项目编号：T030PRP36102

上海交通大學

本科生研究计划（PRP）研究论文

（第 36 期）

论文题目：自动化构建企业知识图谱与股权穿透分析研究

项目负责人：金耀辉 学院（系）：电子信息与电气工程学院

指 导 教 师：金耀辉 学院（系）： 电子信息与电气工程学院

参与学生：周展鹏 密西根学院 F1837009

项目执行时间：2019 年 07 月 至 2020 年 09 月

摘要

2012 年 5 月 16 日 Google 为了提升搜索引擎的性能发布了知识图谱（Knowledge Graph）。近年来，随着大数据时代的到来，将海量的信息结构化、知识化，构建由实体和关系构成的语义网络——知识图谱，已经成为大势所趋。根据应用场景的不同，又可以将知识图谱分成百科全书式知识图谱和领域知识图谱。随着搜索引擎性能的不断提升，通用知识图谱的研究也在不断深入，领域知识图谱却往往被忽略。然而对于很多特定场景问题，通过构建领域知识图谱然后进行推理却是解决特定场景问题的有效办法。因此，本文通过构建企业领域知识图谱探究构建特定领域知识图谱的自动化构建方法，建立了包含数据清洗、图谱定义、实体统一算法的一整套企业知识图谱构建的自动化系统。本文的主要贡献如下：

1. 以构建企业知识图谱为例，应用数学方法定义了企业知识图谱的一般模型。该模型具有高扩展性，同时可以泛化至更多其他领域的知识图谱中。

2. 在企业投资关系中，企业高管名称存在一词多义，一义多词等现象。我们利用公司信息里的地域代码和行业代码、通过企业之间投资关系形成的弱连接组件等信息得到了判别不同企业的同名高管的重名消歧算法。

3. 结合上述图谱模型和实体消歧算法，设计出一个应用于企业领域的图谱构建系统，并将该系统成功应用到构建知识图谱的过程中，使用 Neo4j 成功构建企业领域的知识图谱。

4. 通过已构建好的企业知识图谱，我们解决了金融领域的股权穿透问题，即输入一个公司名称，我们可以返回上游（投资对象）、下游（被投资对象）的股权信息。

关键词：知识图谱，实体消歧，数据清洗，股权穿透

Abstract

On May 16, 2012, Google released the Knowledge Graph in order to improve the performance of its search engine. In recent years, with the advent of the era f big data, it has become an irresistible trend to structure the knowledge-based massive information and build a semantic network composed of entities and relationships, Knowledge Graph. According to different application scenarios, Knowledge Graph can be divided into Open-domain Knowledge Graph and Domain-specific Knowledge Graph. With the continuous improvement of search engine performance, the research on Open-domain Knowledge Graph is also deepening, but Domain-specific Knowledge Graph is often neglected. However, for many scenarios-specific problems, it is an effective way to solve the scenarios-specific problems by building Domain-specific Knowledge Graph and then reasoning. Therefore, this paper explores the automated construction method of building Domain-specific Knowledge Graphs by building enterprise domain knowledge graph, and establishes an automated system of building a whole set of enterprise knowledge graph including data cleaning, graph model definition and entity unified algorithm. The specific work of this paper can be summarized as follows:

1. Taking the construction of enterprise knowledge graph as an example, the general model of the enterprise knowledge graph is defined by applying mathematical method. The model is highly scalable and can be generalized to knowledge graphs in other fields.
2. In the enterprise investment relationship, the name of enterprise senior executives has the phenomenon of polysemy. We use the region code and industry code in the company information, the weak connection components formed by the investment relationship between enterprises and other information to get the name disambiguating algorithm to classify senior executives of different companies with the same name.
3. Combined with the above graph model and entity disambiguation algorithm, a automated graph-building system for enterprise domain was designed, which was successfully applied to the process of constructing enterprise knowledge graph.
4. Through the established enterprise knowledge graph, we solve the problem of enterprise holding in the financial field, that is, by entering a company name, we can return the information of 10 levels of equity of the investment object and the invested object.

