**7.质量评估**

与所创立的知识图谱中的信息无关，从初始知识图谱中提取的信息通常是不完整的，它会含有重复的，相互矛盾的甚至是错误的部分，尤其数据是多源时。在使用外部资源创立和丰富知识图谱之后，关键性的一步是评价成果知识图谱的质量。通过质量，我们就可以推测出该知识图谱是否适合于我们的想法。接下来，质量评估就能帮助我们确认可以有效使用知识图谱的情况。

下面我们讨论充斥于各领域多元化数据质量中的质量维度，它从传统数据库领域逐渐发展到知识图谱领域，其中一部分是普通的，但其余部分更加是知识图谱所独有的。在质量维度致力于评价数据质量方面的同时，我们也会探讨质量度量，它提供了衡量这些维度质量方面的方法。我们所探讨的维度和度量分组的灵感来自于Batini和Scannapieco。

**7.1 准确率**

准确率涉及到实体和相关层面——在图谱中由节点和边进行编码——准确反映了真实生活中的现象。准确率还可以进一步细分为三个方面：句法准确度，语义准确度以及及时性。

**7.1.1 句法准确度**是指原始数据和由定义域和/或数据模型定义的语法规则相比之下的准确性。在句法方面不准确的一个普遍情况出现在数据类型节点中，相比于某个定义范围，它与之相矛盾或者根本不理想。例如，假设一个属性的开始时间定义在xsd:dateTime中，那么就会与定义域矛盾，同样也不理想（假设期望值是）。句法准确度的相关度量是指一个给定属性的错值数量和相同属性对应值的总数之间的比值。诸如这样的句法准确度一般都使用验证工具进行度量。

**7.1.2 语义准确度**是指数据值准确反映了真实现象，而它通常会被不精确的提取值，不精确的蕴涵以及破坏等因素影响。例如，给定智力国会位于Valparaíso这条信息就会产生边（通过蕴涵，提取，完成等步骤），但这实际上是不正确的：智利首都应该是圣地亚哥。评估语义错误的程度是非常具有挑战性的。一个方案是提供人工修正，如果要实现自动化则需检查当前所述关系是否与某些资源冲突。另一个方案选择相信基于预测等措施使用个体过程所生成的知识图谱的质量，在这个过程中需要借助专家或特定标准。

**7.1.3 及时性**是指知识图谱包含当前真实状态的最新信息；换句话说，一个知识图谱也许当前是语义准确的，但如果不采取措施及时更新使其保持最新状态，那它很快就会变得不精确。例如，考虑一个用户使用旅行知识图谱查询从一个城市到另一个城市航班的情况。假设随着当前的航班信息，航空时刻表每分钟更新一次，但是知识图谱每一小时才会更新一次，在这种情况下，我们可以看到在知识图谱的及时性方面出现了质量问题。潜在来源依靠知识图谱中时间注释的改变和影响数据时间有效性的上下文表示方法来完成，而及时性的评估正建立在知识图谱基于此的更新频率的基础上。

**7.2 覆盖率**

覆盖率的提出是为了避免范围相关元素的遗漏，一旦遗漏就会产生不完整的查询结果或蕴含，有偏差的模型等。

**7.2.1 完成度**是指所有需要的信息都呈现在一个特定的数据集中，其包含以下方面：（i）模式完成度是指相同图中的类和属性都呈现在数据图谱中；（ii）实体完成度是指对特定实体中遗失值的比率；（iii）总体完成度表示特定类型的全部真实实体在数据集中所占的比例；（iv）关联完成度是指数据集中的实例都是内部关联的。直接衡量完成度是不简单的，因为这需要理想知识图谱的假设，其在问题中包含的知识图谱中的所有元素都应该是“理想”的代表。具体的方法是与给出理想知识图谱样例的黄金标准（可能基于完成度陈述）比较或衡量来自于完备资源的抽取法所对应的召回率等。

