DOI:10.19322/j.cnki.issn.1006-4710.2025.04.014 https://xuebao.xaut.edu.cn
引文格式:张艺博,何敏,杨明松,赵钦.基于本体的勘察报告合规性自动审查方法研究[J].西安理工大学学报,2025,41(4):594-602.
ZHANG Yibo,HE Min,YANG Mingsong,ZHAO Qin. The method for survey standard retrieval and inference combining knowledge base and ontology[J]. Journal of Xi'an University of Technology,2025,41(4):594-602.

基于本体的勘察报告合规性自动审查方法研究

张艺博[1],何　敏[1],杨明松[2],赵　钦[1]
(1.西安理工大学 土木建筑工程学院,陕西 西安 710048;
2.西安理工大学 计算机科学与工程学院,陕西 西安 710048)

摘要:针对当前勘察报告评审主要依靠专家知识经验进行人工审查,审查过程效率低、智能化水平差的问题,以湿陷性黄土地区岩土工程勘察报告的审查为例,提出基于本体的勘察报告合规性审查方法。结合规范的语义结构,设计了条文的知识表达模式,构建了基于语言技术平台(language technology platform, LTP)信息抽取的勘察规范领域本体。改进了基于词频-逆文本频率(term frequency-inverse document frequency, TF-IDF)算法,完成了勘察报告实体关系抽取与结构化存储,构建了勘察规范领域本体实例层。制定了 SPARQL(protocol and RDF query language)审查规则,实现勘察规范语义检索和推理,进而实现勘察报告的合规性审查。

关键词:勘察规范;合规性审查;信息抽取;知识建模
中图分类号:TU195　　**文献标志码**:A　　**文章编号**:1006-4710(2025)04-0594-09

The method for survey standard retrieval and inference combining knowledge base and ontology

ZHANG Yibo[1], HE Min[1], YANG Mingsong[2], ZHAO Qin[1]
(1. Faculty of Civil Engineering and Architecture, Xi'an University of Technology, Xi'an 710048, China;
2. Faculty of Computer Science and Engineering, Xi'an University of Technology, Xi'an 710048, China)

Abstract: The current survey report review relies primarily on experts' knowledge and experience for manual review. This process is inefficient and lacks intelligence. Taking the review of a geotechnical engineering survey report in a humid Loess Area as an example, we propose a method for survey report compliance review based on ontology. Combined with the semantic structure of the specification, the knowledge expression mode of the provisions is designed, and the domain ontology of the survey specification based on LTP information extraction is constructed. The TF-IDF algorithm is enhanced to accomplish entity relationship extraction and structured storage of the survey report. Additionally, SPARQL review rules have been developed. Based on rule-based reasoning, the semantic retrieval of the survey specification and the reasoning are carried out, followed by the compliance review of the survey report.

Key words: survey specification; compliance review; information extraction; knowledge modeling

收稿日期:2023-12-13;　**网络首发日期**:2024-06-11
网络首发地址:https://link.cnki.net/urlid/61.1294.N.20240611.1133.002
基金项目:国家重点研发计划资助项目(2022YFB2602200);陕西省自然科学基础研究计划资助项目(2023-JC-YB-219)
第一作者:张艺博,女,硕士,研究方向为智能建造与精益管理。E-mail:2220720078@stu.xaut.edu.cn
通信作者:何敏,男,博士,教授,研究方向为智能建造与精益管理、地下工程、生态水利工程优化。E-mail:Hem@xaut.edu.cn

勘察报告是结构设计的依据性文件，勘察报告审查能够确保结构设计基本数据的准确，确保结构设计的安全可靠、经济合理。勘察报告审查的过程依据大量相关规范标准，需要对勘察报告进行全面细致的评审和审核[1]。当前勘察报告评审主要依靠专家知识经验进行人工审查，审查过程效率低、智能化水平差[2]。随着人工智能技术的发展，机器学习、自然语言处理等新兴技术为实现规范的便捷化、智能化查阅及知识推理提供新的方向。通过人工智能技术实现勘察报告自动化审查，对保障工程的安全和可持续发展具有重要意义。

