精益知识图谱方法论

文因互联 鲍捷

知识图谱技术的组成与源流

- 知识图谱是一个交叉的工程
- 四大源流：
  - 知识提取
  - 知识表现
  - 知识存储
  - 知识检索

知识提取
（自然语言处理，机器学习）

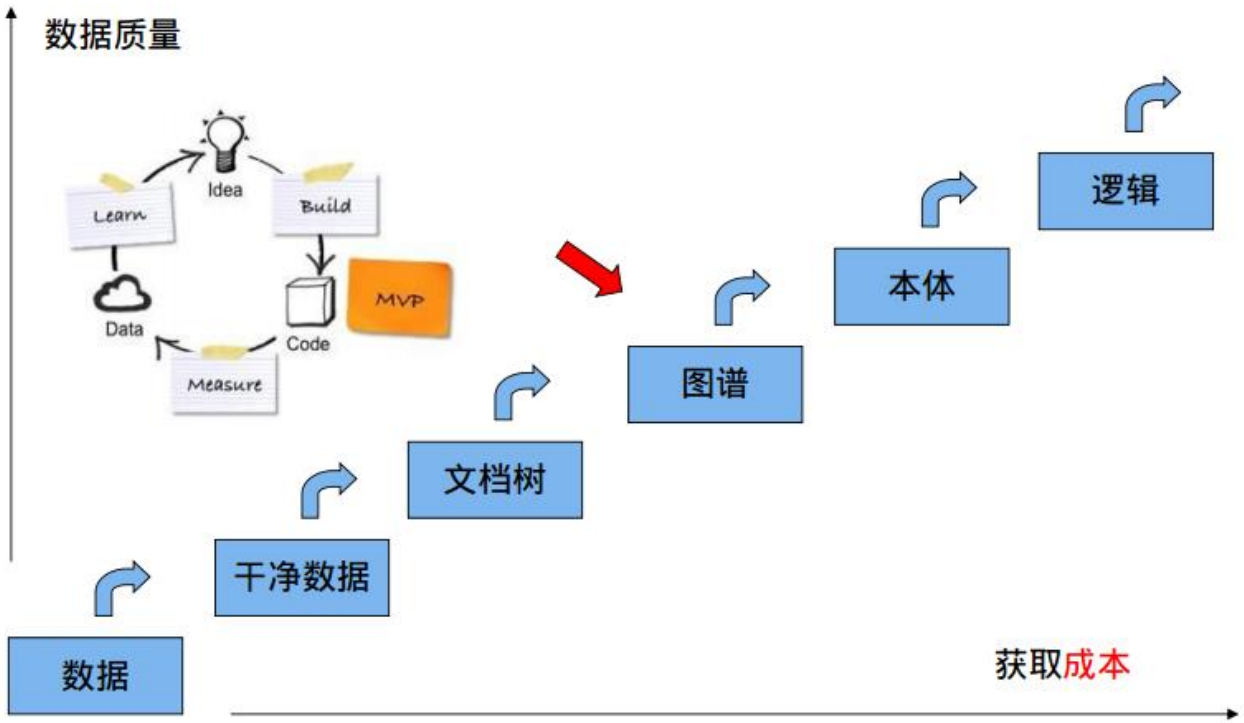
知识存储
（数据库）

知识表现
（逻辑，语义网）

知识检索
（信息检索，人机交互）

数据 vs 知识

- 知识和数据没有非白即黑的界线
- 知识是一种高质量的数据，能够产生新的数据的数据



Twine的尸检报告

- 一次只做一件事，革新的步子不要太大。
  - 开发分布式语义数据库，Web级别的语义搜索，对于一个40个人的小公司，战线太长了。
- 不要坚持和W3C标准兼容。
- 图数据库是未来，但不要指望一步到位，要演化 。
- 产品功能要集中，不要挑战用户稀缺的注意力和理解力。
- 产品和市场要循序渐进，周期要切短。

知识提取

- 知识提取是要解决结构化数据生成的问题。广义上讲，知识提取是数据质量提升中的一环，各种提升数据质量的方法，都可以视为某种知识提取。
- 知识提取的方法
  - 正则
  - 结构提取
  - 实体提取和链接
  - 关系抽取

知识表现

- 知识表示（Knowledge Representation，KR，也译为知识表现）是如何将结构化数据组织，以便于机器处理和人的理解的方法。
- 原则：数据优先，逻辑靠边，我们在学习知识表现方法的时候，要始终牢记知识的可读性、可维护性要远远比它的表达力、计算速度重要
- 知识表现语言：RDF、OWL（ XML、JSON、YAML算不算表现语言），不同于知识交换语言（什么是知识交换，传输吗？）
- 知识推理是成本极高的

知识表现

- 知识表现为数据结构，最常见的是“图”或者“树”
- 图的两个流派：一个是RDF图，一个是属性图（Property Graph）
  - RDF图是W3C的官方标准，科学顶层设计出来的，但是最终市场表现平平
  - 属性图是工程实战中总结出来的，最终得到了市场的认可
  - JSON-LD是RDF的JSON语法，其中LD代表Linked Data。它要解决的是RDF没有好的Web兼容语法问题；体现了两个对人友好的特性：可读性和模块化

知识存储

- 当我们经过知识提取得到了结构化的数据，并选择了适当的知识表现语法后，下一步就是如何持久化存储这些数据
- 知识存储解决如何管理大量的结构化数据。
- 大多数情况下，我们只需要几G、几十G、几百G规模的数据，甚至更少，纯内存处理都可以。不需要去想“大数据”问题
- 现代的关系数据库可能可以解决大多数需要知识图谱的场合。某些特殊场合，我们需要图数据库

知识存储

- 没有一个足够好的数据库解决所有问题
- 大多数图的存储可以归结为为关系型数据库存储；如果只是短层（最多3层）关系查询，不用图数据库。关系数据库+JSON是小规模知识图谱存储的较好选择
- 三种选择
  - 带JSON扩展的关系数据库（PostgreSQL）
  - 图数据库（Neo4j, OrientDB）
  - RDF数据库/其他选择

知识检索

- 知识检索是一个人机交换的过程，人和机器交换知识的过程
- 数据摩擦力
- 元数据促进了知识检索。元数据减少数据流动摩擦，加快数据流动速度的手段
  - 减少人机界面的摩擦：facebook opengraph、faceted browser
  - 机器与机器的摩擦：二维码

小结

- 知识图谱是一系列结构化数据的处理方法，它涉及知识的提取、表示、存储、检索等诸多技术。知识提取是要解决结构化数据生成的问题；知识表示是如何将结构化数据组织，以便于机器处理和人的理解的方法；知识存储解决如何管理大量的结构化数据；知识检索是一个人和机器交换知识的过程
- 从渊源上讲，知识图谱是知识表示与推理、数据库、信息检索、自然语言处理等多种技术发展的融合。
- 实战中的知识图谱，需要充分利用成熟的工业技术，不拘泥于特定的工具和方法，特别是不盲目追求标准化、技术的先进性或者新颖性，以实际的业务出发，循序渐进推进工程的实施。
- 知识存储，能用关系型数据库先用关系型数据库，再考虑图数据库，再考虑RDF数据库，尽量选成熟的解决方案