知识图谱的应用

李文哲
机器学习 话题的优秀回答者
关注他

297 人赞了该文章

> 原文链接（同一个作者， 大数据中心公众号）： 知识图谱的应用

导读

知识图谱 (Knowledge Graph) 是当前的研究热点。自从2012年Google推出自己第一版知识图谱以来，它在学术界和工业界掀起了一股热潮。各大互联网企业在之后的短短一年内纷纷推出了自己的知识图谱产品以作为回应。比如在国内，互联网巨头百度和搜狗分别推出”知心“和”知立方”来改进其搜索质量。那么与这些传统的互联网公司相比，对处于当今风口浪尖上的行业 - 互联网金融，知识图谱可以有哪方面的应用呢?

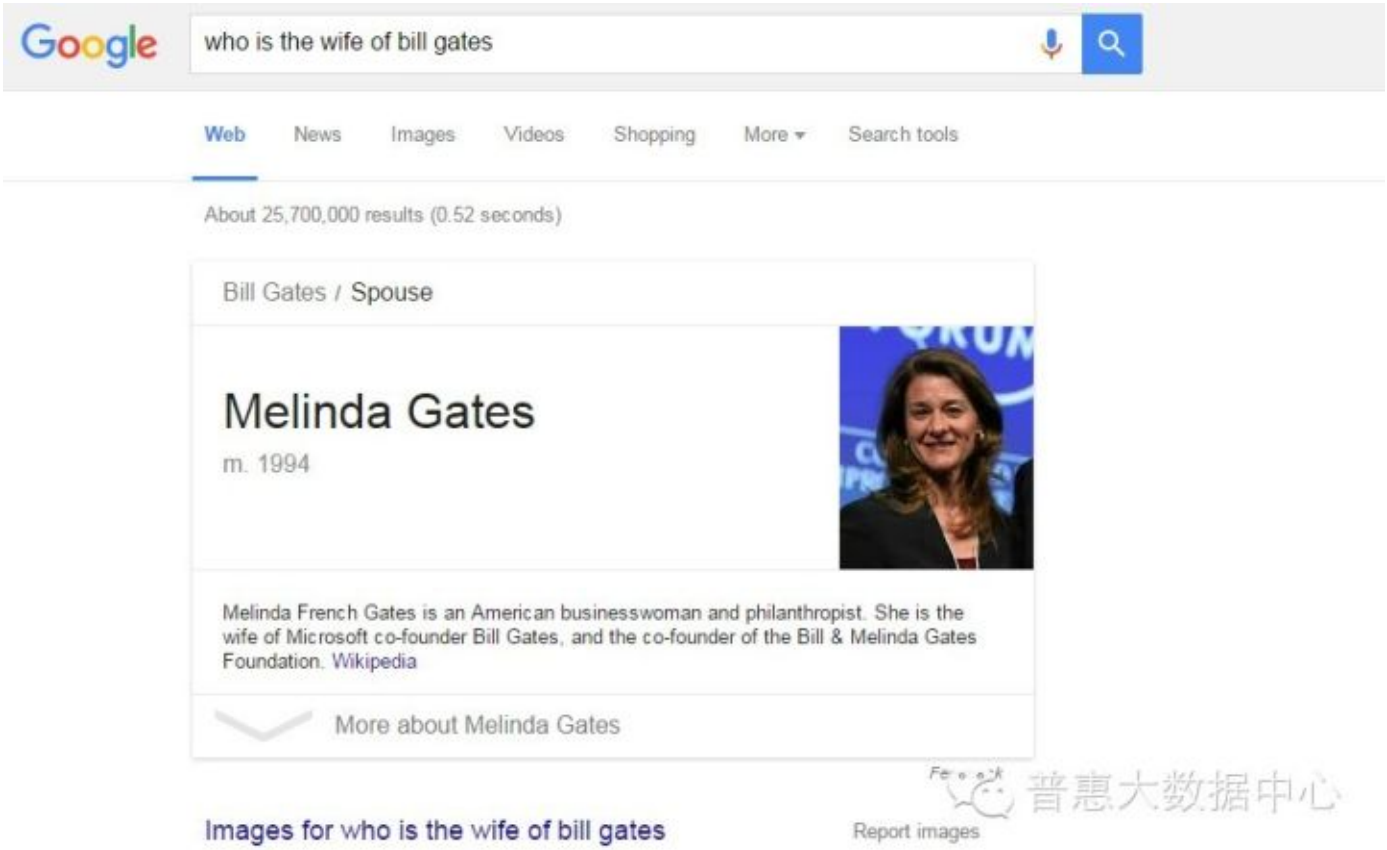
> 目录
>
>

1. 什么是知识图谱?

知识图谱本质上是语义网络，是一种基于图的数据结构，由节点(Point)和边(Edge)组成。在知识图谱里，每个节点表示现实世界中存在的“实体”，每条边为实体与实体之间的“关系”。知识图谱是关系的最有效的表示方式。通俗地讲，**知识图谱就是把所有不同种类的信息（Heterogeneous Information）连接在一起而得到的一个关系网络。知识图谱提供了从“关系”的角度去分析问题的能力。**

索质量。比如在Google的搜索框里输入Bill Gates的时候，搜索结果页面的右侧还会出现Bill Gates相关的信息比如出生年月，家庭情况等等。

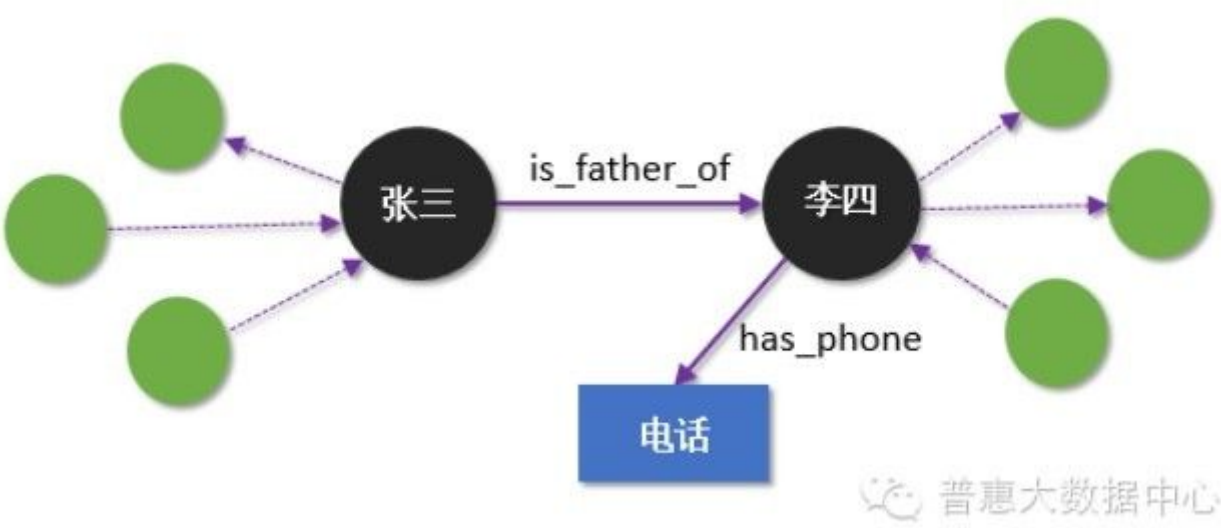
另外，对于稍微复杂的搜索语句比如 ”Who is the wife of Bill Gates“，Google能准确返回他的妻子Melinda Gates。这就说明搜索引擎通过知识图谱真正理解了用户的意图。



