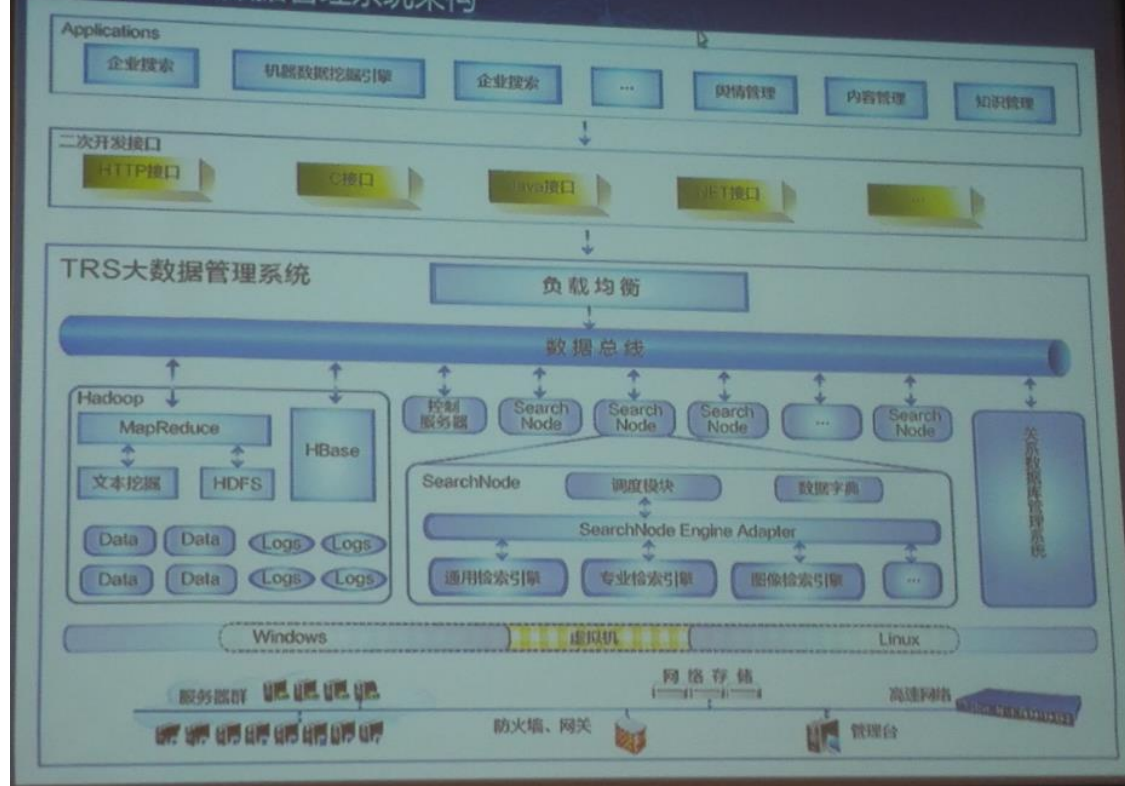


五、海量数据管理——TRS海贝

- 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持。

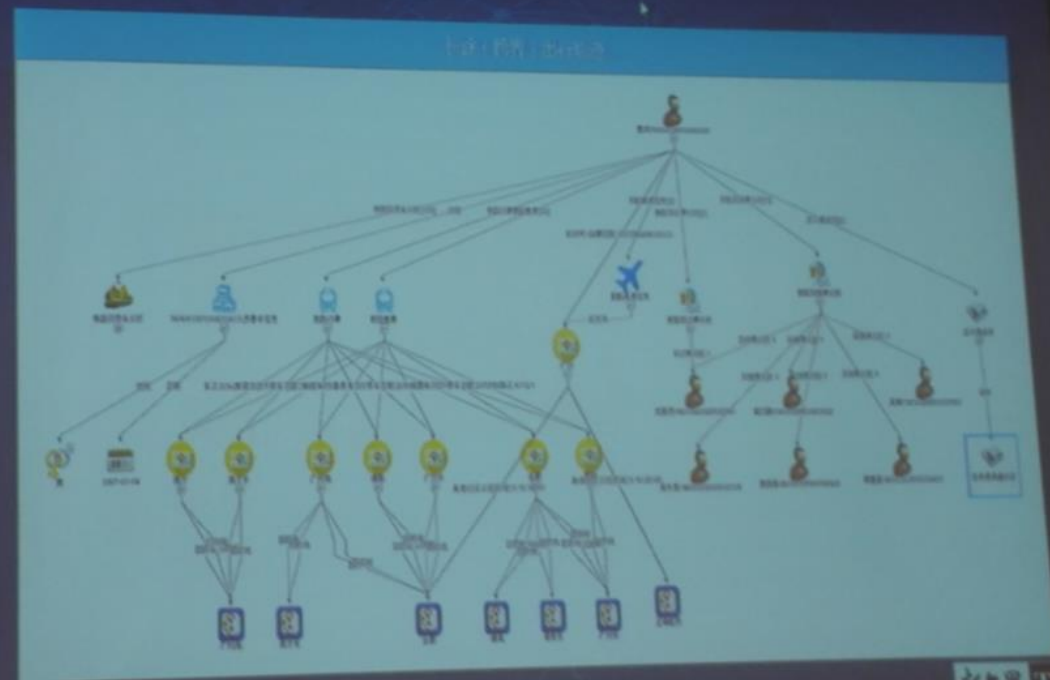


TRS海贝大数据管理系统架构



产品介绍：TRS 海贝：一款基于弹性扩展架构的海量数据存储于检索系统，定位为企业级 NoSQL，企业级检索平台和大会数据管理集成平台

