深交所知识图谱实践

为提高公司管理部门“科技监管”水平，立足于为上市公司监管提供数据支持服务，深交所自主研发了知识图谱一期工程项目，并于2018年7月上线。基于知识图谱，可以查询任意多个主体之间的关系网络，如企业与企业、企业与自然人、自然人与自然人之间的关系，并支持多个目标主体之间n度以内路径的自动探寻，可以辅助用户在多源异构的大数据中发现异常点。

证券期货业金融科技研究发展中心（深圳）（以下简称“金融科技中心”）是经证监会批准，由深交所建设运营，以服务行业为导向，聚焦金融科技创新发展的行业公共研究平台。为了更好地推进监管科技工程项目，深交所依托金融科技中心于2018年开展了《面向上市公司监管的知识图谱构建与应用研究》课题研究。本文对深交所知识图谱的内外部数据融合、本体构建及部分应用进行介绍。

1.内外部数据融合

知识图谱构建过程中需要对内部数据和外部数据进行融合，并考虑数据的时间维度。内外部数据融合包括数据选取与准备、数据预处理、实体对齐、数据聚合。实体对齐是数据融合中的关键技术，分为企业实体对齐和自然人实体对齐。

企业实体对齐可以通过名称、工商注册号、组织机构代码等进行精确对齐，也可以通过名称相似度、地址相似度等进行模糊对齐。相比于企业实体对齐，由于自然人名字同名现象普遍存在，自然人实体对齐更为困难。自然人的实体对齐分为精确对齐和模糊对齐。精确对齐除了基于证件号码外，还可以基于公司关联关系进行对齐，即两个有关联可达路径的公司含有多个同名自然人实体，则将这些自然人实体进行对齐。模糊对齐可以基于公司属性信息进行融合，即如果具有名称相似度、地址相似度或者电话号码相同的两个企业具有同名自然人实体，则将这些自然人实体进行对齐。在实体对齐中，可以采用多种相

似度算法计算相似度，如加权编辑距离等，实体对齐中的阈值可以调节。

2.本体构建

本体是描述客观世界的抽象模型，以形式化方式对概念及其之间的关系给出明确的定义。本体的构建依赖领域知识，目前构建出的本体包括自然人、公司、行业、地域、概念板块、企业主营产品类型等。本体反映的概念更为抽象，通过本体，可以描述知识图谱的数据模式，并有效地发现知识图谱中的不同实体之间的隐含关联关系。

3.知识图谱在上市公司监管中的应用

知识图谱在上市公司监管中的应用很多，本文重点介绍基于知识图谱的资本系挖掘以及风险事件传导。

（1）资本系挖掘。资本系的说法源自于财经媒体对“一控多”现象的关注。“一控多”自1999年之后大量涌现，成为境内证券市场普遍存在的现象。基于股权等信息可以挖掘出上市公司所属资本系以及资本系核心的实际控制人等信息。

通过知识图谱，可以直观地展现上市公司与资本系核心及资本系内其他公司的关联路径，这些信息可以有效辅助日常监管、公司治理、上市公司风险排查等工作。同时，还可以结合具体的监管业务需求对资本系进行全面且多方位的解读。由于资本系的自动命名技术还有待提高，为了更好地结合业务需求，系统应支持手工配置资本系的名称。资本系结果评价对于资本系挖掘算法的改进非常有必要，但资本系结果评价也存在一些难点，比如，需要花费较多的时间来评价资本系挖掘算法。

（2）风险事件传导。知识图谱由大量的节点与边组成，风险事件可以通过节点与边在知识图谱中传导，研究基于知识图谱的风险事件传导模型算法、路径及关键节点可以有效地防范上市公司风险。风险事件传导分析离不开领域本体的支持，包括上市公司所在行业、地域、产品、概念等信息。这些信息有助于知识图谱中的实体泛化，进而更准确地分析风险事件的传导过程。风险事件传导需要考虑的因素很多，比如相同的持股比例对于上市公司和非上市公司的差异化影响、不同公司对于风险的承受不同能力、不同类型关系的权重设置、传导衰减因子计算等。

另外，产业链数据可以将知识图谱中的不同实体关联到一起，反映出企业之间的财务及业务关系，是风险事件传导分析的重要环节。通过风险事件传导分析，输出受影响的上市公司名单，可辅助上市公司风险排查。