Eastman 及其团队[3]是合规性审查领域的先驱之一，他们将建设工程领域自动规则检查过程总结为 3 个阶段：规范解读、模型准备、模型审查。

工程建设规范知识表示及知识建模方面，学者们主要研究了通过自动化的方法实现规范条文的结构化转化。钟波涛等[4]在分析标准条文基础上构建本体，利用本体对条文进行标注，结合本体实现对规范知识的建模。胡云忠[5]通过构建本体对规范条文情景信息进行建模，使用本体来对条文所表达的信息进行表征，实现了规范条文的语义检索系统。Xu 等[6]采用本体论和基于规则的自然语言处理(natural language processing，NLP)框架来自动解释公用事业法规，提取法规信息并将其转化为逻辑条款。Moon 等[7]通过对规范条款的语义、语句分析，不断完善建筑法规信息自动化合规性检查技术。

工程建设领域模型准备方面，学者们主要以包含项目的几何、物理和功能信息的建筑信息模型(building information modeling，BIM)为基础，通过信息扩充的方法进行 BIM 的语义丰富及提取[8]，以提供全面的信息支持后续分析。陈培智等[9]从 BIM 模型中获取建筑基本信息，识别构件拓扑关系为自动化评估做准备。张吉松等[10]提出了 BIM 模型信息处理子模块，实现 BIM 模型信息的提取、转换与映射。

工程建设领域模型审查阶段的方面，学者们主要通过构建推理规则来检查模型是否符合之前阶段解读的规范要求，包括检查模型的建筑尺寸、材料使用、安全标准等。甘晨[11]通过人工完善的方式构建了 SPARQL 审查规则，开发施工图合规性审查的原型系统。陈远等[12]以 BIM 模型为检查对象建立了逻辑策略下推理机制，将规则库中的规则信息与 BIM 信息进行推理，反映每个构件的规范符合情况。刘平等[13]提出一种基于 BIM 的地铁车站设计自动审查系统，识别设计不符合规则要求的模型构件。

从以上分析中可以看出，目前知识建模主要集中于建筑设计、消防等领域，勘察领域尚无成熟的本体结构和知识网络。同时，既有文献对合规性审查的研究主要集中在 BIM 模型的审查方面，主要以 BIM 模型对应的元素及属性信息等对象作为实体构建本体[14]。勘察领域的对象显然与之不同，勘察报告中与审查要点有关的数据和信息等对象的分布比较均匀，且表述联系紧密，通过提取勘察报告关键词得到审查要点是一个潜在途径。

针对岩土勘察的应用场景和需求，拟以湿陷性黄土地区勘察报告审查为研究对象，研究勘察条文知识的结构化抽取方法与勘察领域知识库的概念层构建方法，分析勘察报告文本中审查要点提取、实体关系抽取与存储、勘察领域本体实例层构建方法，设计推理规则，探讨勘察规范语义检索和推理方法，辅助管理者实现基于本体的勘察报告审查。

1 勘察领域知识建模

1.1 勘察规范知识分析

岩土勘察领域的知识主要来自于相关规范和标准，存在规范种类多、描述范围广等问题。分类梳理湿陷性黄土地区岩土工程勘察领域所依据的规范和标准，识别、归纳、提取岩土勘察规范知识，分析规范语义知识结构，结合自然语言处理技术实现规范文本挖掘，为本体开发提供数据来源。

依据专家工程经验总结出了湿陷性黄土地区岩土工程勘察所依据的 14 本规范规程，具体信息见表 1。

1) 岩土勘察领域规范的知识特点

①不断更新变化。由于技术条件或政策法规的变化，规范会随之进行相应的更新变化。

②规范之间互相引用。规范涉及领域专业内容时，会推荐参考其具体规范，规范之间互相引用。例如：《岩土工程勘察规范》5.7.2 关于地震效应中对《建筑抗震设计规范》的引用，又如 6.1.1 条文中对《湿陷性黄土地区建筑规范》的引用，因此在进行勘察时需同时依据多本规范。

③约束内容分散。针对不同的勘察项目如勘探点布置、勘探点深度、地下水监测等，单一规范不能满足勘察要求，因此需要勘察人员和管理人员同时熟知多本规范，工程管理效率不高。