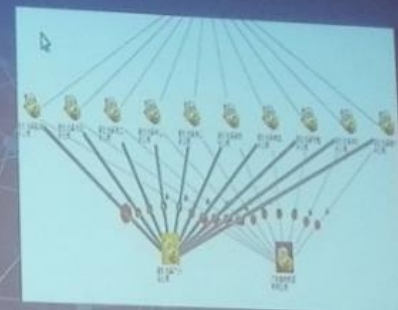
六、知识图谱可视化综合挖掘



可视化让结果更直观



社交媒体可视化



数据流可视化

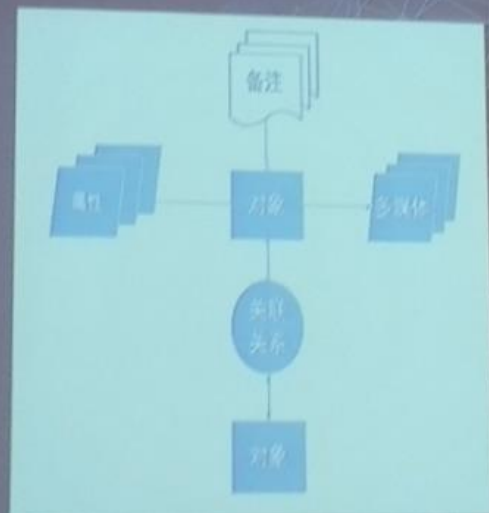


访问行为可视化



可视化大屏指挥

TRS水晶球动态本体的知识管理

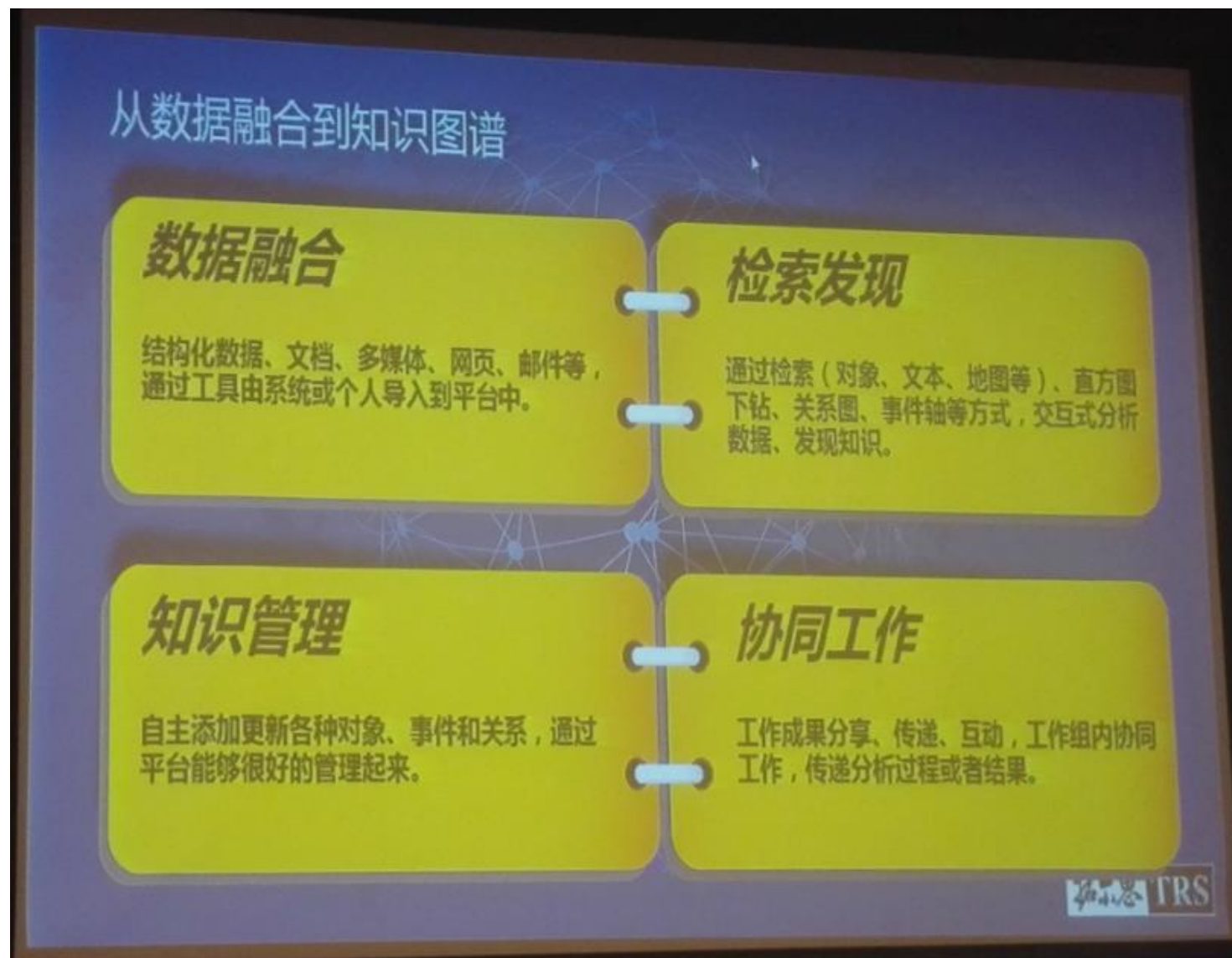


