数据赋能：知识图谱崛起

林志刚 51195100036

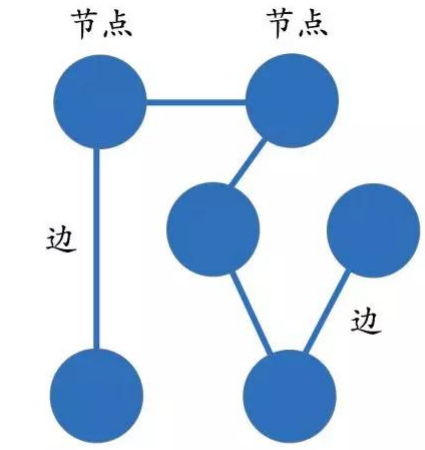
1. 概论

随着移动互联网的发展，万物互联成为了可能，这种互联所产生的数据也在爆发式地增长，而且这些数据恰好可以作为分析关系的有效原料。如果说以往的智能分析专注在每一个个体上，在移动互联网时代则除了个体，这种个体之间的关系也必然成为我们需要深入分析的很重要一部分。 在一项任务中，只要有关系分析的需求，知识图谱就“有可能”派的上用场。

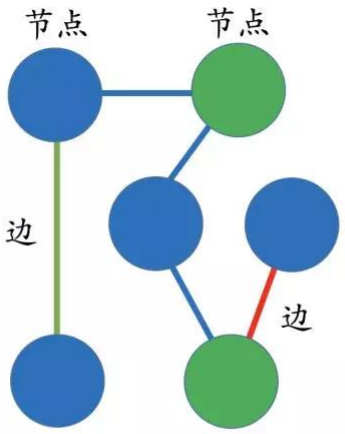
2. 知识图谱的定义

知识图谱是由 Google 公司在 2012 年提出来的一个新的概念。从学术的角度，我们可以对知识图谱给一个这样的定义：**“知识图谱本质上是语义网络 (Semantic Network) 的知识库”**。但这有点抽象，所以换个角度，从实际应用的角度出发其实**可以简单地把知识图谱理解成多关系图（Multi-relational Graph）。**

那什么叫多关系图呢？ 学过数据结构的都应该知道什么是图（Graph）。图是由节点（Vertex）和边（Edge）来构成，但这些图通常只包含一种类型的节点和边。但相反，**多关系图一般包含多种类型的节点和多种类型的边**。比如左下图表示一个经典的图结构，右边的图则表示多关系图，因为图里包含了多种类型的节点和边。这些类型由不同的颜色来标记。