表1 规范汇总

Tab.1 Specification summary

编号	名称	编号	名称
GB 50021—2009	岩土工程勘察规范	GB/T 50123—1999	土工试验方法标准
GB 50025—2018	湿陷性黄土地区建筑标准	JGJ 120—2012	建筑基坑支护技术规程
GB 50007—2011	建筑地基基础设计规范	GB/T 50783—2012	复合地基技术规范
JGJ 79—2012	建筑地基处理技术规范	DBJ 61—6—2006	西安地裂缝场地勘察与工程设计规程
GB5 0011—2010	建筑抗震设计规范	JGJ 167—2009	湿陷性黄土地区建筑基坑工程安全技术规程
JGJ 94—2008	建筑桩基技术规范	GB 50585—2010	岩土工程勘察安全规范
JGJ/T 72—2017	高层建筑岩土工程勘察标准	建质[2010]215号	房屋建筑和市政基础设施工程勘察文件编制深度规定(2010年版)

2) 约束条文归类分析

通过对规范条文内容进行总结分析可以发现,规范条文一般的表述形式为使用场景描述结合具体的约束内容,特点在于丰富的场景和多样的约束形式和内容。本研究中根据实际研究对象将约束条文分为:属性值约束、条件约束、引用约束[15-16]。

属性值约束:针对不同的勘察对象存在不同的属性值约束,这些属性对勘察对象进行了全面的描述。例如:"详细勘察的单栋高层建筑勘探点布置,应满足地基均匀性要求,且不应小于4个"这一规范条文对单栋高层建筑这一勘察对象的勘探点布置数量进行了约束,要求数量上不少于4个;又如"条形基础的勘探孔深度不应小于基础底面宽度3倍且不应小于5 m"这一条文对勘探孔的深度进行了限制性描述。本研究将这类通过不同的属性值对勘察对象不同的属性进行约束的条文,定义为属性值约束。

条件约束:除了确定的数值约束外,还有部分条文是考虑由于不同的工况而引起的勘察手段的区别所设立的条款。例如:"当土层性质不均匀时,应增加取土试样或原位测试数量"这一条文约束了当土层不均匀的情况,应在先前规范约束的基础上,增加取土试样的数量;再如:"遇基岩或厚层碎石土等稳定地层时,勘探孔深度可适当调整",规定了稳定底层时的探测孔深度。此类规范条文可以看作是属性值约束类条文的深化约束,将此类规范定义为条件约束。

引用约束:本次的研究对象黄土地区高层建筑岩土工程勘察涉及到不同的土壤性质和建筑类型等问题,不仅需要参考综合性规范《岩土工程勘察规范》,还需参考《高层建筑岩土工程勘察规范》和《湿陷性黄土地区建筑标准》等专业规范中的详细条纹约束,因此将此类条文定义为引用约束。如规范条文:"对湿陷性粉砂,不扰动土样的试验方法和评定标准按照《湿陷性黄土地区建筑规范》执行"。

对规范条文进行归类分析后,还需对不同的约束类型所涵盖的知识表达形式进行总结,用以体现条文描述的知识架构,从而使条文所描述的实体约束更加清晰,方便领域本体的构建。

3) 约束条文的知识表达模式

分别总结属性值约束、条件约束、引用约束条文的表达形式的规律,形成统一的规范化知识表达模式,见表2。

表2 约束条文的知识表达汇总

Tab.2 Summary of forms of intellectual expressions of binding provisions

类型	知识表达模式	示例
属性值约束	(情景描述 SD)(勘察对象 Ob)(勘察属性 Vv)(情态词 Ph)(比较词 Pr)(属性值 Va)	(当存在相对软弱下卧层时/SD),(持力层/Ob)(厚度/Vv)(不宜/Ph)(小于/Pr)(5m/Va); (单栋高层建筑的勘察点/Ob)(布置/Vv)(不应/Ph)(小于/Pr)(4个/Va)。
条件约束	(勘察对象 Ob)(属性 Ve)(动词 Vt)(属性 Ve) (情态词 Ph)(勘察构件 Cg)(勘察手段 Cs)	当(高层建筑/Ob)(平面/Ve)(为/Vt)(矩形/Ve)时,(应/Ph)按(双排/Cg)(布设/Cs); 当(高层建筑/Ob)(平面/Ve)(为/Vt)(不规则形状/Ve)时,(应/Ph)在(凸出部位的角点和凹进的阴角/Cg)(布设勘探点/Cs)。
引用约束	(勘察属性 Vv)(情态词 Ph)(引用规范 St)	凡判别为(可液化的场地/Vv),(应/Ph)按现行国家标准(《建筑抗震设计规范》/St)的规定确定其液化指数和液化等级; 当场地位于(抗震危险地段/Vv)时,(应/Ph)根据现行国家标准(《建筑抗震设计规范》/St)的要求,提出专门研究的建议。