谱表示方式和应用，这也是工业界比较关心的话题。

2. 知识图谱的表示

假设我们用知识图谱来描述一个事实（Fact） - “张三是李四的父亲”。这里的实体是张三和李四，关系是“父亲”（is_father_of）。当然，张三和李四也可能会跟其他人存在着某种类型的关系（暂时不考虑）。当我们把电话号码也作为节点加入到知识图谱以后（电话号码也是实体），人和电话之间也可以定义一种关系叫 has_phone，就是说某个电话号码是属于某个人。下面的图就展示了这两种不同的关系。



另外，我们可以把时间作为属性（Property）添加到 has_phone 关系里来表示开通电话号码的时间。这种属性不仅可以加到关系里，还可以加到实体当中，当我们把所有这些信息作为关系或者实体的属性添加后，所得到的图谱称之为属性图 （Property Graph）。属性图和传统的RDF格式都可以作为知识图谱的表示和存储方式，但二者还是有区别的，这将在后面章节做简单说明。

3. 知识图谱的存储

知识图谱是基于图的数据结构，它的存储方式主要有两种形式：RDF存储格式和图数据库(Graph Database)。至于它们有哪些区别，请参考【1】。下面的曲线表示各种数据存储类型在最近几年的发展情况。从这里我们可以明显地看到基于图的存储方式在整个数据库存储领域的飞速发展。这幅曲线图来源于 Graph DBMS increased their popularity by 500% within the last 2 years

下面的列表表示的是目前比较流行的基于图存储的数据库排名。从这个排名中可以看出**neo4j在整个图存储领域里占据着NO.1的地位，而且在RDF领域里Jena还是目前为止最为流行的存储框架。**这部分数据来源于 DB-Engines Ranking

21	Neo4j（图）
32	MarkLogic（XML）
42	Titan（图）
46	OrientDB（图，文档）
61	Virtuoso（RDF,关系等）
80	Jena（RDF）
88	Sesame（RDF）
90	ArangoDB（图）
120	AllegroGraph（RDF）

当然，如果需要设计的知识图谱非常简单，而且查询也不会涉及到1度以上的关联查询，我们也可以选择用关系型数据存储格式来保存知识图谱。但对那些稍微复杂的关系网络（现实生活中的实体和关系普遍都比较复杂），知识图谱的优点还是非常明显的。首先，在关联查询的效率上会比传统的存储方式有显著的提高。当我们涉及到2,3度的关联查询，基于知识图谱的查询效率会高出几千倍甚至几百万倍。其次，基于图的存储在设计上会非常灵活，一般只需要局部的改动即可。比如我们有一个新的数据源，我们只需要在已有的图谱上插入就可以。于此相反，关系型存储方式灵活性方面比较差，它所有的Schema都是提前定义好的，如果后续要改变，它的代价是非常高的。最后，**把实体和关系存储在图数据结构是一种符合整个故事逻辑的最好的方式**。