包含一种类型的节点和边



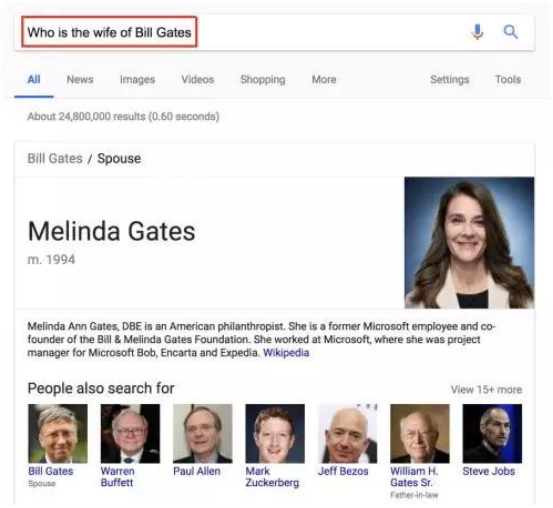
包含多种类型的节点和边

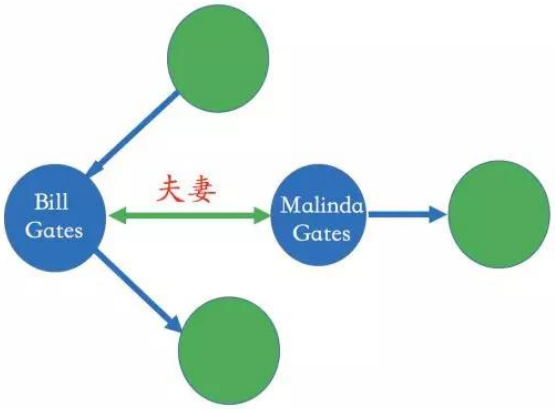
（不同形状和颜色代表不同种类的节点和边）

在知识图谱里，我们通常用“实体（Entity）”来表达图里的节点、用“关系（Relation）”来表达图里的“边”。**实体指的是现实世界中的事物比如人、地名、概念、药物、公司等，关系则用来表达不同实体之间的某种联系，**比如人-“居住在”-北京、张三和李四是“朋友”、逻辑回归是深度学习的“先导知识”等等。

3. 知识图谱的表示

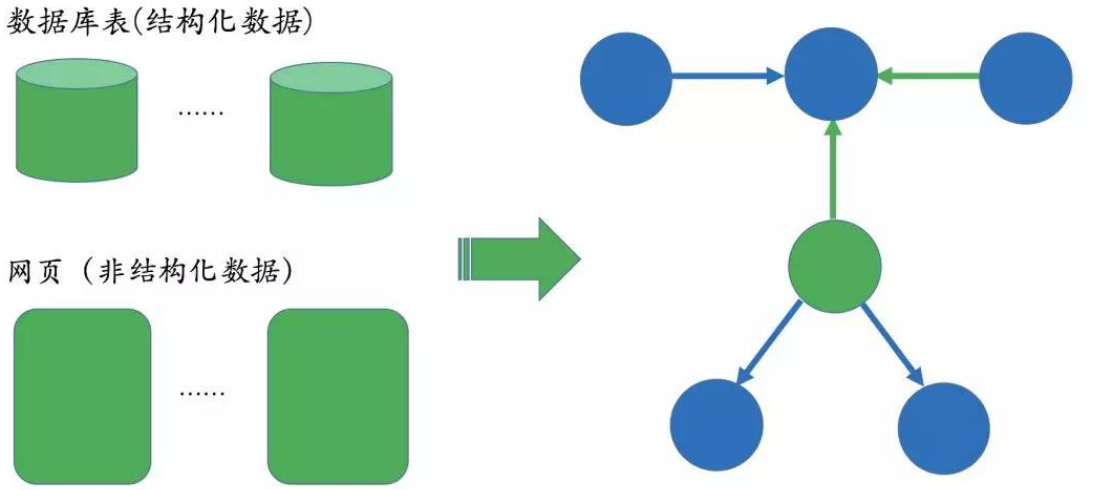
知识图谱应用的前提是已经构建好了知识图谱,可以认为是一个知识库。这也是为什么它可以用来回答一些搜索相关问题的原因，比如在 Google 搜索引擎里输入“Who is the wife of Bill Gates?”，可以得到答案-“Melinda Gates”。这是因为在系统层面上已经创建好了一个包含“Bill Gates”和“Melinda Gates”的实体以及他俩之间关系的知识库。所以，当执行搜索的时候，就可以通过关键词提取（"Bill Gates", "Melinda Gates", "wife"）以及知识库上的匹配可以直接获得最终的答案。这种搜索方式跟传统的搜索引擎是不一样的，一个传统的搜索引擎它返回的是网页、而不是最终的答案，所以就多了一层用户自己筛选并过滤信息的过程。





4. 知识抽取

知识图谱的构建是后续应用的基础，而且构建的前提是需要把数据从不同的数据源中抽取出来。对于垂直领域的知识图谱来说，它数据源主要来自两种渠道：一种是业务本身的数据，这部分数据通常包含在公司内的数据库表并以结构化的方式存储；另一种是网络上公开、抓取的数据，这些数据通常是以网页的形式存在所以是非结构化的数据。前者一般只需要简单预处理即可以作为后续 AI 系统的输入，但后者一般需要借助于自然语言处理等技术来提取出结构化信息。比如在上面的搜索例子里，Bill Gates 和 Malinda Gate 的关系就可以从非结构化数据中提炼出来，比如维基百科等数据源。