1.2 勘察规范领域本体构建

1.2.1 本体构建方法

黄土地区高层建筑岩土勘察无可重用的领域本体。基于七步法和 Protégé 软件，提出黄土地区高层建筑勘察规范领域本体构建方法如下。

1）抽取关键概念和术语。从黄土地区高层建筑领域知识勘察所涉及的规范文档中将重要术语概念梳理出来，发掘术语的隐含的专业经验、技术细节等知识，将其共同作为本体建模中的关键概念。

2）定义类以及类于类之间的关系。将所梳理出来的概念术语进行分类，并依据条文的领域知识和复杂逻辑对不同的类定义它们之间的从属关系。整个过程采取自上而下的方法，逐层建立子类。

3）添加属性和约束。根据所建立好实体层次，对实体添加对象属性或数据属性，丰富整个语义网。

4）本体的形式化表示。对所建立好的本体采用 RDF（resource description framework）语言进行描述，从而可以实现基于本体的检索和推理。

1.2.2 LTP 信息抽取

LTP 是哈尔滨工业大学提供的一套中文自然语言处理工具，提供了中文分词、词性标注、命名实体识别、依存句法分析、语义角色标注、语义依存分析 6 个模块[17]。为从非结构化规范文本识别实体和分类实体之间的特定关系，通过 LTP 实现勘察规范结构化处理。根据黄土地区高层建筑勘察规范术语生成的自定义词典定义黄土地区高层建筑勘察规范领域中的勘察对象、勘察手段等专有名词，协助勘察条文中实体的识别和抽取。结合实际研究特点，提出基于 LTP 的黄土地区高层建筑勘察规范结构化处理流程见图 1。

1.2.3 概念提取

对勘查规范进行知识抽取后，梳理出可以对条文进行结构化描述的关键术语概念，每个概念都代表一个实体类。根据不同的知识约束类型将不同实体分为不同的类并定义类与类之间的关系，确定父类和子类。本研究中共梳理出 256 个术语概念，共分为七大类：勘察规范、勘察条件、勘察对象、岩土性质、勘察手段、勘察属性、环境条件。列举部分类见表 3。

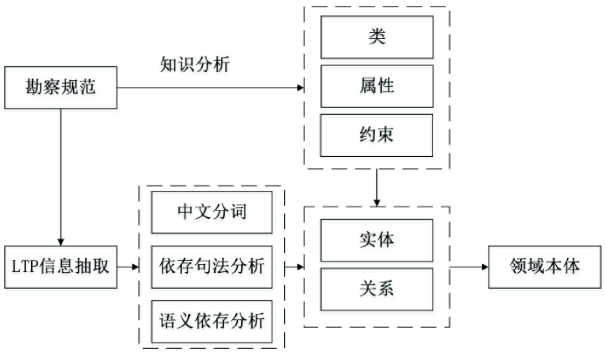

图 1 基于 LTP 的勘察规范结构化处理流程图

Fig. 1 Flowchart for structured processing of survey specifications based on LTP

表 3 实体关系表

Tab. 3 Entity relation table

父类	子类
岩土性质	湿陷性、湿陷性黄土、湿陷性土层、湿陷性黄土层、原土、地层结构、天然黄土、不扰动土样…
勘察手段	勘探点、勘探孔、控制性勘探点、静力触探、动力触探、探井、原位测试、土试样、原状土试样、不扰动土样、浸水试验、静荷载试验、物探、钻探…
勘察属性	勘探点间距、勘察深度、密实度、湿陷等级、控制性勘探点、液化判别深度、竖向间距、直径、勘探孔深度、地基主要受力层、基础底面宽度、变形验算、地基变形计算深度、建筑轮廓…

1.2.4 属性建立

在建立勘察规范的实体知识框架后，根据类与类之间的约束联系定义不同的属性类型如下。

1）has_contain：对象属性。表示一个类是下一个类的下游知识，即对其进行进一步约束的实体，主要表示例如“详细勘察的单栋高层建筑勘探点布置，应满足地基均匀性要求，且不应小于 4 个”此类约束条文中的详细勘察、单栋高层建筑以及勘探点这三个实体之间的关系，属性特征为单值属性。