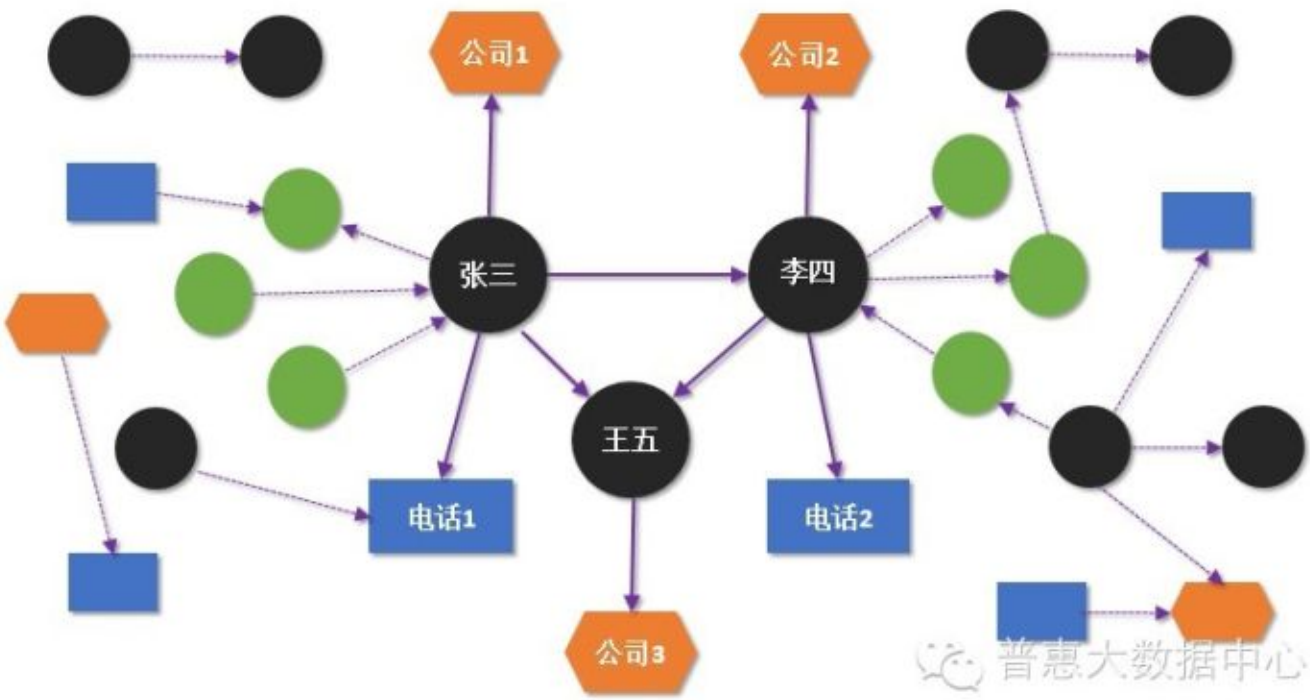
4. 应用

然可以发挥它潜在的价值， 我们在后续的文章中会继续讨论。

反欺诈

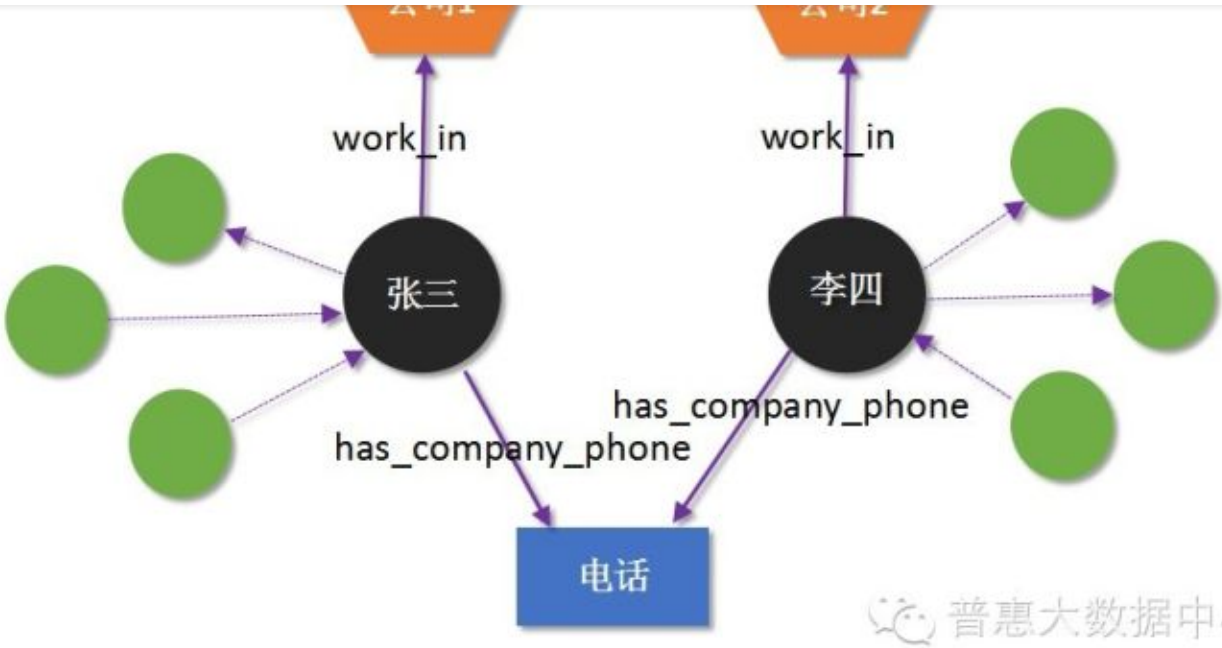
反欺诈是风控中非常重要的一道环节。基于大数据的反欺诈的难点在于如何把不同来源的数据（结构化，非结构）整合在一起，并构建反欺诈引擎，从而有效地识别出欺诈案件（比如身份造假，团体欺诈，代办包装等）。而且不少欺诈案件会涉及到复杂的关系网络，这也给欺诈审核带来了新的挑战。 知识图谱，作为关系的直接表示方式，可以很好地解决这两个问题。 首先，知识图谱提供非常便捷的方式来添加新的数据源，这一点在前面提到过。其次，知识图谱本身就是用来表示关系的，这种直观的表示方法可以帮助我们更有效地分析复杂关系中存在的特定的潜在风险。

反欺诈的核心是人，首先需要把与借款人相关的所有的数据源打通，并构建包含多数据源的知识图谱，从而整合成为一台机器可以理解的结构化的知识。在这里，我们不仅可以整合借款人的基本信息（比如申请时填写的信息），还可以把借款人的消费记录、行为记录、网上的浏览记录等整合到整个知识图谱里，从而进行分析和预测。这里的一个难点是很多的数据都是从网络上获取的非结构化数据，需要利用机器学习、自然语言处理技术把这些数据变成结构化的数据。

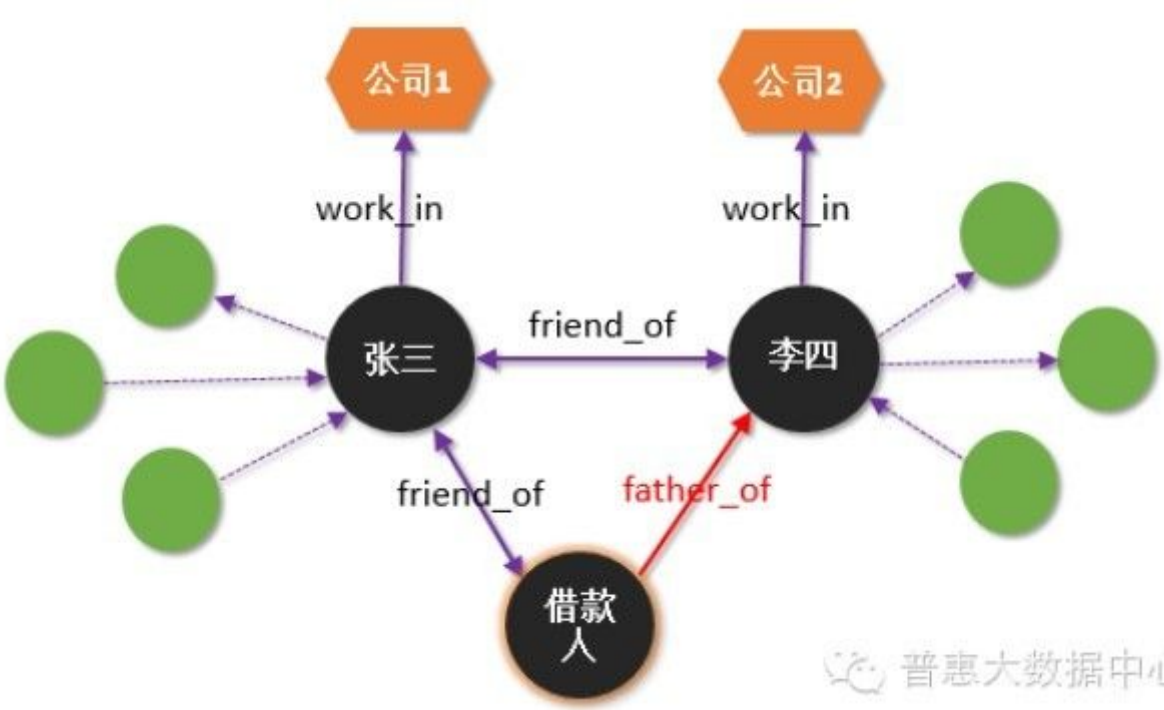


不一致性验证

不一致性验证可以用来判断一个借款人的欺诈风险，这个跟交叉验证类似。比如借款人张三和借款人李四填写的是同一个公司电话，但张三填写的公司和李四填写的公司完全不一样，这就成了风险点，需要审核人员格外的注意。

