面向法规智能的消防规范图谱构建及应用初探

林佳瑞[1]，廖盘宇[2,1]

（1 清华大学土木系，北京 100084；2 中山大学土木工程学院，广州 510275）

【摘　要】消防规范对保障生命财产安全、推动韧性城市发展具有重要意义。在人工智能蓬勃发展的今天，传统非结构化文本形式的规范已难以满足合规审查、关联分析等法规智能应用场景。针对该问题，本研究提出一种基于 XML 的规范文本结构化与规范图谱构建方法，实现了非结构化文本向结构化 XML 的自动转换与基于 neo4j 的规范图谱生成。验证表明方法可行，可支持规范关联检索、冲突分析以及设计审查等多个法规智能场景，具有重要理论及应用价值。

【关键词】法规智能；消防审查；图数据库；关联检索；文本分析

1　引言

随着城市化高速发展，我国消防安全形势日趋复杂、严峻。自改革开放以来，我国经历了人类历史上最大规模、最快速度的城市化过程，城市规模、人口爆炸式增长，大量高层及超高层建筑、医疗及商业综合体等作为城市的核心构成，同时也成为重大消防安全事故的主要发生场所[1]。根据应急管理部年鉴数据，2017 年全国共接报火灾 28.1 万起，伤亡 2271 人，直接财产损失 36 亿元，其中城市火灾 13.1 万起，约占全国火灾总数的 47%。2018 年，我国先后发生“8.25”哈尔滨酒店火灾等重大事故，给人民生命财产安全带来巨大损失。国际上，2017 年 6 月 14 日英国伦敦“格兰菲尔塔”火灾致 81 人死亡，2019 年 2 月 12 日印度新德里市中心一酒店失火伤亡 22 人，教训惨痛。有关统计表明，尽管火灾起数及人员伤亡得以控制，但直接经济损失仍在逐年增长[1]。

消防规范是有关消防理论、实践经验的进一步凝练和总结，是建筑设计、建造乃至使用的重要依据，对保障人民生命财产安全、韧性城市发展具有重要意义。历经七十余年发展，我国已形成了系统完善、体系庞大的建筑领域标准体系。作为城镇安全的重要保障，消防规范与建筑领域其他规范相互引用、相互支撑，形成了复杂的标准规范关联网络。然

【基金项目】国家自然科学基金(51908323)，清华大学自主科研计划(2019Z02UOT) ，北京市自然科学基金(8194067)

【作者简介】林佳瑞(1987–)，男，助理研究员。主要研究方向为建筑信息化、智慧建造、数字防灾。E-mail：lin611@tsinghua.edu.cn

而，海量复杂的规范体系也为建筑设计、建造及管理人员带来了极大的挑战，如何高效的检索、查询和利用消防规范中蕴含的知识和经验已成为当前亟待解决的难题之一。

近年来，以深度学习为代表的人工智能（AI）蓬勃发展，融入和推动了各个行业的变革。法律 AI、法律法规数据库等的法规智能探索方兴未艾[2, 3]。尽管国内外建筑行业已在合规性自动审查[4]、多规合一图审系统等方面做了广泛的探索，但规范数据结构化、规范数据库构建等方面的研究、探索尚显不足。目前，有关规范仍以传统 PDF 文档、网页等形式存储和管理，检索和利用效率低，已成为消防乃至建筑领域的智能化水平和转型升级的瓶颈。

针对上述问题，本研究旨在以消防规范为研究对象，研究提出传统规范文档的结构化处理及图数据库构建方法，并结合关联规范检索、规范关联分析等场景探索分析规范图谱的应用场景，以期推动建筑领域法规智能的发展、提升行业信息化水平。

2 研究目标

如图 1 所示，本研究旨在提出一种从 PDF 格式消防规范构建消防规范图谱的方法，并探索其应用场景。具体包括三部分：1）建立基于 XML 的规范条文结构化表达并提出 PDF 规范的格式化处理方法；2）建立基于 neo4j 图数据库的规范图谱结构并，3）基于结构化 XML 规范数据建立消防规范图谱并探索潜在应用价值。有关研究将为建筑领域规范结构化处理、规范图谱构建、规范关联检索分析提供了方法和工具。