2）has_unite：对称属性。表示两个实体具有相同层次的含义，主要表示“采取土试样和进行原位测试的勘探孔的数量，应根据地层结构、地基土的均匀性和工程特点确定”此类约束条文中的地层结构、地基土的均匀性以及工程特点三个实体之间的关系，

属性特征为对称属性。

3) has_compare:数据属性。表示对类的抽象数值约束的一类属性。例如表示“钻探取土试样孔的数量不应少于勘探孔总数的 1/3”中钻探取土试样孔此类实体的数值。数据类型为 string。

4) has_date:数据属性。表示类的具体数值大小。例如“地震液化判别应在地面以下 15 m 的范围内进行”中的地震液化判别等实体。数据类型选择:int、float。

1.2.5 可视化表示

在 Protégé 中将不同的数据属性和对象属性丰富到建立的本体概念中,对所建立好的黄土地区高层建筑勘察领域本体进行可视化的表示,见图 2。

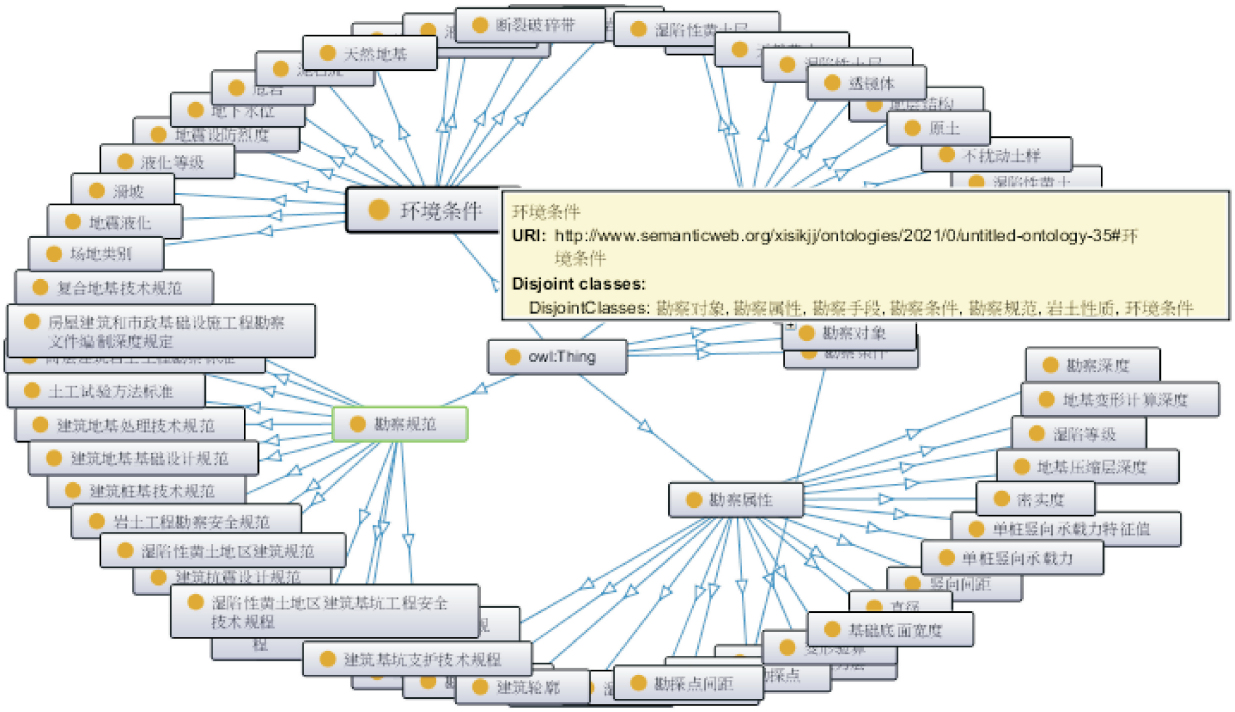

图 2 领域本体的可视化

Fig. 2 Visualization of domain ontology

2 勘察领域实例创建

2.1 勘察报告文本关键词提取

勘察报告是各个勘察项目数据的集合总结,反映各项目的基本情况,在 TF-IDF 基本算法的基础上加入一定人工干预机制,考虑词语在不同文本中的分布信息、近似表述以及位置信息,实现对勘察报告关键信息的概括性描述,以 30 本黄土地区高层建筑的岩土勘察报告为数据集。基于 TF-IDF 的实现文本关键词提取思路见图 3。