Key words: knowledge graph, entity disambiguation, data cleaning, equity analysis

第一章 绪论

知识图谱(Knowledge Graph)以结构化的形式描述客观世界中概念、实体及其关系,将互联网的信息表达成更接近人类认知世界的形式,提供了一种更好地组织、管理和理解互联网海量信息的能力。而金融机构需要了解企业与关联方的各种关系，洞察企业的风险传导、异常交易等信息。因此构建企业领域的知识图谱将为解决金融问题提供一个有效的解决方案，这类企业知识图谱应当包含企业的基本信息和企业的投资关系信息，从而可以帮助我们解决很多金融领域内的存在的推理问题。

然而构建企业知识图谱仍然存在很多的挑战：

1. 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、历史名称等情况，不同企业高管同名不同人的情况。事实上，实体消歧是构建知识图谱中常见的难点，也是重点研究方向之一，在我们的研究中，通过使用更多的信息构建规则，通过已构建图谱的结构实现了效果良好的实体消歧算法，同样的思路也可应用到其他领域的图谱构建中。
2. 同时，企业数据非常庞大且稀疏，利用这样的信息构建上亿节点的图谱往往具有很大的困难，需要定义适合的图谱来处理这样庞大而稀疏的数据，同时支持上亿节点与关系图谱的高效构建。在研究中，我们通过定义 Broad Graph 实现了对稀疏数据的处理，同时利用图数据库 neo4j 的原生工具实现了高效构建上亿图谱的算法，这对未来其他领域知识图谱的构建也能够起到参考作用。

这篇论文将会介绍一种自动化构建企业知识图谱的系统，可以同时处理实体消歧和数据稀疏的问题，最后利用已构建的企业知识图谱，可以完成金融领域股权穿透问题的推理。接下来，本文将会首先从整体介绍图谱的定义和构建企业知识图谱所需要的系统。

第二章 图谱定义和系统构建

2.1 图谱定义

知识图谱由一条条知识组成，每条知识都可以称作一个三元组（Subject，Predicate，Object）。Subject 是一类实体，这类实体拥有唯一 ID。Object 可以是实体，也可以是某些特征值，例如一串字符或者一个数字。Predicate 描述了 Subject 和 Object 之间的关系，例如，（公司 A，投资，公司 B）就代表了公司 A 投资了公司 B，而（公司 A，位于，北京）则代表了公司 A 的所在地是北京，这里北京就属于一类特征值。在知识图谱中，我们可以把 subject 和 object 理解成为图的节点，而 predicate 是连接两个节点的关系。

为了简化问题的定义，本文使用了一种特殊的图谱定义，广图（Broad Graph）。广图可以用下面简单的数学符号表示：

$$G = (N_1, E, N_2)$$

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则是关系 E，这种关系往往会标注有关节点特征的信息。在企业知识图谱种，我们往往使用广图描述企业的基本信息。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。

使用广图来描述企业的基本信息的意义在于，现实中得到的企业工商信息数据往往非常稀疏，如果有关企业的某类数据存在缺失，就使用缺失值存贮在对应节点的属性中，那么我们将会在数据库中填充存储大量无意义的数据，而这样的无意义数据无疑大大降低了图谱的信息密度。同时，使用节点的属性值储存类似的信息限制了节点的可扩展性，例如，当需要为图谱中的每一个企业增加一个新的属性值时，我们就需要重新构建每一个节点和每一个节点拥有的属性列表。然而使用广图，面对相似的问题，我们只需要为每一个企业节点添加一个新的三元组来代表新的属性信息即可，这大大加强了图谱的可扩展性。

同时，在图谱中，我们还需要定义节点与节点之间的关系，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。

总的来说，企业知识图谱由两部分子图构成，$C = (O, N)$:

1. $O = \{NID, \{I, L\}\}$: 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。
2. $N = \{NID, \{A, V\}\}$: $\{A, V\}$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。

根据这样的定义，我们制作如下图（Figure 1）所示的示意图，值得注意的是，根据我们广图的定义，这里所有的姓名信息也都当作属性值作为一个三元组储存在图谱中。

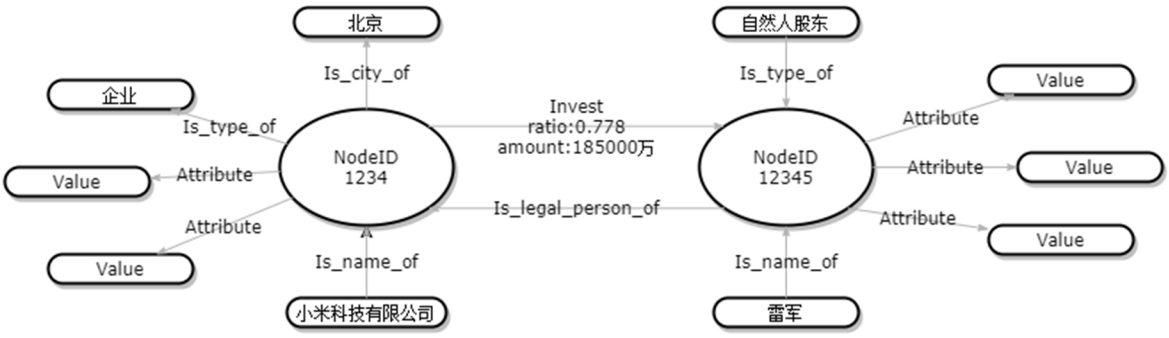

Figure 1 图谱模型示意图

同时，如 Figure 2 所示，根据我们从国家企业信用信息公示系统中采集到的所有数据，利用图数据库 neo4j，同样制作了一张拥有上千万节点的企业知识图谱，然而囿于篇幅限制，我们这里只截取了图中的两个节点。

图谱的定义和需求是相互适应的。在第一章绪论，这篇论文谈到了金融领域中常见的股权穿透问题，解决这类问题往往需要企业与企业之间、企业与投资人之间的投资关系，因此我们在图谱中也同样定义了企业与企业之间、企业与投资人之间同样可以存在投资关系。同时，在最开始的时候，我们同样定义了一个特殊的问题，即企业高管、投资人的重名问题，然而根据所采集到的信息来看，我们无法直接区分两个同名者是否为同一个人，因此我们需要额外的实体消歧算法来辅助我们解决这类问题，

例如，直观地来看，投资了同一地域的相同产业的不同公司的两个同名投资人有很大可能是同一个人，因为他们投资的企业位于同一个城市，同时他们关心的产业也属于同一方向。类似的例子还有很多，因此这就需要我们在图谱的定义中添加额外的信息，经过我们的研究后，我们总结了几种可以帮助我们进行实体消歧，尤其是消除企业高管重名的歧义，的几种信息，如投资人类型，企业所在地域还有行业信息等。