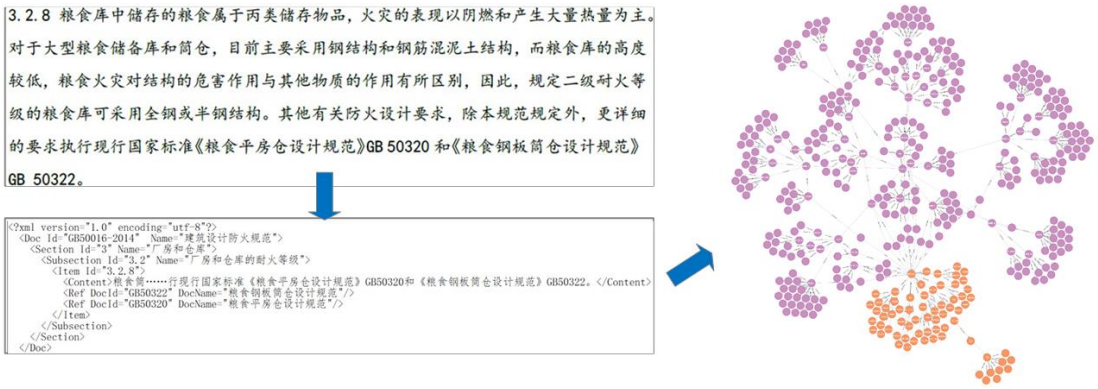

图 1 研究目标

3 规范文本结构化及消防规范图谱构建方法

当前我国有关规范的发布和公开仍然以 PDF 文件形式为主，这种形式虽然便于工程人员浏览和阅读，但缺乏章节条文关联信息，结构化程度低，不利于计算机处理和使用。如图 1 左上角所示，《建筑设计防火规范》GB 50016-2014 的 3.2.8 条文中规定了防火设计还需额外满足国家标准《粮食平房仓设计规范》GB 50320 和《粮食钢板筒仓设计规范》GB

50322。此处既隐含本条文属于规范的第 3 章第 2 节，也隐含当前条文与其他规范的关联关系。尽管工程人员可以从中了解有关信息，但这种格式处理起来并不方便。如果采用图 1 左下角所示的 XML 格式，则可清晰的表达规范和标准的章节层级结构及有关条文对其他规范的引用关系，从而极大地方便计算机的处理，也便于更准确地搜索内容、传输内容和描述事物。在此基础上，可进一步整合领域多部相关规范构建规范图谱，基于图数据库强大的关联查询能力实现规范关联检索及分析。

3.1 基于 XML 的规范条文结构化模型

为实现规范条文的结构化、规范化表达，研究首先建立了如图 2 所示的规范条文结构化模型。考虑规范特点，模型主要包括 4 类实体对象，分别是：规范文档、章、节、条文及条文引用。其中，规范文档(Doc)包括名称(Name)、编号(Id)、编制部门(CreatedBy)、颁布部门(IssuedBy)、施行时间(Date)等属性并可包括一系列章节，章(Section)及节(Subsection)主要包括标题(Name)和编号(Id)信息，且二者分别包含一系列的节和条文，其中章也可直接包含条文实体。最后，条文(Item)则包括编号(Id)、内容(Content)及规范引用信息(Ref)。其中规范引用信息主要包括引用的规范编号(DocId)、名称(DocName)以及章节、条文编号(SectionId、SubsectionId、ItemId)等信息。