2.1.1 文本预处理

由于勘察报告专业性较强,在分词时以构建的领域本体实体为自定义词典进行辅助分词,同时根据勘查要点对一些关键词进行手动添加。

经过多次对数据集的预训练发现,需预先对一些候选词进行预设、过滤。对于具有同义描述的候选词,需预先定义它们属于同一关键词,并将他们的计算值合并。选择候选关键词时仅考虑名词、自定义词典、动词、专有名词和字符串,将其他词性的词汇进行剔除。由于动词表述了客体到客体的过程,是文本含义表述一个比较重要的点,但文本中也经常会出现一些实际含义不大的动词,例如:“进行”,“完成”等,设定对词长小于 2 的非标题动词进行剔除。

2.1.2 TF-IDF 改进措施

由于 TF-IDF 算法未考虑词汇所处的位置等因素,在基于 TF-IDF 算法提取到的关键词基础上对词汇的性质、所处位置、词长三个因素进行综合考虑并附以一定权重,对所得到的关键词进行进一步计算优化。

1) 位置权重

结合勘察报告的表述逻辑,标题词往往是本段文字所描述的主体含义,对标题词赋予更高的权重,在进行词汇遍历时跳过非标题词汇,非标题词汇乘以其权重,权重公式为:

$$\text{weight}_{\text{title}}(\omega_i) = 0.5 + \frac{100}{L_j} \tag{1}$$

式中：i 表示词汇出现在文档中的位置；L_j 表示文档 j 内包含词汇的总数；设其基准权重为 0.5。

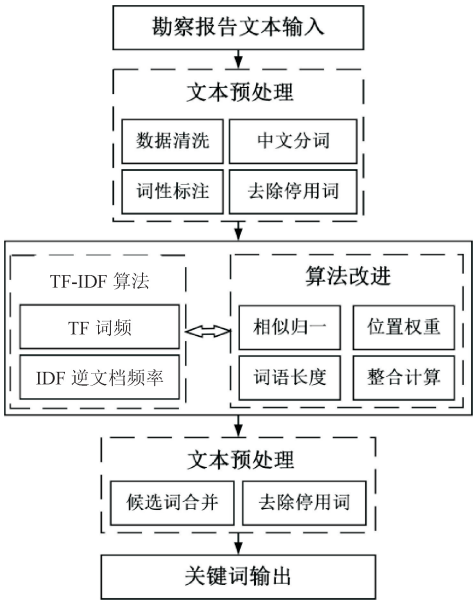

图 3 关键词 TF-IDF 计算流程

Fig. 3 Key words TF-IDF calculation process

2）长度权重

考虑到长度越长的词表述的含义越丰富，因此依据词汇的长度对其附以一定的权重，权重公式见式(2)。同时为了减少对长词的过分依赖，将文档中的最长词长度作为分母对分子进行归一化处理。

$$\text{weight}_{\text{len}}(\omega_i) = \frac{\text{len}(\omega_i)}{\max_{\text{len}}} \tag{2}$$

式中：len 表示词汇的字符数；$\max_{\text{len}}$ 表示文档所有词汇中最长词长度。

3）特征权重

在文本预处理阶段，已构建的领域本体实体词典在关键词选取时应具有更高的权重。字符串、名词、动词以及一些专有名词的重要性也应当进行排序并体现出来，不同词汇的权重见表 4。

表 4 特征词汇权重表

Tab. 4 Feature vocabulary weight table

词性	词性标注	权重
自定义词典	n_new	1.2
专有名词	name	1.0
字符串	string	0.8
名词	n	0.8
动词	v	0.6

在确定以上三个因素不同情况下设定的权重后，对候选词的 TF-IDF 值进行重新计算，还需考虑因素融合的权重。为了实现对勘察报告待审查内容的提取，将以上三个因素中的特征权重提高，使其能更大概率的提取有关勘察要点的关键词。经过对实验结果的多次模拟，以上三个因素的权重取值确定为：1.1、1.1、1.3。