介绍了图谱的定义后，接下来在章节 1.2，本文将会具体介绍自动化构建企业知识图谱的系统构造。

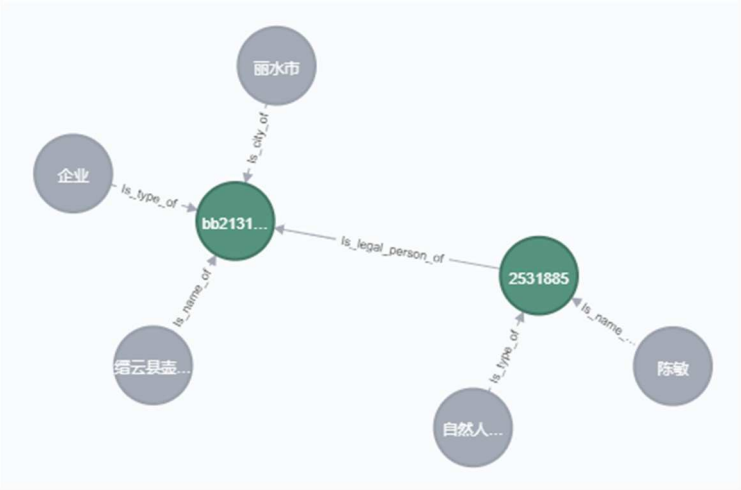

Figure 2 利用图数据库 Neo4j 和从国家企业信用公示系统采集的数据构建的企业知识图谱（部分）

2.2 系统构建

下图（Figure 3）介绍了自动化系统的基本组件。它拥有 5 个基本组件，分别是数据采集、数据清洗、构建图谱、实体统一和股权穿透算法。

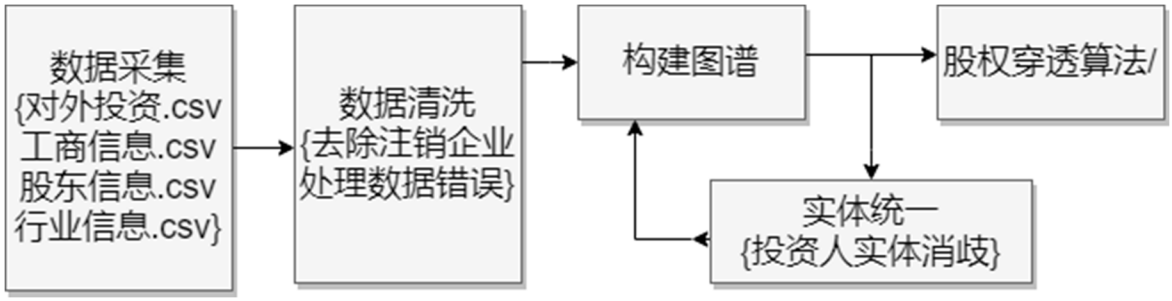

Figure 3 自动化构建企业知识图谱系统的基本组件

2.2.1 数据采集

构建系统的主要信息全部来自国家企业信用公示系统。国家企业信用公示系统提供全国企业、农民专业合作社、个体工商户等市场主体信用信息的填报、公示和查询服务。根据《中华人民共和国政府信息公开条例》和《企业信息公示暂行条例》，我们从该系统所收集的信息完全合理合法。我们将从国家企业信用公示系统中采集的数据分成了四张表格，分别是对外投资表、工商信息表、股东信息表和行业信息表。其中对外投资表和股东信息表主要包含的是企业与企业之间、企业和投资人之间投资关系的信息，而工商信息表中主要包含企业的基本信息，如企业 ID 等，这三张表格主要用于构建企业知识图谱，而行业信息表包含企业的行业信息，这对我们做实体

统一时有很大帮助。

然而囿于技术限制和公示系统所提供信息的质量问题，所采集到的信息大多非常稀疏，经过了简单的处理，信息中仍然存在较多质量问题，因此在自动化系统构建中，我们引入了数据清洗的组件，帮助我们提升数据的密度和质量。

2.2.2 数据清洗

首先，采集到的原始数据的质量问题可以分为以下几类：

1. 全空数据条目以及关键字段缺失条目；
2. 数据条目重复；
3. 字段数据不合理，例如，“股东类型”：“江苏省”。

因此，针对这几类普遍存在的问题，数据清洗的过程包含了：缺失处理，去重和不合法字段去除。

同时，数据中还有一类特殊的问题需要处理，即数据中包含已注销企业的信息。然而从国家企业信用公示系统中，我们可以得到对应企业当前的企业状态，依靠企业状态所提供的数据，我们可以直接删除企业状态为注销等异常状态的企业的数据条目。

最后，如下表所示（Table 1），以工商信息表为例，经过清洗，仍然有 997 万条、占原始数据 99%左右的数据条目得以保留，然后使用清洗得到的数据，即可构建拥有上千万节点和关系的原始图谱。