**7.2.2 典型性**聚焦于评估包含在知识图谱中或从知识图谱中排除的内容所对应的高级偏差，而不是关注遗失的领域相关元素的比例。同样地，这个维度假设知识图谱是不完整的——即它是理想知识图谱中的一个样本——然后会问这个样本是如何出现偏差的。偏差也许会出现在数据中，途中或者推理过程中。数据偏差的实例包含不能充分代表真实世界特定部分的实体/关系，语言偏差不能充分反映特定语言的资源（比如标签和描述），社会偏差不能充分反映特定性别或种族等。与之相比，图偏差也许会造成高级定义从偏差数据中提取出来，语义定义不会包括非常规情况等。未意识到的偏差也许会产生不利影响；例如，对于与圣地亚哥紧密相关的事件和关联，如果我们的旅行知识图谱存在一个地理上的偏差——也许是由于创造城市所用到的资源，抑或旅行策划者的雇佣等——这会导致围绕圣地亚哥的旅行得到不均衡的提升（潜在地混合了未来的偏差）。典型性的度量包含那些知识图谱已知统计分布的比较，例如，把已知的人口密度与实体定位相比，已知的演说家的分布与语言分布比较等。另外一个选择是比较一般的统计学法则和知识图谱，比如Soulet et al.使用本福德法的（非）一致性来衡量知识图谱中的代表性。

**7.3 连贯性**

连贯性是指知识图谱是不是很好地符合——或者说一致于由图层面定义的规范语义和限制。

**7.3.1 一致性**意味着知识图谱摆脱了与所考虑的特定逻辑蕴涵之间的（逻辑/规范）矛盾。例如，在我们的知识图谱中可以定义，当与边结合时会产生一个不一致现象，蕴含着属于两个不相交的集合和。更一般地，如果蕴含着否定情况，那么结合一个“不”状态，表2-4中的任何语义特征都可以产生不一致现象。一致性可以由知识图谱中发现的不一致数量来衡量，其进一步还可以细分为由每个语义特征识别的不一致数量。

**7.3.2 有效性**是指知识图谱摆脱了违背约束，譬如被图形表示法捕捉到（详见3.1.2）。比如我们指定一个图形，其目标节点最多只包含一个国家。接着，给定边，假定是的一个目标，此时我们就会违背约束。相反地，即使我们在知识图谱中定义了类似的基数限制，也不一定会导致不一致，由此，在没有UNA的情况下，我们首先应该推断和涉及相同实体。有效性一个最直接的度量就是计数每个约束的违背数量。

**7.4 简洁性**

简洁性以一个简练、易理解的方式仅展示相关内容（避免“信息过载”）。

**7.4.1 简练性**是指避免出现与域无关的模式和数据元素。Mendes et al.区分了内涵简练性（图层面）和外延简练性（数据层面），其中前者涉及数据不包含冗余图元素（属性、类别、图形等）的情况，后者不包含冗余的实体和关系。例如，在我们的知识图谱中，包含在中的事件都关联在智力的旅行，这会影响知识图谱的外延简练性，对于给定的域也可能会返回不相关的结果。一般说来，简练性可以由关联于域的属性、类别、图形、实体、关系等的比例来度量，也就是说需要一个黄金标准或是技术来评估域相关性。

**7.4.2 表示简练性**涉及内容在知识图谱中展现的简练程度，这也许会重复地内涵或外延。例如，有服务于相同目的地两个属性flight和flies to，它们会消极影响典型简练性的内在形式，同时有两个代表智利首都的节点（与其他节点均不相连）和会影响典型简练性的外延形式。表示简练性的另一个例子是使用不必要的复杂模型结构，例如不必要的具体化，或者在元素的顺序不重要时使用关联列表。尽管表示简练性很难评估，但我们依然可以使用冗余节点数来衡量。

**7.4.3 易理解性**是指简化人类用户可以理解的、不模糊的数据，其中至少包含人类可读标签和描述的条款（最好是几种不同语言）以便于理解所描述的内容。返回表1，尽管节点和用来确保事件标识符的独特性，它们也应该会与诸如和这样的标签有联系。理想情况下，人类可读信息足够消除一个特定的节点，譬如将描述和联系起来以将其从城市紧密联系中消除。易理解性的度量包含人类可读标签和描述的节点数量的比例，这些标签和描述的独特性，语言支持等。

**7.5 其他质量维度**

为了一般性地应用到知识图谱中，我们已经讨论了一些关键的质量维度。更深层次的维度也许与特定领域，特定应用或特定的图数据模型中的上下文语境有关。要了解更多细节，我们提供了Zaveri et al.的综述和Batini、Scannapieco的书。