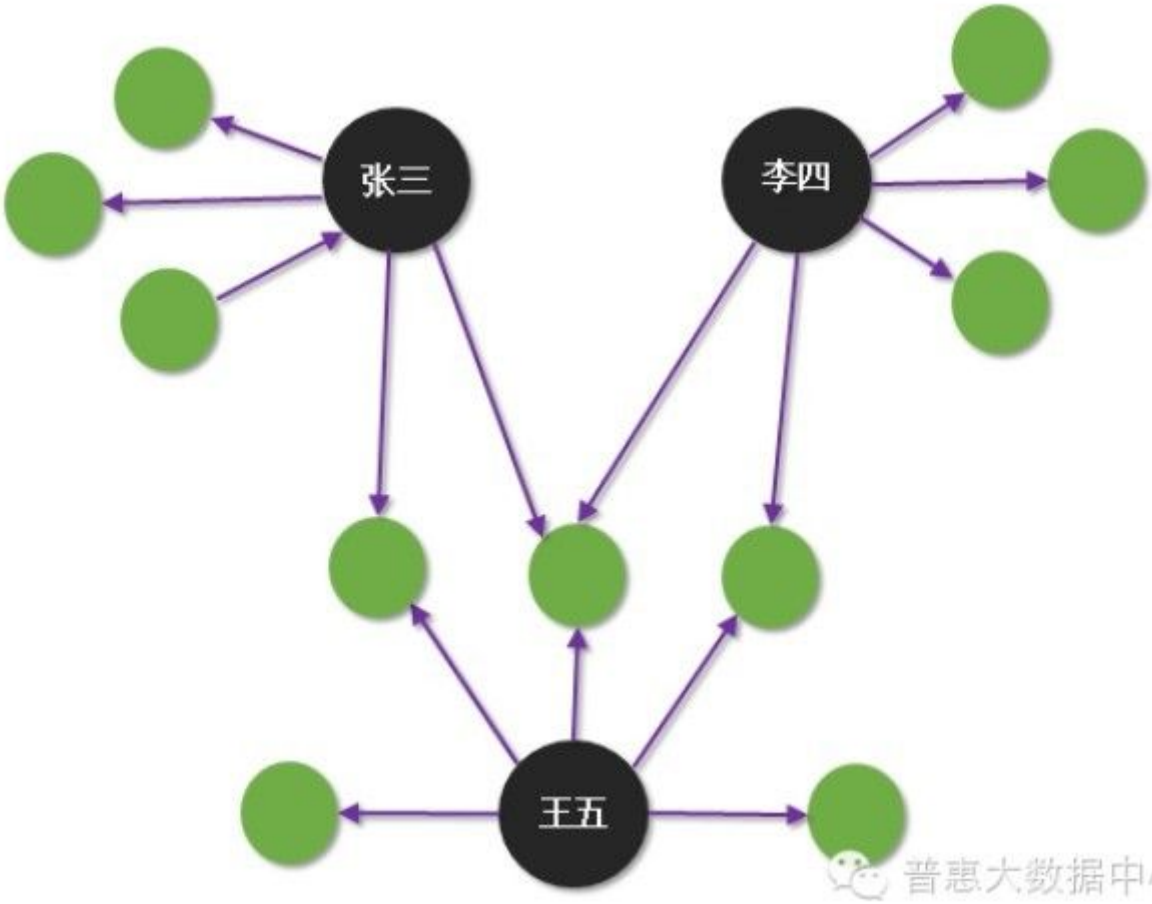


再比如，借款人说跟张三是朋友关系，跟李四是父子关系。当我们试图把借款人的信息添加到知识图谱里的时候，“一致性验证”引擎会触发。引擎首先会去读取张三和李四的关系，从而去验证这个“三角关系”是否正确。很显然，朋友的朋友不是父子关系，所以存在着明显的不一致性。



不一致性验证涉及到知识的推理。通俗地讲，**知识的推理可以理解成“链接预测”，也就是从已有的关系图谱里推导出新的关系或链接。** 比如在上面的例子，假设张三和李四是朋友关系，而且张三和借款人也是朋友关系，那我们可以推理出借款人和李四也是朋友关系。

相比虚假身份的识别，组团欺诈的挖掘难度更大。这种组织在非常复杂的关系网络里隐藏着，不容易被发现。当我们只有把其中隐含的关系网络梳理清楚，才有可能去分析并发现其中潜在的风险。知识图谱，作为天然的关系网络的分析工具，可以帮助我们更容易地去识别这种潜在的风险。举一个简单的例子，有些组团欺诈的成员会用虚假的身份去申请贷款，但部分信息是共享的。下面的图大概说明了这种情形。从图中可以看出张三、李四和王五之间没有直接的关系，但通过关系网络我们很容易看出这三者之间都共享着某一部分信息，这就让我们马上联想到欺诈风险。虽然组团欺诈的形式众多，但有一点值得肯定的是知识图谱一定会比其他任何的工具提供更佳便捷的分析手段。