信息抽取的难点在于处理非结构化数据。在下面的图中，我们给出了一个实例。左边是一段非结构化的英文文本，右边是从这些文本中抽取出来的实体和关系。在构建类似的图谱过程当中，主要涉及以下几个方面的自然语言处理技术：

a. 实体命名识别（Name Entity Recognition）
b. 关系抽取（Relation Extraction）
c. 实体统一（Entity Resolution）
d. 指代消解（Coreference Resolution）

5. 知识图谱的存储

知识图谱主要有两种存储方式：一种是基于 RDF 的存储；另一种是基于图数据库的存储。它们之间的区别如下图所示。RDF 一个重要的设计原则是数据的易发布以及共享，图数据库则把重点放在了高效的图查询和搜索上。其次，RDF 以三元组的方式来存储数据而且不包含属性信息，但图数据库一般以属性图为基本的表示形式，所以实体和关系可以包含属性，这就意味着更容易表达现实的业务场景。

RDF	图数据库
○ 存储三元组（Triple）	○ 节点和关系可以带有属性
○ 标准的推理引擎	○ 没有标准的推理引擎
○ W3C标准	○ 图的遍历效率高
○ 易于发布数据	○ 事务管理
○ 多数为学术界场景	○ 基本为工业界场景

6. 金融知识图谱的搭建

一个实际的具体案例，讲解怎么一步步搭建可落地的金融风控领域的知识图谱系统。

1. 业务理解	10%
2. 知识图谱设计	10%
3. 算法	50%
4. 开发	30%

很多人以为…

1. 业务理解	30%
2. 知识图谱设计	30%
3. 算法	20%
4. 开发	20%

其实…

一个完整的知识图谱的构建包含以下几个步骤：**1. 定义具体的业务问题 2. 数据的收集 & 预处理 3. 知识图谱的设计 4. 把数据存入知识图谱 5. 上层应用的开发，以及系统的评估。**下面我们就按照这个流程来讲一下每个步骤所需要做的事情以及需要思考的问题。

6.1 定义具体的业务问题

在 P2P 网贷环境下，最核心的问题是风控，也就是怎么去评估一个借款人的风险。在线上的环境下，欺诈风险尤其为严重，而且很多这种风险隐藏在复杂的关系网络之中，而且知识图谱正好是为这类问题所设计的，所以我们“有可能”期待它能在欺诈，这个问题上带来一些价值。