	保留数据量	**占比（%）**
原始数据	10085149	100.00
删除企业状态异常	10074030	99.89
删除企业 ID 缺失	9981387	98.97
删除重复数据	9970597	98.86

Table 1 工商信息表的数据清洗过程和保留数据量

2.2.3 构建图谱

根据清洗得到的数据和预先定义的模型，使用 neo4j 的原生工具即可帮助我们迅速完成上千万节点与关系的大型图谱的快速构建。然而仅通过当前数据得到的图谱并非正确，因为我们还没有区分同名企业高管是否是同一个人，因此当前我们将所有重名者均当作不同的人，并且分配了不同的 NodeID。结合已构建图谱的信息和其他表单提供的信息，利用实体消歧算法，我们可以得到需要合并的 NodeID 节点，进而更新我们当前的原始图谱，最后经过多轮迭代，即可得到一张信息准确度较高的企业知识图谱。

2.2.4 实体统一

目前，实体统一主要使用原始图谱中存在的弱连接组件信息、行业信息表中提供的行业信息和工商信息表中提供的地域信息构建规则进行消歧，该算法可以达到一定的效果，具体细节将会在第三章中具体阐述。

2.2.5 股权穿透算法

股权穿透基于利用已构建的企业知识图谱，直观地反应目标企业的股权状况。而这一部分的内容主要使用 neo4j 图数据库中自带的 cypher 语句返回所需要的股权穿透信息。

选择目标企业名称，如下图所示（Figure 4），在 neo4j 中提供的可视化工具中输入 cypher 语句，即可完成对目前企业股权信息的检索。

```
MATCH q = (a:Name{Name:" [target_enterprise_name] "})-
[:Is_name_of]->(b:NodeID)
MATCH p = (:Name)-[:Is_name_of]->(:NodeID)-[:Invest*1..]->
(b:NodeID)-[:Invest*0..]->(:NodeID)<-[:Is_name_of]-(:Name)
RETURN q,p
```