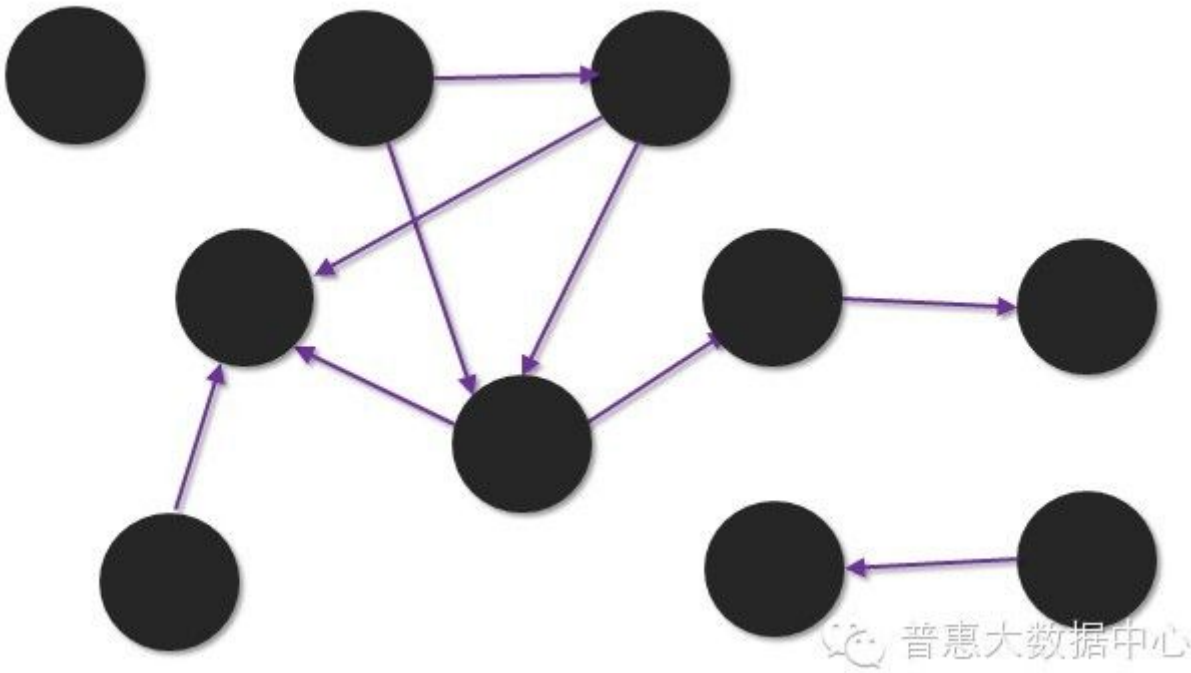


异常分析（Anomaly Detection）

异常分析是数据挖掘研究领域里比较重要的课题。我们可以把它简单理解成从给定的数据中找出“异常”点。在我们的应用中，这些”异常“点可能会关联到欺诈。既然知识图谱可以看做是一个图（Graph），知识图谱的异常分析也大都是基于图的结构。由于知识图谱里的实体类型、关系类型不同，异常分析也需要把这些额外的信息考虑进去。大多数基于图的异常分析的计算量比较大，可以选择做离线计算。在我们的应用框架中，可以把异常分析分为两大类： 静态分析和动态分析，后面会逐一讲到。