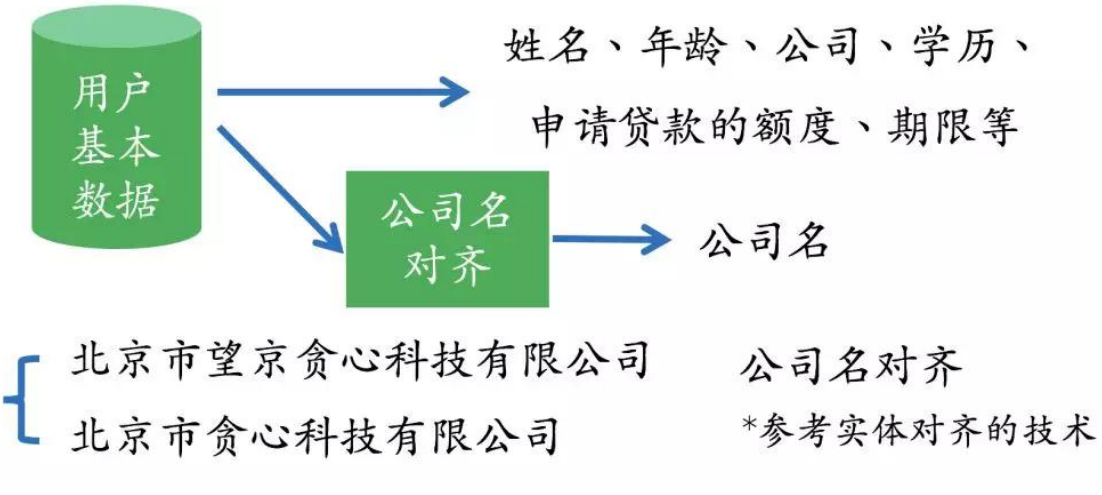
6.2 数据收集 & 预处理

下一步就是要确定数据源以及做必要的数据预处理。针对于数据源，我们需要考虑以下几点：1. 我们已经有哪些数据？ 2. 虽然现在没有，但有可能拿到哪些数据？ 3. 其中哪部分数据可以用来降低风险？ 4. 哪部分数据可以用来构建知识图谱？

对于反欺诈，有几个数据源是我们很容易想得到的，包括用户的基本信息、行为数据、运营商数据、网络上的公开信息等等。假设我们已经有了一个数据源的列表清单，则下一步就要看哪些数据需要进一步的处理，比如对于非结构化数据我们或多或少都需要用到跟自然语言处理相关的技术。 用户填写的基本信息基本上会存储在业务表里，除了个别字段需要进一步处理，很多字段则直接可以用于建模或者添加到知识图谱系统里。对于行为数据来说，我们则需要通过一些

简单的处理，并从中提取有效的信息比如“用户在某个页面停留时长”等等。 对于网络上公开的网页数据，则需要一些信息抽取相关的技术。

举个例子，对于用户的基本信息，我们很可能需要如下的操作。一方面，用户信息比如姓名、年龄、学历等字段可以直接从结构化数据库中提取并使用。但另一方面，对于填写的公司名来说，我们有可能需要做进一步的处理。比如部分用户填写“北京贪心科技有限公司”，另外一部分用户填写“北京望京贪心科技有限公司”，其实指向的都是同一家公司。所以，这时候我们需要做公司名的对齐，用到的技术细节可以参考前面讲到的实体对齐技术。



6.3 知识图谱的设计

图谱的设计是一门艺术，不仅要对业务有很深的理解、也需要对未来业务可能的变化有一定预估，从而设计出最贴近现状并且性能高效的系统。在知识图谱设计的问题上，我们肯定会面临以下几个常见的问题：**1. 需要哪些实体、关系和属性？ 2. 哪些属性可以做为实体，哪些实体可以作为属性？ 3. 哪些信息不需要放在知识图谱中？**

6.4 把数据存入知识图谱

存储上我们要面临存储系统的选择，但由于我们设计的知识图谱带有属性，图数据库可以作为首选。但至于选择哪个图数据库也要看业务量以及对效率的要求。如果数据量特别庞大，则 Neo4j 很可能满足不了业务的需求，这时候不得不去选择支持准分布式的系统比如 OrientDB, JanusGraph 等，或者通过效率、冗余

原则把信息存放在传统数据库中，从而减少知识图谱所承载的信息量。 通常来讲，对于 10 亿节点以下规模的图谱来说 Neo4j 已经足够了。

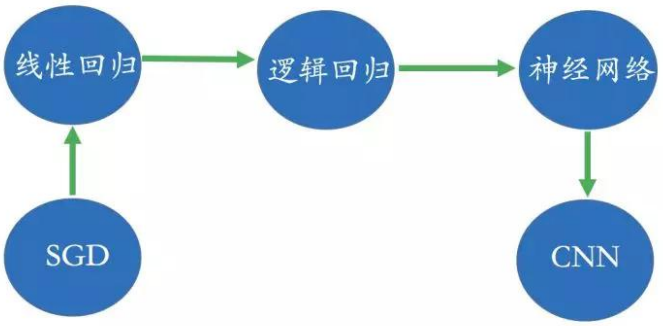
6.5 上层应用的开发

等构建好知识图谱之后，接下来就要使用它来解决具体的问题。对于风控知识图谱来说，首要任务就是挖掘关系网络中隐藏的欺诈风险。**从算法的角度来讲，有两种不同的场景：一种是基于规则的；另一种是基于概率的**。鉴于目前 AI 技术的现状，基于规则的方法论还是在垂直领域的应用中占据主导地位，但随着数据量的增加以及方法论的提升，基于概率的模型也将会逐步带来更大的价值。

