



自动化构建企业知识图谱与股权穿透 分析研究

学生 周展鹏

指导教师 金耀辉

2020年9月



上海交通大学

SHANGHAI JIAO TONG UNIVERSITY

目录 Contents

1

问题定义

2

图谱定义

3

系统构建

4

实体统一

5

个人总结



上海交通大学
SHANGHAI JIAO TONG UNIVERSITY

1

问题定义

2

图谱定义

3

系统构建

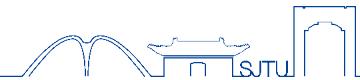
4

实体统一

5

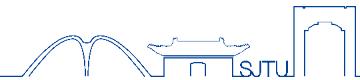
个人总结





问题定义

- Google于2012年提出知识图谱的概念。
- 知识图谱（Knowledge Graph）以结构化的形式描述客观世界中的概念、实体及其关系。
- 根据应用场景的不同，知识图谱可以简单分成两种
 - 百科全书知识图谱
 - 领域知识图谱
- 相比于百科全书式的知识图谱，有关领域知识图谱的研究并没有非常丰富，但是对于解决特定领域内的问题来说，领域知识图谱常常提供很多方便。



问题定义

- 金融机构需要了解企业于关联方的各种关系，洞察企业的风险传导、异常交易等信息。基于这样的需求，构建包含企业基本信息和企业投资关系的企业知识图谱是很有必要的。
- 然而构建知识图谱还存在很多挑战
 - 企业名称、高管名称存在一词多义、一义多词的奇异性，如同一企业简称、全称、历史名称和不同企业高管同名不同人的情况。事实上我们更加关注，不同企业高管同命不同人的情况，因为相比于企业名称，自然人姓名更易重复。
 - 数据量巨大且稀疏，构建图谱和依据图谱进行推理都需要花费大量时间。
- 最后利用已构建的图谱，进行股权穿透分析。

1

问题定义

2

图谱定义

3

系统构建

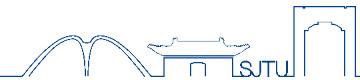
4

实体统一

5

个人总结

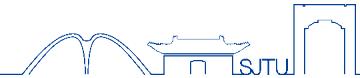




图谱定义

节点属性

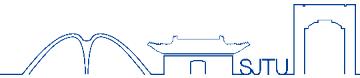
- 为了处理稀疏的企业数据，这里我们借用了广图的概念定义部分图谱。
- 广图可以用下面的简单数学符号表示：
 - $G = (N_1, E, N_2)$
 - 这里的 N_1 代表了某类实体或节点， N_2 则代表了某一类特征值的存在，连接了实体和特征值的关系 E ，这种关系往往标注有关节点特征的信息。
- 例如，小米科技有限责任公司的公司类型属于有限责任公司，那么在图谱中就会相应地建立（小米科技有限责任公司的公司ID，的公司类型是，有限责任公司）三元组来表示这类信息。
- 同时，使用广图增强了图谱的可扩展性，给图谱的每一个企业节点添加新的属性时，只需要在原有基础上创建三元组即可。



图谱定义

节点之间

- 直到刚刚，我们定义的图谱都属于一个企业的attribute-value pair。
- 图谱中，还应该出现**企业和企业之间的关系**：
 - 例如小米科技有限责任公司的法人代表时雷军，那我们就需要定义类似（小米科技有限责任公司的ID，的法人代表是，雷军的ID）的三元组。
 - 同时，企业与企业之间，企业与自然人之间也存在投资关系，我们也需要相应的定义，例如（公司A，投资，公司B）即代表公司A投资了公司B，（自然人A，投资，公司C）即代表自然人A投资了公司C。