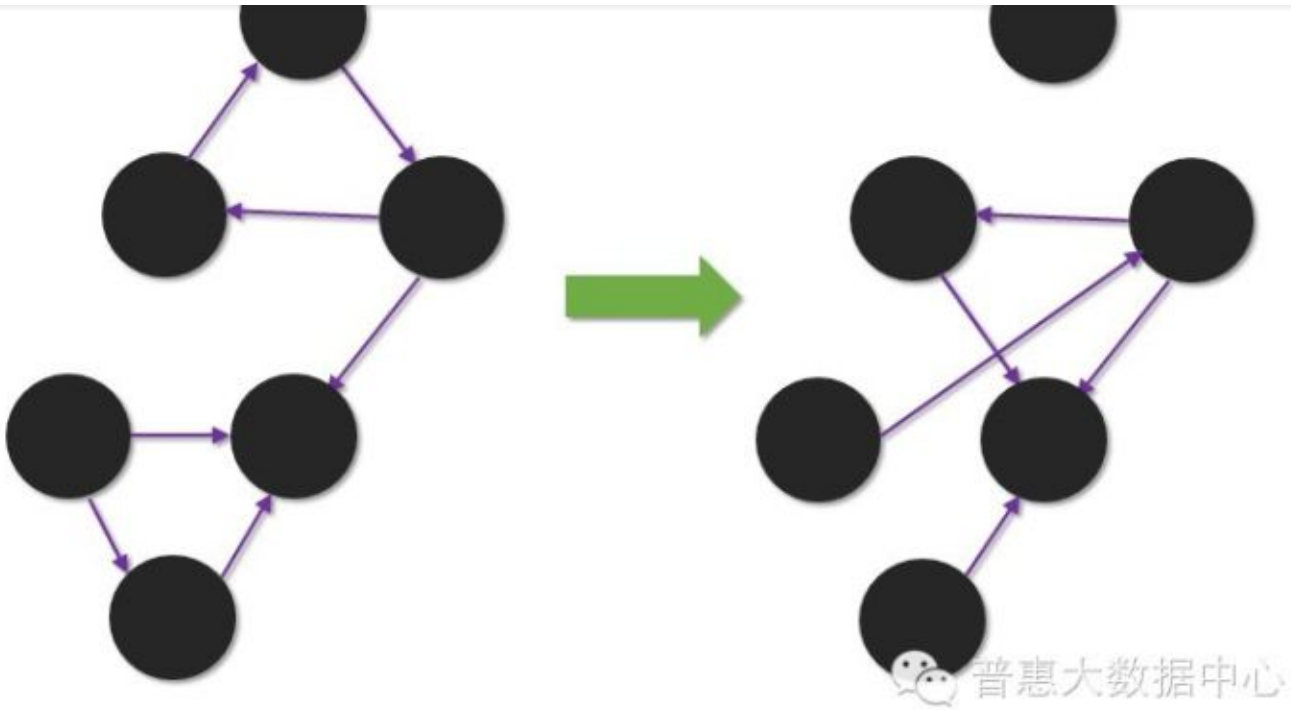
- 静态分析

以针对这些异常的结构，我们可以做出进一步的分析。



- 动态分析

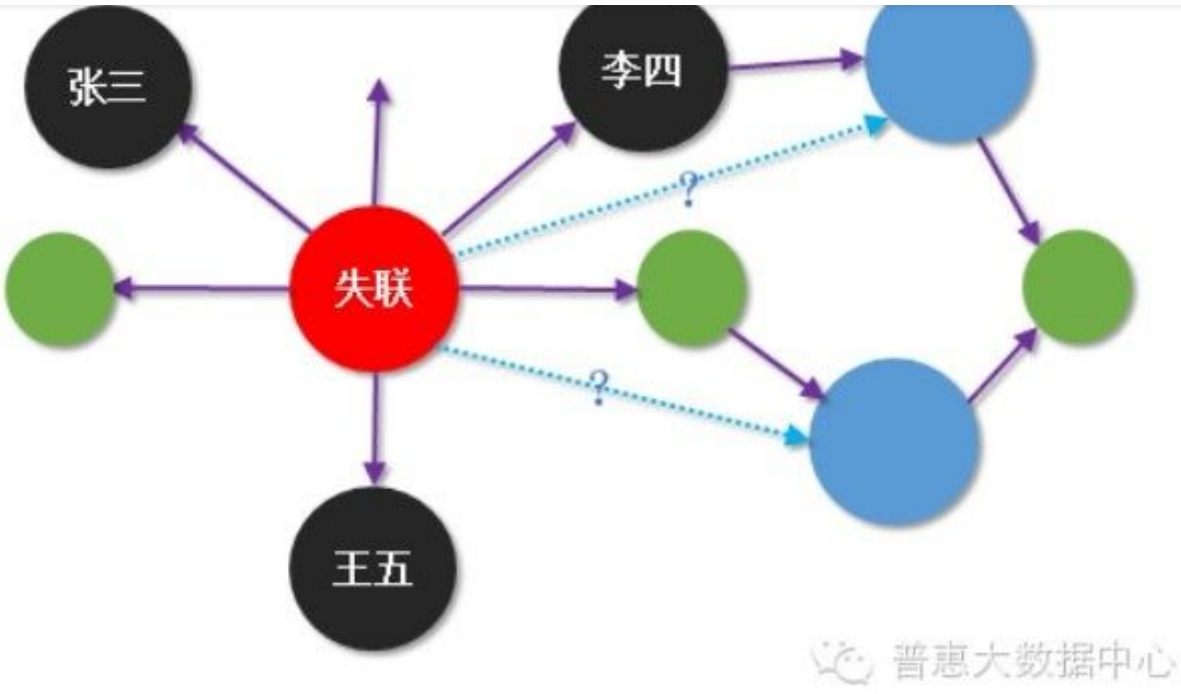
所谓的动态分析指的是分析其结构随时间变化的趋势。我们的假设是，在短时间内知识图谱结构的变化不会太大，如果它的变化很大，就说明可能存在异常，需要进一步的关注。分析结构随时间的变化会涉及到时序分析技术和图相似性计算技术。有兴趣的读者可以去参考这方面的资料【2】。



> 失联客户管理

除了贷前的风险控制，知识图谱也可以在贷后发挥其强大的作用。比如在贷后失联客户管理的问题上，**知识图谱可以帮助我们挖掘出更多潜在的新的联系人**，从而提高催收的成功率。

现实中，不少借款人在借款成功后出现不还款现象，而且玩“捉迷藏”，联系不上本人。即便试图去联系借款人曾经提供过的其他联系人，但还是没有办法联系到本人。这就进入了所谓的“失联”状态，使得催收人员也无从下手。那接下来的问题是，在失联的情况下，我们有没有办法去挖掘跟借款人有关系的新的联系人？ 而且这部分人群并没有以关联联系人的身份出现在我们的知识图谱里。如果我们能够挖掘出更多潜在的新的联系人，就会大大地提高催收成功率。举个例子，在下面的关系图中，借款人跟李四有直接的关系，但我们却联系不上李四。那有没有可能通过2度关系的分析，预测并判断哪些李四的联系人可能会认识借款人。这就涉及到图谱结构的分析。



智能搜索及可视化展示

基于知识图谱，我们也可以提供智能搜索和数据可视化的服务。智能搜索的功能类似于知识图谱在Google, Baidu上的应用。也就是说，对于每一个搜索的关键词，我们可以通过知识图谱来返回更丰富，更全面的信息。比如搜索一个人的身份证号，我们的智能搜索引擎可以返回与这个人相关的所有历史借款记录、联系人信息、行为特征和每一个实体的标签（比如黑名单，同业等）。另外，可视化的好处不言而喻，通过可视化把复杂的信息以非常直观的方式呈现出来， 使得我们对隐藏信息的来龙去脉一目了然。

精准营销

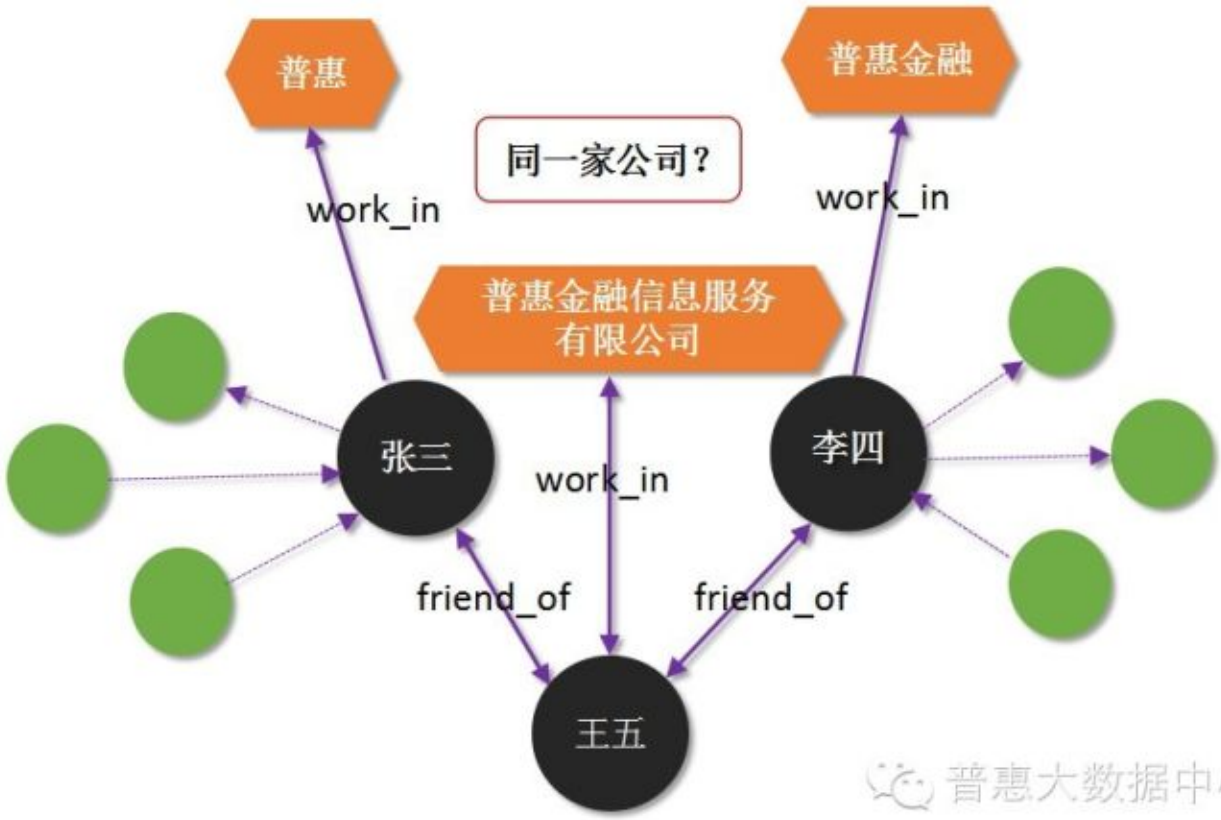
“A knowledge graph allows you to take core information about your customer—their name, where they reside, how to contact them—and relate it to who else they know, how they interact on the web, and more”-- Michele Goetz, a Principal Analyst at Forrester Research