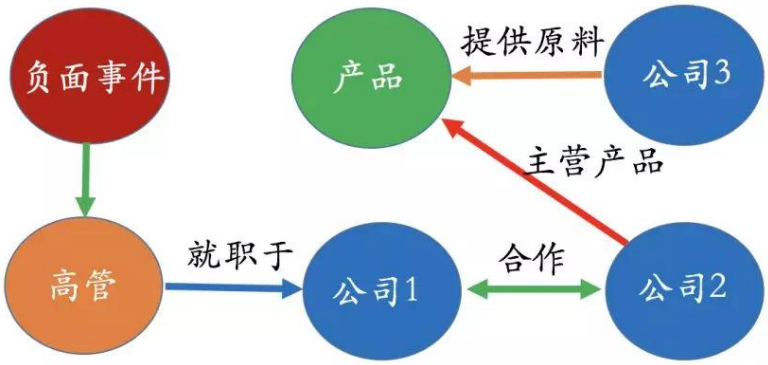
7. 知识图谱在其他行业中的应用

除了金融领域，知识图谱的应用可以涉及到很多其他的行业，包括医疗、教育、证券投资、推荐等等。其实，只要有关系存在，则有知识图谱可发挥价值的地方。 在这里简单举几个垂直行业中的应用。

比如对于教育行业，我们经常谈论个性化教育、因材施教的理念。其核心在于理解学生当前的知识体系，而且这种知识体系依赖于我们所获取到的数据比如交互数据、评测数据、互动数据等等。为了分析学习路径以及知识结构，我们则需要针对于一个领域的概念知识图谱，简单来讲就是概念拓扑结构。在下面的图中，我们给出了一个非常简单的概念图谱：比如为了学习逻辑回归则需要先理解线性回归；为了学习 CNN，得对神经网络有所理解等等。所有对学生的评测、互动分析都离不开概念图谱这个底层的数据。



在证券领域，我们经常会关心比如“一个事件发生了，对哪些公司产生什么样的影响？”比如有一个负面消息是关于公司 1 的高管，而且我们知道公司 1 和公司 2 有种很密切的合作关系，公司 2 有个主营产品是由公司 3 提供的原料基础上做出来的。



其实有了这样的一个知识图谱，我们很容易回答哪些公司有可能会被这次的负面事件所影响。当然，仅仅是“有可能”，具体会不会有强相关性必须由数据来验证。所以在这里，知识图谱的好处就是把我们所需要关注的范围很快给我们圈定。接下来的问题会更复杂一些，比如既然我们知道公司 3 有可能被这次事件所影响，那具体影响程度有多大？ 对于这个问题，光靠知识图谱是很难回答的，必须要有一个影响模型、以及需要一些历史数据才能在知识图谱中做进一步推理以及计算。

8. 实践上的几点建议

知识图谱是一个比较新的工具，它的主要作用还是在于分析关系，尤其是深度的关系。所以在业务上，首先要确保它的必要性，其实很多问题可以用非知识图谱的方式来解决。知识图谱领域一个最重要的话题是知识的推理。 而且知识的推理是走向强人工智能的必经之路。但很遗憾的，目前很多语义网络的角度讨论的推理技术（比如基于深度学习，概率统计）很难在实际的垂直应用中落地。其实目前最有效的方式还是基于一些规则的方法论，除非我们有非常庞大的数据集。

参考文献

Wang Z, Zhang J, Feng J, et al. Knowledge graph embedding by translating on hyperplanes[C]//Twenty-Eighth AAAI conference on artificial intelligence. 2014.]

Lin Y, Liu Z, Sun M, et al. Learning entity and relation embeddings for knowledge graph completion[C]//Twenty-ninth AAAI conference on artificial intelligence. 2015.

[1]雷新强.基于知识图谱的互联网金融研究主体、研究热点与演进分析[J].辽宁工业大学学报(社会科学版),2019,21(06):39-42.