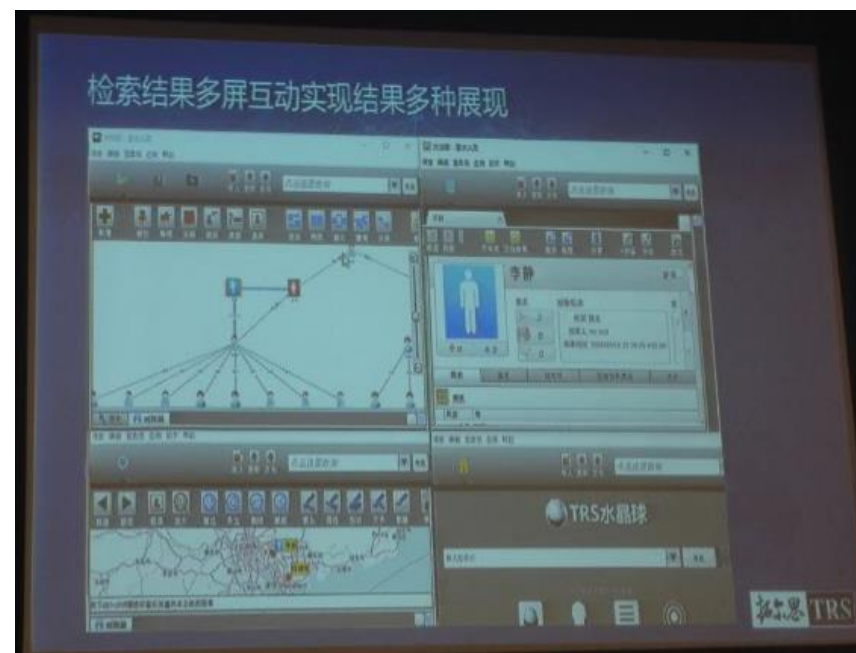
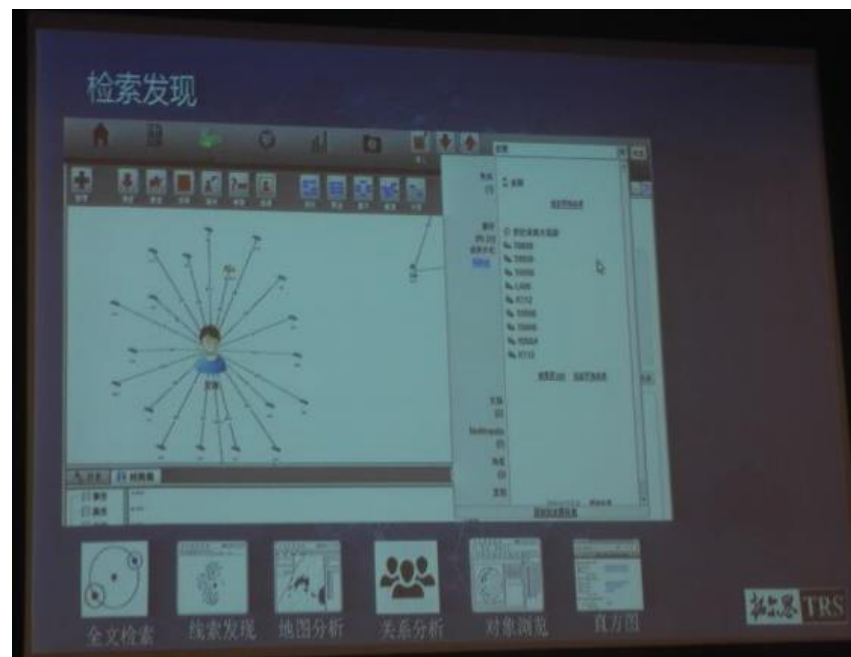
- 对象—代表世界上的事物。
 - 对象包括：实体、事件、文档、多媒体
- 属性—对象的属性
- 关系—对象之间的联系
- 备注—对象相关的文字备注
- 多媒体—对象的多媒体附件
- 通过“实体 - 关系 - 实体”三元组描述物理世界中的概念及其相互关系