一个聪明的企业可以比它的竞争对手以更为有效的方式去挖掘其潜在的客户。在互联网时代，营销手段多种多样，但不管有多少种方式，都离不开一个核心 - 分析用户和理解用户。知识图谱可以结合多种数据源去分析实体之间的关系，从而对用户的行为有更好的理解。比如一个公司的市场经理用知识图谱来分析用户之间的关系，去发现一个组织的共同喜好，从而可以有针对性的对某一类人群制定营销策略。只有我们能更好的、更深入的（Deep understanding）理解用户的需求，我们才能更好地去做营销。

知识图谱在工业界还没有形成大规模的应用。即便有部分企业试图往这个方向发展，但很多仍处于调研阶段。主要的原因是很多企业对知识图谱并不了解，或者理解不深。但有一点可以肯定的是，**知识图谱在未来几年内必将成为工业界的热门工具**，这也是从目前的趋势中很容易预测到的。当然，知识图谱毕竟是一个比较新的工具，所以在实际应用中一定会涉及到或多或少的挑战。

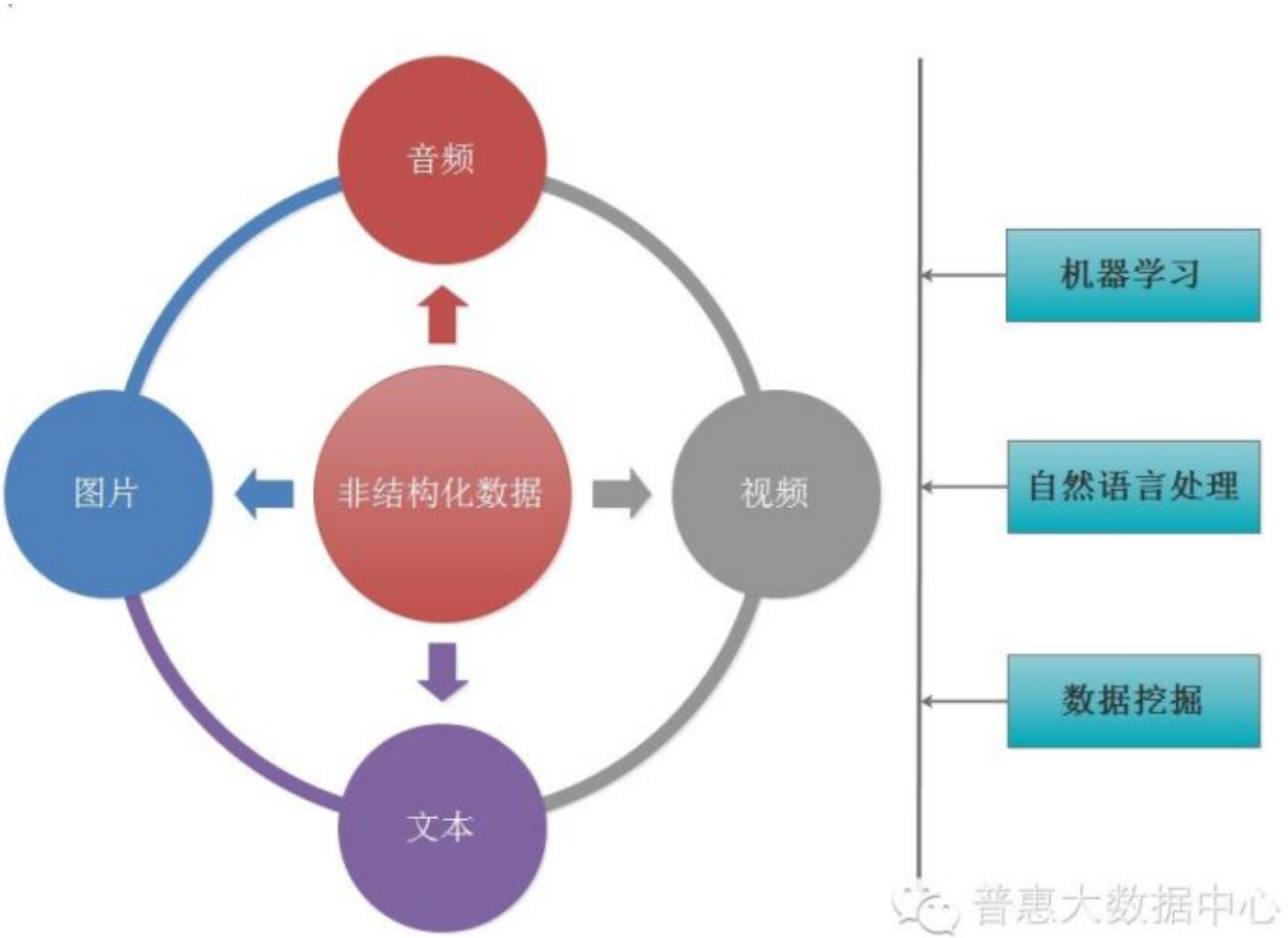
> 数据的噪声

首先，数据中存在着很多的噪声。即便是已经存在库里的数据，我们也不能保证它有100%的准确性。在这里主要从两个方面说起。第一，目前积累的数据本身有错误，所以这部分错误数据需要纠正。 最简单的纠正办法就是做离线的不一致性验证，这点在前面提过。第二， 数据的冗余。比如借款人张三填写公司名字为”普惠“，借款人李四填写的名字为”普惠金融“，借款人王五则填写成”普惠金融信息服务有限公司“。虽然这三个人都隶属于一家公司，但由于他们填写的名字不同，计算机则会认为他们三个是来自不同的公司。那接下来的问题是，怎么从海量的数据中找出这些存在歧义的名字并将它们合并成一个名字？ 这就涉及到自然语言处理中的”消歧分析”技术。