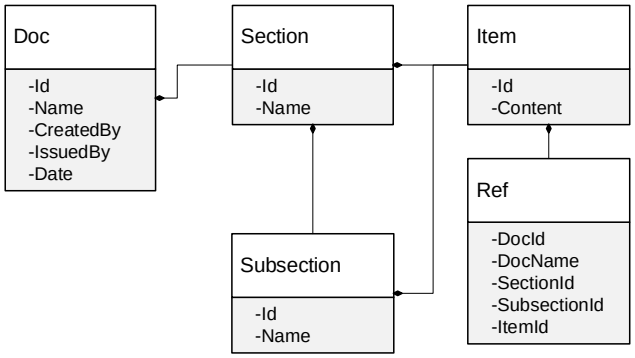

图 2　基于 XML 的规范条文结构化模型

3.2 规范条文结构化 XML 模型构建方法

为构建规范条文的结构化 XML 文档，本研究采用如图 3 所示的处理步骤，具体包括：文本提取、文本清洗、条文关联识别、XML 文档构建四个步骤。由于 python 提供了丰富的基础模块和便捷的编程环境，本研究采用 python 及有关模块作为基础实现规范条文结构化 XML 文档的自动构建。具体介绍如下：

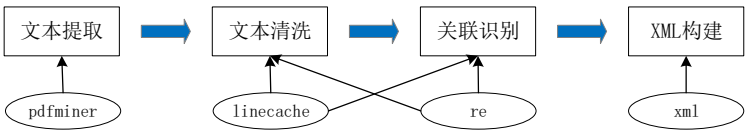

图 3　规范条文结构化 XML 文档构建方法

(1) 文本提取：首先，基于 python 模块库提供的 pdfminer 模块读取 PDF 文档信息，并逐页提取文本数据，并最终整合各页的文本数据，得到整体规范的文本文件。

(2) 文本清洗：接着，考虑到前述步骤从 PDF 文档提取的文本数据包括目录、空行、空格等无关内容，因此需对其进行进一步的清洗和处理。该步骤采用 python 提供的 linecache 模块和 re 模块完成。其中，linecache 模块可将文本文件内容缓存到内存中、提高文件的读取效率，re 模块则提供了正则表达式匹配操作，可基于用户自定义的规则匹配具有特定特征的字符串。基于这两个模块，首先通过字符串匹配识别目录、说明等内容，并将其删除，然后进一步去除空行、页眉、页脚等文本，最终，可得到清洗后的规范正文内容。

(3) 关联识别：同时，由于需要识别文本中的规范名称、编号以及章节编号、名称等信息，本研究通过分析有关规范及其章节的编号方式建立了相应的正则表达式。考虑本研究收集到的规范的特点，研究过程主要处理了采用“字母缩写 5 位数字编号-颁布年”形式进行规范编号和采用“章号.节号.条文号”形式进行章节和条文编号的规范文本数据。基于前述模式，可形成相应的正则表达式匹配有关信息。此外，鉴于规范通常包括主编部门、颁布部门、施行时间等信息，研究也采用文本匹配的方法对有关信息进行了提取。基于前述提取的信息，就可以建立规范各条文及其与规范章节的关联关系，并可从规范条文中提取引用的有关规范编号等信息，建立规范的引用关系。

(4) XML 构建：最后，基于前述提取的规范条文关联关系和引用关系，可逐步输出 XML 文档的各部分内容。该部分主要包括：规范基本信息生成，逐章、逐节输出条文信息以及输出条文对其他规范的引用信息等步骤。

3.3 基于 neo4j 的规范图谱构建方法

前述 XML 形式的结构化规范文档可清晰准确的表达单个规范各部分组成及其相互关系，但在多个规范及其条文的相互关联信息管理、查询方面仍存在明显不足。鉴于图数据库 neo4j 在关联关系查询、分析方面的优势[5]，本研究建立了如图 4 所示的规范条文的图数据库模型，并在此基础上构建了规范图谱。图 4 中，以节点的形式表达了规范文档、章、节及条文，各节点的属性基本与前述 XML 模型一致，并通过:refTo 关联关系表达条文对其他规范、章节及条文的引用关系。同时，规范文档及章节、条文之间的包含关系也采用了:hasChild 关系表达。此外，考虑规范图谱包括多个规范，因此前述节点均增加了规范编号 code_id 以区分不同规范对应的章节及条文节点。