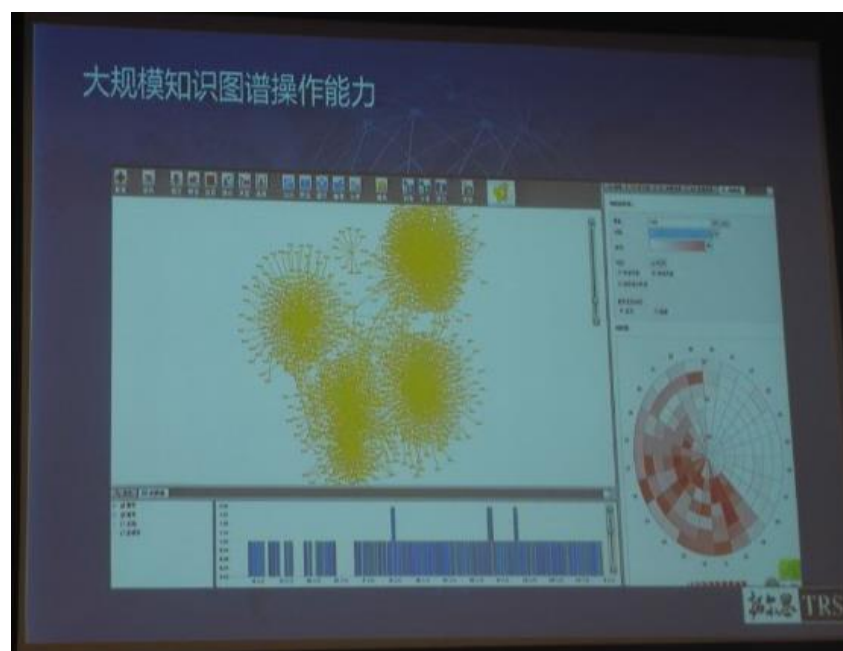
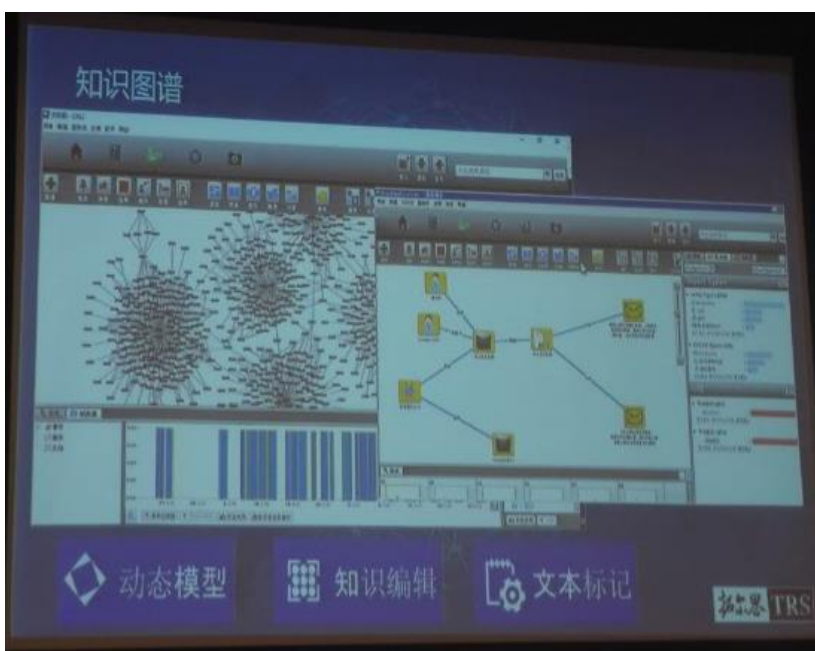