> 非结构化数据处理能力

的信息是一件非常有挑战性的任务，这对掌握的机器学习，数据挖掘，自然语言处理能力提出了更高的门槛。



> 知识推理

推理能力是人类智能的重要特征，使得我们可以从已有的知识中发现隐含的知识， 一般的推理往往需要一些规则的支持【3】。例如“朋友”的“朋友”，可以推理出“朋友”关系，“父亲”的“父亲”可以推理出“祖父”的关系。再比如张三的朋友很多也是李四的朋友，那我们可以推测张三和李四也很有可能是朋友关系。当然，这里会涉及到概率的问题。当信息量特别多的时候，怎么把这些信息（side information）有效地与推理算法结合在一起才是最关键的。**常用的推理算法包括基于逻辑（Logic） 的推理和基于分布式表示方法（Distributed Representation）的推理**。随着深度学习在人工智能领域的地位变得越来越重要，基于分布式表示方法的推理也成为目前研究的热点。如果有兴趣可以参考一下这方面目前的工作进展【4,5,6,7】。

> 大数据、小样本、构建有效的生态闭环是关键

的欺诈样本数量不多，即便有几百万个贷款申请，最后被我们标记为欺诈的样本很可能也就几万个而已。这对机器学习的建模提出了更高的挑战。每一个欺诈样本我们都是以很高昂的“代价”得到的。随着时间的推移，我们必然会收集到更多的样本，但样本的增长空间还是有局限的。这有区别于传统的机器学习系统，比如图像识别，不难拿到好几十万甚至几百万的样本。

在这种小样本条件下，构建有效的生态闭环尤其的重要。**所谓的生态闭环，指的是构建有效的自反馈系统使其能够实时地反馈给我们的模型，并使得模型不断地自优化从而提升准确率**。为了搭建这种自学习系统，我们不仅要完善已有的数据流系统，而且要深入到各个业务线，并对相应的流程进行优化。这也是整个反欺诈环节必要的过程，我们要知道整个过程都充满着博弈。所以我们需要不断地通过反馈信号来调整我们的策略。

6. 结语

知识图谱在学术界和工业界受到越来越多的关注。除了本文中所提到的应用，知识图谱还可以应用在权限管理，人力资源管理等不同的领域。在后续的文章中会详细地讲到这方面的应用。

参考文献

【1】De Abreu, D., Flores, A., Palma, G., Pestana, V., Pinero, J., Queipo, J., ... & Vidal, M. E. (2013). Choosing Between Graph Databases and RDF Engines for Consuming and Mining Linked Data. In *COLD*.

【2】User Behavior Tutorial

【3】刘知远 知识图谱——机器大脑中的知识库 第二章 知识图谱——机器大脑中的知识库

【4】Nickel, M., Murphy, K., Tresp, V., & Gabrilovich, E. A Review of Relational Machine Learning for Knowledge Graphs.

【5】Socher, R., Chen, D., Manning, C. D., & Ng, A. (2013). Reasoning with neural tensor networks for knowledge base completion. In *Advances in Neural Information Processing Systems* (pp. 926-934).

【6】Bordes, A., Usunier, N., Garcia-Duran, A., Weston, J., & Yakhnenko, O. (2013). Translating embeddings for modeling multi-relational data. In *Advances in Neural Information Processing Systems* (pp. 2787-2795).

3175).

编辑于 2015-12-05

「真诚赞赏，手留余香」

赞赏

3 人已赞赏

文章被以下专栏收录

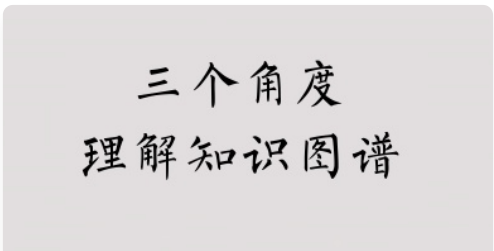
用技术改变世界
专栏讨论人工智能技术、行业应用以及探讨如何成为一名合格的AI人才

关注专栏

推荐阅读

今天咱们来绘制一个知识图谱怎么样?

云渡



三个角度理解知识图谱

徐超

【cs类
网络分析

【昨天玩
录，今天
知识图谱
料，部分
专栏 | 这
谱技术与

芳知

18 条评论

切换为时间排序

Alex Wang 3 年前

两年前做complex networks研究influence maximization，有类似于或可应用于精准营销的想法，今天看到您的介绍印证了当初的想法。写得很棒很友好，学习了:D

2

黄羽 回复 Alex Wang 1 年前

我也在做complex networks，企业图谱和风险分析，想用传染病的模型应用到一夜风险里面，不知道这个能行的通不？模型能不能解释和预测现实世界？doing

赞

Alex Wang 回复 黄羽 1 年前

挠挠头...不做complex networks好多年了，不敢妄言啊.

赞

楚之白 3 年前

关于知识图谱目前有哪些开源工具呢?

赞

贺斯琪 3 年前

碰到内涵丰富、讲得清楚、读得懂的文章很不容易，感谢专栏作者，期待后续的文章

赞

李文哲 (作者) 回复 贺斯琪 3 年前

谢谢支持，节后会发布另一篇文章

赞

果果是枚开心果 2 年前

同问开源工具

2

杨海宏 2 年前

nice!

赞

聂广洋 2 年前

才接触到图谱相关内容，很有帮助和启发

赞

如何深度利用知识图谱进行关联查找

赞

XIAO FENG 2 年前

看来学习知识图谱，还是要把图计算技术捡起来了！文章很棒，深入浅出，贴近应用，赞！

赞

shi davide 2 年前

文章思路很清晰，希望能介绍下如何从百科或者开放文本中建立关系

1

wang liang 1 年前

写的很棒,思路很清晰.读完后能明白图谱的作用,以及重要性.

赞

空leo 1 年前

假设张三和李四是朋友关系，而且张三和借款人也是朋友关系，那我们可以推理出借款人和李四也是朋友关系。

看到这里我觉得不对，朋友的朋友并不一定是朋友，这个关系有点类似于skos:related。

1

上树的鱼 回复 空leo 1 个月前

这里应该有有一个反向的关系描述，或者可传递的关系描述，例如可传递表示(南京位于江苏，江苏位于中国，就可以推理出南京位于中国)，例如双向关系(小明是张阿姨的儿子，就可以知道张阿姨的孩子包括小明)

赞

张小果 1 年前

已粉

赞

双乔巴巴 1 年前

公司为了建设智慧大脑，建起了一个大数据平台和人工智能平台，上面做了很多的应用，例如客户画像、关系网、反欺诈等等，另外，也引入了人脸识别、语音识别、语义分析等，问题来了，怎么把这些信息汇聚起来呢？现在没有想到一个好的载体，不知道构建一个知识图谱作为基础存储工程，是否靠谱？请专家们指点一二。

如果你是用作知识存储，那么用知识图谱准没错，但是看你的意思应该是作为数据和知识的混和存储，建议数据存储还是放在分布式存储中。