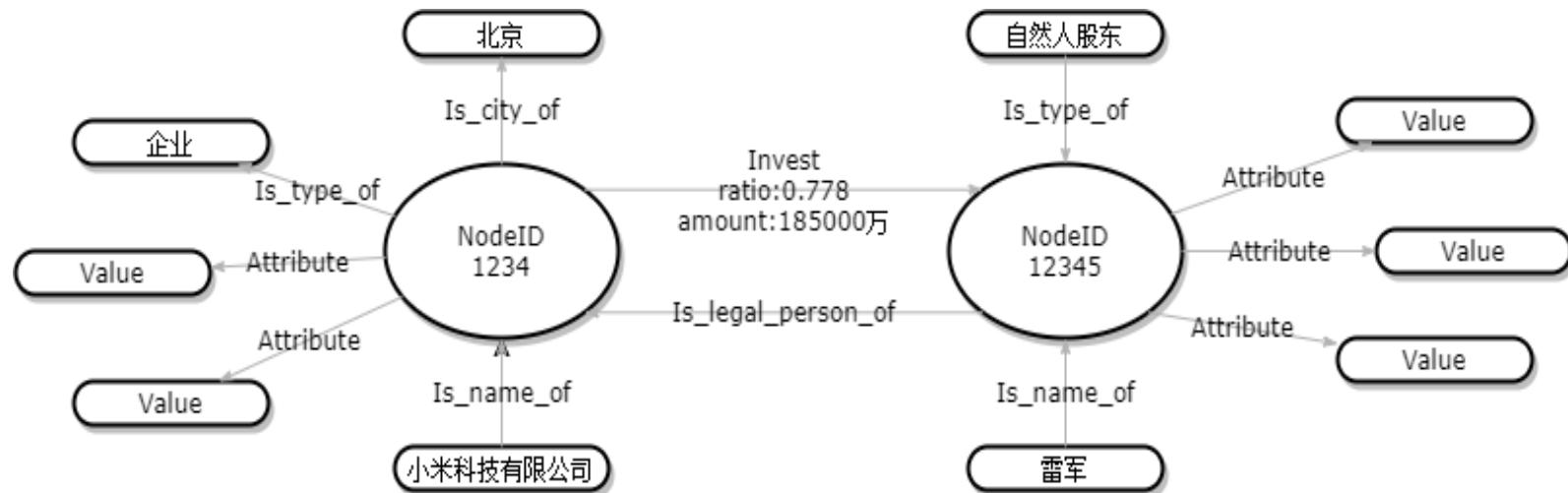


图谱定义

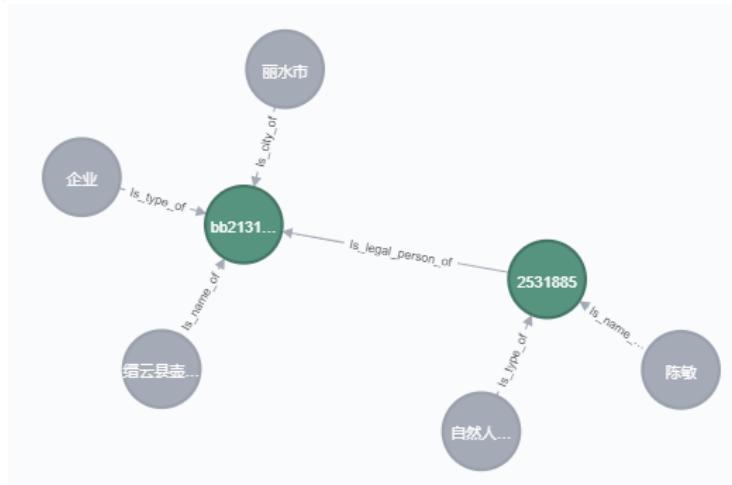
完整数学定义

- 总的来说，企业知识图谱由两部分子图构成， $C = (O, N)$ ：
 - $O = \{NID, \{I, L\}\}$: NID (NodeID) 代表企业的唯一ID或者投资人的唯一ID，而I(Invest)代表企业和投资人，或者企业和企业之间存在的关系。同时，每一条投资关系中都存在两种属性，(Ratio, Value)，分别代表投资关系重大控股比例和实缴金额，最后L(Legal Person)代表企业和投资人之间存在的法人代表关系。
 - $N = \{NID, \{A, V\}\}$: $\{A, V\}$ 代表对于企业唯一ID或者投资人唯一ID而言的一组属性值关系，例如企业名称，企业注册城市，投资人类型等等。