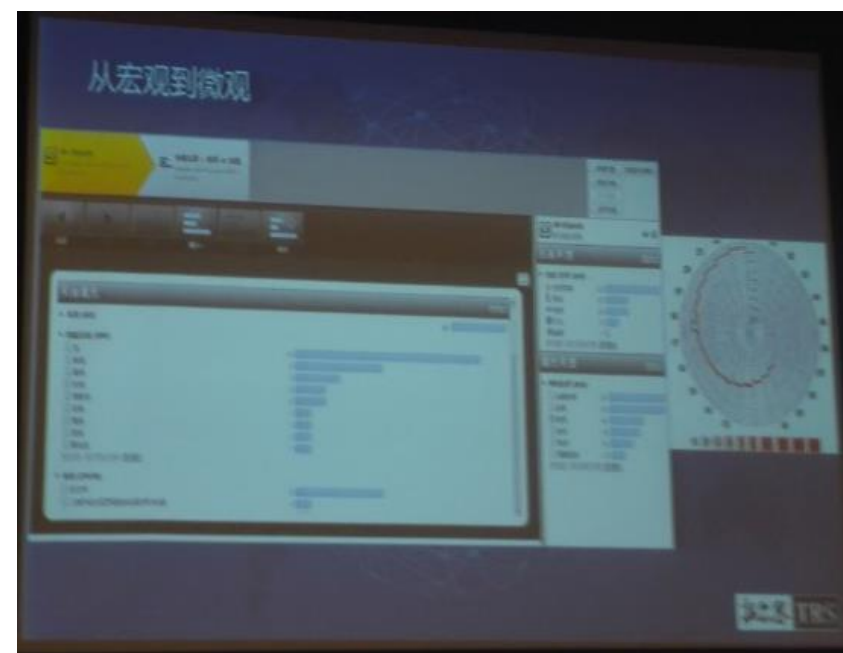
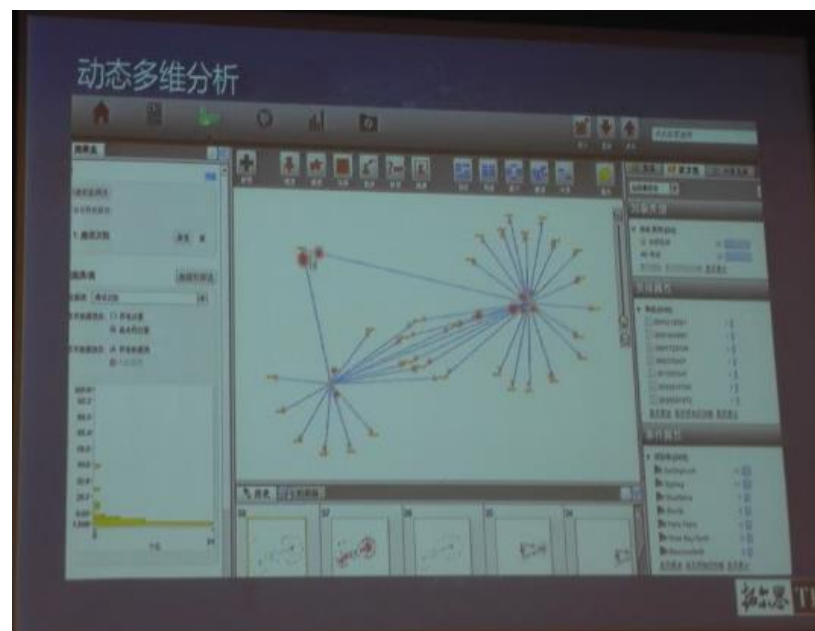
面向领域的分析师工具：本体+数据+扩展