基于图 4 所示模型，规范图谱的构建主要包括三个步骤：1）创建节点：遍历前述步骤生成各规范 XML 文档创建 Doc、Section、SubSection、Item 等节点；2）创建节点包含关系：遍历各规范 XML 文档，根据各节点的包含关系生成:hasChild 关联；3）创建规范引用关系：遍历所有规范 XML 文档，提取各文档条文的引用关系（Ref）信息，创建条文对其他规范的引用关系。最后，我们基于 Python 模块库中的 py2neo 模块、xml 模块、re 模块和 os 模块实现了前述方法：首先，使用 os 模块历遍所有需要导入数据库的 XML 规范文件；其次，使用 xml 模块解析每个 XML 文件，获得结构化树状数据；然后，利用 py2neo 模块

进行数据操作，把数据导入 neo4j 数据库中，生成规范图谱。

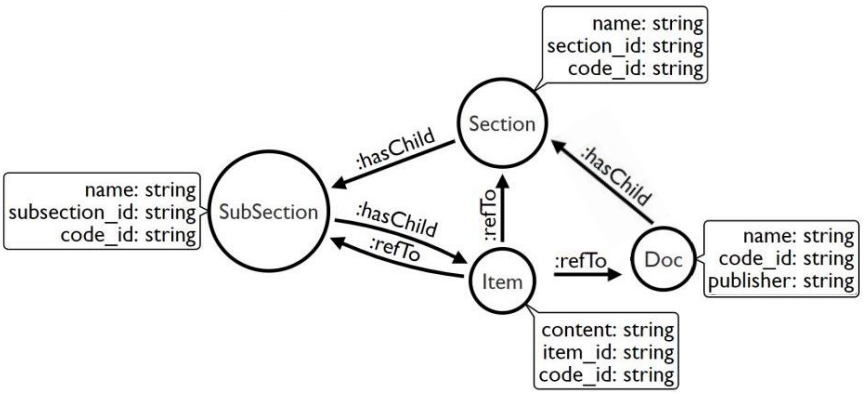

图 4　规范条文 neo4j 数据库存储模型

3　基于规范图谱的关联关系初探

基于前述方法和步骤，我们将收集到的 24 部消防相关规范的 PDF 文档进行了处理，生成了图 5(a)所示的规范图谱。整个规范共包括节点 3964 个，关联关系 4377 个，其中包含关系 4047 个，规范引用关系 330 个。基于该规范图谱，我们可采用如下语句查询规范内部条文的互相引用关系，查询所得部分结果如图 5(b)所示。

```
MATCH p=(d:Doc)-[:hasChild*..3]->()-[:refTo]-()<-[:hasChild*..3]-(d) RETURN p
```