- 图谱的简单示意图



- 图谱的简单Demo图



目录 Contents

1

问题定义

2

图谱定义

3

系统构建

4

实体统一

5

个人总结

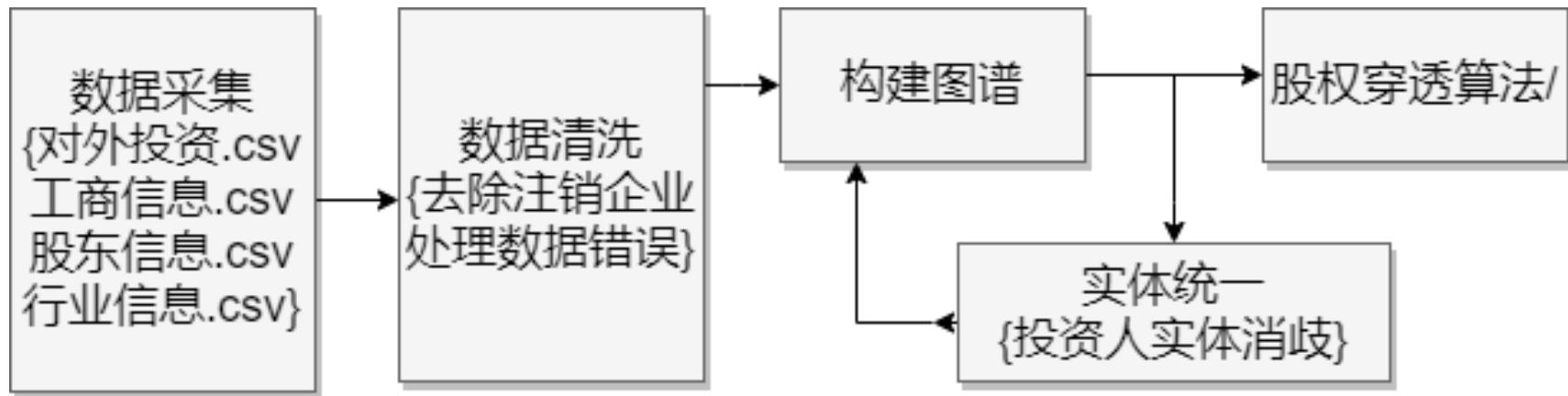


上海交通大学
SHANGHAI JIAO TONG UNIVERSITY

系统构建



- 定义图谱之后，即可进行图谱的构建。
- 下图是自动化构建企业知识图谱系统的基本组件，包含数据采集、数据清洗、构建图谱、实体统一和股权穿透算法等五个基本组件。



- 接下来，将会对这五个组件做具体的介绍，其中实体统一将会放到下一个章节单独介绍。

系统构建（数据采集）



- 构建图谱的主要信息来自国家企业信用公示系统。
- 根据国家的法律法规，该信息渠道完全合理合法。
- 我们将从该系统中收集的数据分成了4张表：
 - 对外投资表
 - 股东信息表
 - 工商信息表
 - 行业信息表
- 其中对外投资表和股东信息表主要提供了企业与企业之间、企业与自然人之间的投资关系的信息，而工商信息表提供了企业的基本信息，最后行业信息表主要提供了该企业的经营范围的信息。

系统构建（数据清洗）



- 原始数据的质量问题主要是
 - 全空数据条目或关键字段缺失
 - 数据条目重复
 - 字段数据不合理，如，“股东类型”：“江苏省”
- 因此针对这类普遍存在于四张表的问题，我们进行了相似的操作：关键字段缺失条目去除，重复数据去除和不合理字段去除。
- 同时数据中还有一类特殊的问题，即数据中包含已注销企业的信息。针对这类问题，我们根据从国家企业信用系统中得到的对应企业的企业状态，即可直接删除企业状态异常的企业对应的数据条目。
- 大部分表单在经过清洗之后，仍然有99%的数据得到了保留。