Figure 4 股权穿透算法所使用的 Cypher 语句

例如，想要对“内蒙古乳业技术研究院有限责任公司”进行股权穿透，在输入了对应的 cypher 语句后，我们即可得到如下图（Figure 5）所示的股权穿透结构图。

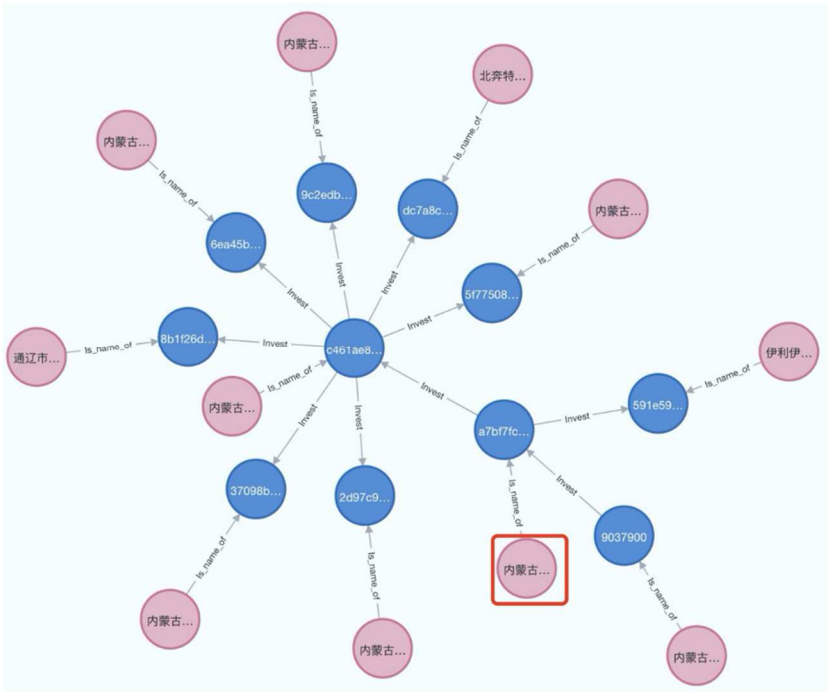

Figure 5 内蒙古乳业技术研究院有限责任公司股权穿透图

2.3 小结

在这样一个构建领域内知识图谱的图谱定义中，我们希望图谱具有足够的可扩展性，因此我们定义了广图的概念，同时，利用广图我们可以实现对稀疏数据的处理。并且，我们认为广图的概念可以应用到更多的领域内知识图谱的构建中，可以泛化到足够多的应用场景中。

同时在系统设计中，我们希望引入最少的人力，在人为干预较少的情况下，高效准确地搭建企业知识图谱，同时我们希望图谱的信息密度和精度可以反复迭代，所以我们在系统设计中引入了一个闭环设计，这样我们的图谱就可以反复提升质量。

最后，通过一个简单的应用领域知识图谱的例子，即股权穿透问题，我们可以发现依靠图谱我们可以更加轻松、高效、直观地回答很多特定领域内的知识型推理问题，相信未来依靠知识图谱，我们也会挖掘更多的应用。

第三章 实体统一

本章节，我们将详细介绍对企业高管、法人代表等同名自然人进行消歧的实体统

一算法。因为此前构建原始图谱的过程中，我们将每一个同名的自然人都当作不同的两个人，并且分配不同的 NodeID，因此我们的实体消歧算法变成了实体统一算法，即找出哪些同名节点需要合并，即将两个同名的不同自然人合并成为同一个人。根据使用信息的不同，我们将实体统一算法分为两个部分：基于地域+行业的合并和基于投资关系的合并。

3.1 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的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。

使用数据：

1. “股东信息表”11365440 条，用于合并节点；
2. “工商信息表”9970587 条，用于查询区域代码；
3. “行业分类表”9548555 条，用于查询行业代码。

股东信息表中能够查询到区域和行业代码的纵沟 6260374 条，在所有数据中占比 55.08%，去除掉的包括：无法查询到相应区域/行业代码、查询到代码不符合规范、投资人不是自然人（只对自然人重名问题进行消歧）等的数据条目。

依靠以上的数据，我们得到了如下表（Table 2）所示的合并数目表：

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%）

Table 2 九级划分合并数目表

从表中可以看出，使用三级行业和三级地域划分出来的企业类别进行实体统一所合并的节点占比最少，同时也是最保守的一种划分。

3.2 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，

3.3 小结

通过使用从商业用的企业信息查询系统中返回的数据来看，我们发现应用这两种信息得到的实体统一结果具有很高的精度。然而使用规则判断仍然存在较大的可能判断失误，需要介入人工对合并之后的节点进行反复验证，未来我们可能应用贝叶斯分析的方法，将原有的规则判断转变为概率分析，对高概率为同一人的节点进行合并，这样的方法在面对未来越来越多的数据时，可靠性更强。

第四章 结论

随着互联网的迅速发展，可以通过互联网获得的数据越来越庞大，海量的非结构化数据成为了人们的宝贵资源，但是如何利用这些资源成为了人们的难题。自从 2012 年 Google 提出知识图谱的概念后，越来越多的研究在专注于将海量的数据构建成图谱，但同时很多有关领域内知识图谱的构建的研究还处于起步阶段。本文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，同时在系统中引入闭环，通过不断迭代图谱进而扩展图谱的信息容量且提高图谱的信息质量。

最后，我们仍然有很多不足，例如，实体消歧目前主要依靠规则判断，这样的方法具有一定的作用，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。同时，在数据清洗的过程中，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，相信未来在这个方面我们会有更好的处理。

致谢

首先感谢我的导师金耀辉老师。本论文的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，在论文完成之际，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！