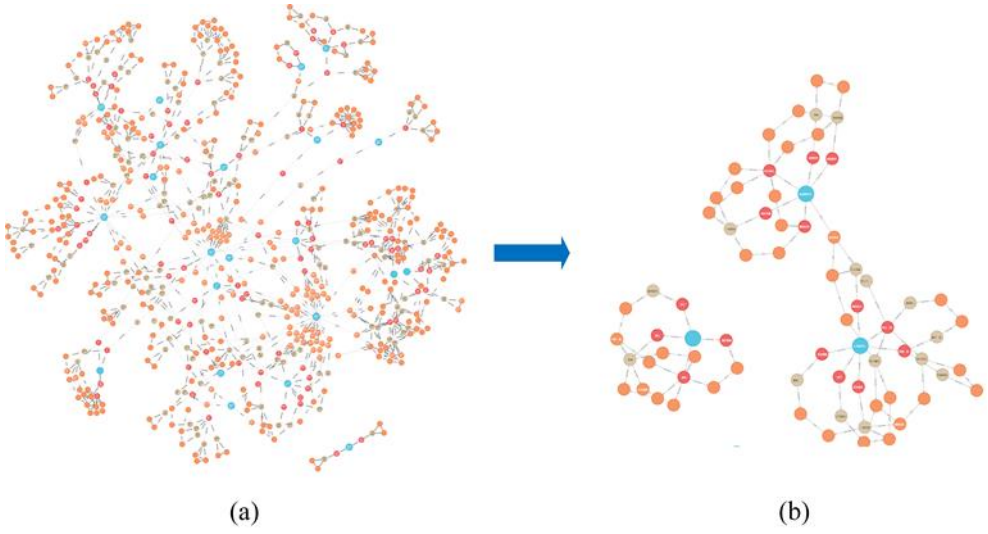

图 5　消防规范图谱及规范自引关系提取

同时，随着规范图谱中的规范数量不断增长，各规范之间的互引关系也将迅速增长。利用规范图谱查询规范之间的互引关系可为分析规范依赖关系、发现可能的规范引用冲突

提供支持。由于本研究采用的数据有限，为体现上述规范互引分析，本研究修改个别规范的互引情况，接着采用如下查询语句可得到图 6 所示规范条文的互引情况。

```
MATCH p=(d:Doc)-[:hasChild*..3]->()-[:refTo]-()<-[:hasChild*..3]-(d1:Doc) WHERE d<>d1 RETURN p
```

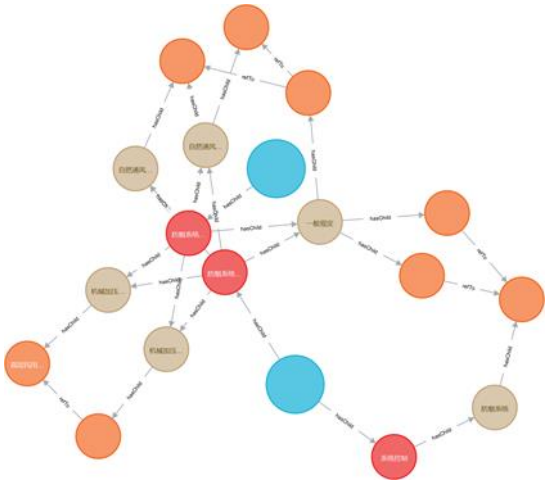

图 6　消防规范互引关系查询

4　结论及展望

本研究面向法规智能，针对传统规范文本结构化程度低、关联检索困难的问题，研究提出一种基于 XML 的规范条文结构化处理方法，并在此基础上构建了基于 neo4j 的规范图谱。最后将有关方法应用于消防规范图谱的构建与关联查询分析，验证了方法的可行性。有关研究可自动对建设领域规范进行结构化处理并自动构建规范图谱，为自动设计审查、智能规范检索、规范冲突分析等法规智能场景提供便于计算机处理的数据或规范图谱数据支持。

限于数据有限，有关算法的通用性仍需进一步加强。同时，研究构建的规范图谱规模尚小且粒度较粗，未来可进一步通过命名实体识别、实体关系抽取等方法建立细粒度的规范图谱，以支持更加广泛的应用场景。

参 考 文 献

[1]　傅智敏. 我国火灾统计数据分析[J]. 安全与环境学报, 2014,14(06):341-345.

[2]　Lindhols Nina. 走进法规智能[J]. 经济导报：医药技术, 2006(4):18-20.

[3]　梁泰鑫, 李斌, 赵光明. 环境保护法律法规智能检索系统研究综述[J]. 山东化工, 2011,40(09):40-43.

[4]　林佳瑞, 郭建锋. 基于BIM的合规性自动审查[J]. 清华大学学报(自然科学版), 2020:1-7.

[5]　陈建峡, 黄煜俊, 曹国金, 等. 基于知识图谱的司法案件可视化研究与实现[J]. 湖北工业大学学报, 2019,34(05):72-77.