系统构建（构建图谱）

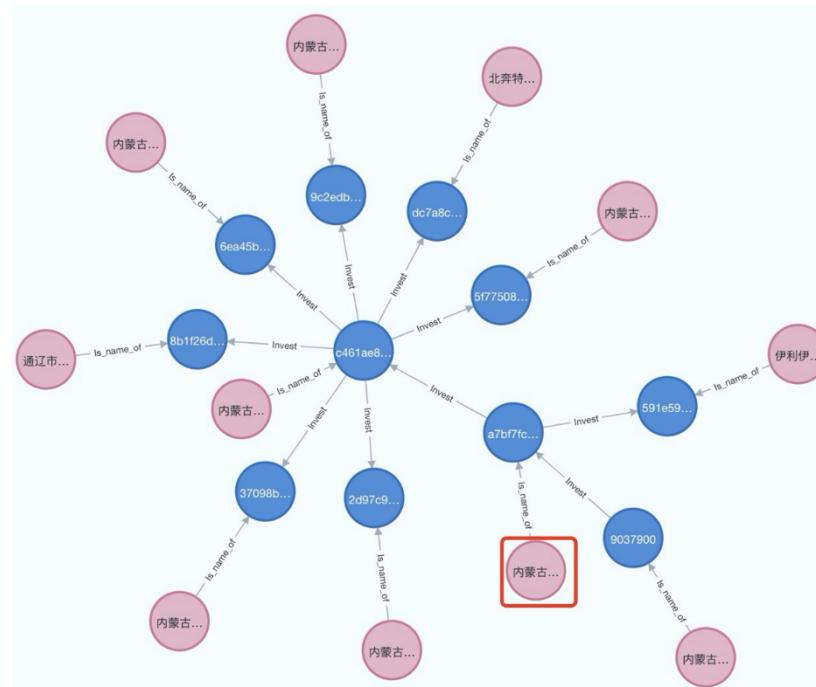


- 利用清洗好的数据和提前定义的图谱模型，利用neo4j图数据库的原生工具，我们能够实现在10min左右的时间内构建拥有上千万节点和关系的图谱。
- 从系统构建的示意图中，可以看到实体统一和构建图谱两个组件之间形成了一个环状结构，即可以经过不断迭代丰富或者加强图谱的准确性。
- 在问题定义的部分，我们提到在实际构建图谱的过程中存在同名高管不同人的情况，因此需要实体消歧的算法帮助我们分辨。
- 利用原始数据第一次构建的图谱中，我们将所有同名者均视作不同人，并且分配唯一的NodeID，随后在后面实体统一算法中，将很有可能是同一个人的同名自然人找到，并在图谱中进行节点合并，即完成了初步的迭代，这样经过多次迭代的图谱，就可以得到很高精度的实体消歧。

系统构建（股权穿透）



- 股权穿透完全依赖于以构建的图谱。利用neo4j的原生工具即可比较直观的返回企业的股权穿透情况。
- 例如，我们对“内蒙古乳业技术研究院有限责任公司”进行股权穿透，即可得到如下的穿透结构图。



1

问题定义

2

图谱定义

3

系统构建

4

实体统一

5

个人总结



实体统一



- 在利用原始数据直接构建的图谱中，我们将每一个同名自然人都算做不同人，因此我们的算法从实体消歧变为实体统一，即，找到那些同名节点需要合并。
- 我们将实体统一的算法，分为两个部分：
 - 基于地域+行业的合并
 - 基于投资关系的合并
- 接下来，我们将分别对这两个部分进行介绍。