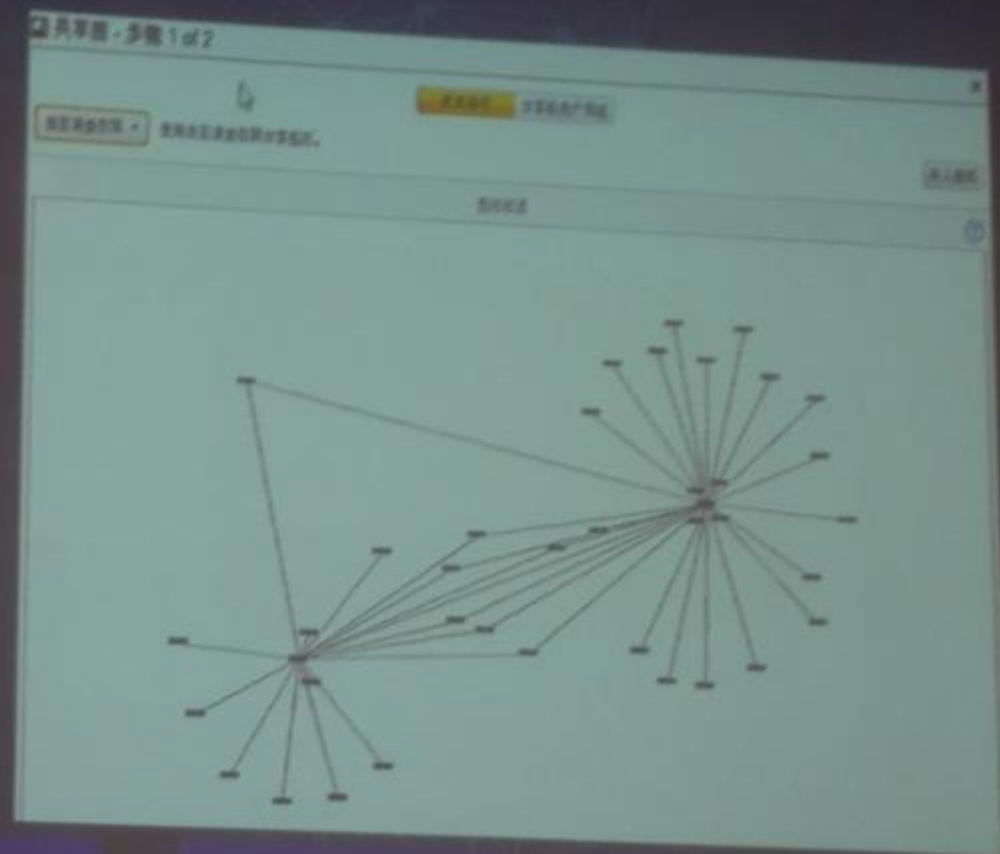
- 水晶球的四大亮点







协同工作



知识链接



协同标注



成果分享

拓尔思 TRS

TRS 水晶球面向领域



线下交流

- 水晶球目前是alpha版本，还没有正式公开推出
- 开发2年，目前是40人团队，开发高峰期人数更多
- 水晶球支持单机版和分布式
- 和palantir不好比对，中文支持，战法更丰富
- G20有应用，已有一些合同

小结

- 更多是产品宣传，没有深度解析
- 从demo看，从界面到功能完全是模仿palantir
- 开发成本比较大，超40人2年
- 此类产品是当前的一个热点，此次工业界论坛还有一家海云数据也推出同类可视化产品
- TRS有较好的整体生态，包括此次介绍的Adapter、CKM、Hybase，据介绍和水晶球有深度整合
- 水晶球的功能亮点：数据融合、检索发现、知识管理、协同工作