基于行业+地域的合并



- 根据我们的常识可以直观地得到，投资了同行业同区域的企业同名自然人有很大可能性为同一人。
- 因此我们利用工商信息表中的地域代码和行业信息表中的行业代码对所有企业进行3级行业（行业可以分为一级行业、二级行业和三级行业，数字越大代表行业细分越精确）和3级地域划分（即省市县三级行政区域划分），同行业同区域的企业归为一类，类内重名则视为同一人。

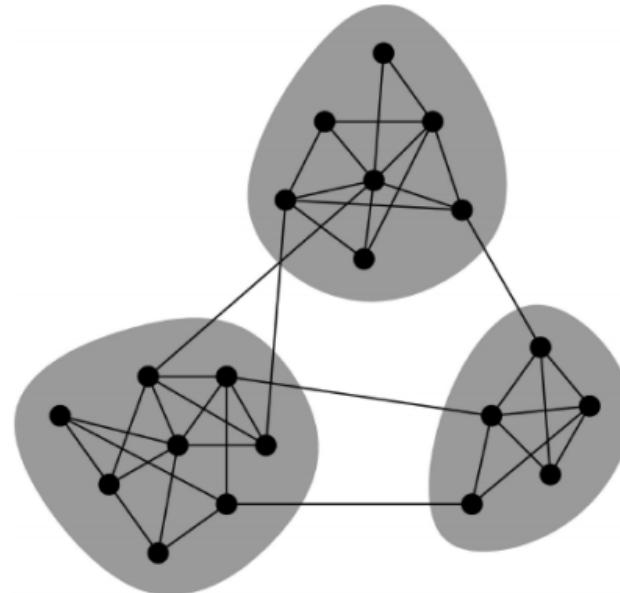
合并数目	一级地域/省	二级地域/市	三级地域/县
一级行业	1269402 (20%)	851612 (14%)	571330 (9%)
二级行业	737658 (12%)	512242 (8%)	358905 (6%)
三级行业	631417 (10%)	516242 (7%)	320834 (5%)

- 从表中可以看到，基于三级行业和地域的合并是最保守的。

基于投资关系的合并



- 各个节点（企业法人，自然人）之间存在相互投资关系，根据相互投资关系对节点进行聚类，类内节点具有紧密的相互投资关系，而类间的联系较少。类内重名则视为同一人。这里的聚类方式采用了 neo4j 算法数据库中的 Weakly Connected Components 算法，即通过判断弱连接来寻找企业知识图谱连接更紧密的类。



1

问题定义

2

图谱定义

3

系统构建

4

实体统一

5

个人总结



上海交通大学
SHANGHAI JIAO TONG UNIVERSITY

总结



- 聚焦构建企业知识图谱，创新性地运用了广图的概念，提高了图谱的可扩展性，同时为未来面对更加稀疏的非结构化数据提供了解决思路。
- 同时，我们开创性地构建了构建企业知识图谱的自动化系统，希望引入最少的人工干预构建信息密度高、信息精度高的企业知识图谱。
- 仍然存在很多不足：
 - 实体消歧目前主要依靠规则判断，这样的方法具有一定作用，但是其准确度很难提升，未来可以结合机器学习的方法提高实体消歧算法的性能和准确度。
 - 数据清洗的过程中，我们讲很多关键信息缺失的数据条目直接丢弃可能会造成数据的浪费等。

致谢



- 感谢我的导师金耀辉老师。本研究是在金老师的悉心栽培和精心指导下完成的。感谢金老师一年以来给我前进的动力，以及在学术、生活等各个方面给予我的无微不至的关怀和照顾。金老师严谨的学术态度、开朗大度的胸怀和高瞻远瞩的目光给我留下了难忘的印象。老师的教诲让我终生受益，在项目完成之际，向金老师表示深深的谢意！
- 感谢实验室的师兄和同学在研究工作和日常生活中给予我的关心和帮助。
- 最后，再次向所有我学习道路上给我关心和照顾的老师，亲人和朋友表示最诚挚 的谢意！

谢谢！