通过对关键词的 TF-IDF 进行加权计算得出新的 IF-IDF 值，得到 7 项勘察要点及 105 条审查内容，包含勘探布点、勘探深度、取岩土样与测试试验、地下水及腐蚀性评价、地震效应及抗震设防、桩基础、基坑工程。

2.2 勘察报告实体关系提取

结合基于 TF-IDF 算法提取出的关键词以及勘察报告文本信息，根据审查要点，构建正则表达式提取模型，从勘察报告中对关键词的描述性知识进行提取和结构化输出，以关系数据的形式存储在 MySQL 中。

调用 Python 正则表达式模块 Re，通过 search 命令将计算 TF-IDF 所得的关键词以及文中的数字、比较词（不大于、不小于等）按照顺序进行输出，并以空格隔开。

通过对实体约束信息进行对比发现，一些实体只有一个约束信息，例如“勘察等级甲级”，此时只需输出其后一个非空格字符，对于有两个或约束信息的实体，例如“试桩数量不宜少于总桩数的 1%，且不应少于 3 根”，则需连续输出其后第二或第二加第三个字符。其使用的正则语句为 regex = '(? <= 勘察等级).[a-zA-z0-9_] *'，根据输出位数的不同稍作更改即可，将输出的字符以csv文件保存，并储存到关系数据库中。

2.3 语义元素映射

为了实现勘察报告审查，需要对关系数据库进行扩充，使其能表达的更丰富的知识。采用自顶向下的方式构建数据库，分别构建 genre、criterion、detail 三个表。其中，genre 中包括两列，分别为 genre_id 和 genre_name，主要用来储存勘察规范目录和顺序；criterion 包括三列，分别为 explore_id、explore_name、code_requirement，主要用来储存实体以及实体对应的条文；detail 包括三列，分别为 detail_id、detail_name、detail_content，主要用来将条文中的约束进行结构化储存，包含通过 LTP 信息抽取得到的结构化数据和部分人工扩充。

3 勘察报告合规性自动审查

3.1 勘察规范审查规则建立

基于勘察规范实例库,构建任务本体,将数据库中的表作为本体中的类,表的列作为属性,对象属性为表之间的关系,数据属性为不同列的格式约束。为了实现基于本体的勘察报告合规性审查,将关系数据库映射到本体上,采用 D2RQ(accessing relational databases as virtual RDF graphs)工具将关系数据转成 RDF。使用 SPARQL 对 RDF 格式的数据进行访问和查询,例如对"各勘察阶段工作要求"这一实体进行查询可输出根据实体之间存在关系的知识。

定义审查规则,描述实体之间存在的关系和属性。为满足审查要求,构建的规则集应涵盖所有审查要点,使用 SPARQL 规则模版将定义好的规则编码成计算机可理解的格式。例如,对于"高层基坑岩土取样试验和原味测试数量对每一主要岩土层不少于 6 组"定义的审查规则为:

?s rdf:type :detail_name.

?s :detail_name '室内压缩试验'.

?s :hasone?o.

?o :explore_id?n.

FILTER (?r >= 7)

为了构建基于本体的勘察规范检索应用,将 SPARQL 集成在代码中,在实现这个过程中,使用 Python 的 SPARQLWrapper 第三方库链接 endpoint 服务向控制台发送查询语句,实现结果的输出。

采用基于规则的推理方法,将勘察规范条文结构化的知识与勘察报告中提取的审查要点进行对比。为实现 RDF 推理,采用 Apache Jena 架构对导出的 RDF 数据进行管理和储存。将 RDF 类型数据转换成 TDB 类型数据、配置及启动 Apache Fuseki、利用 SPARQL 从 Apache Jena 中进行知识检索。具体步骤为:

1) 使用 tdbloader.bat 脚本文件储存 RDF 数据;

2) 手动配置推理引擎文件,添加 TDB 文件、本体文件、以及规则文件路径;

3) 配置规则推理文件,见图 4;

4) 通过浏览器访问"http://localhost:3030/"实现对 RDF 数据的查询和推理。

```
@prefix : <http://www.kancha.com#> .
@prefix owl: <http://www.w3.org/2002/07/owl#> .
@prefix rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#> .
@prefix xsd: <http://www.w3.org/2001/XMLSchema#> .
@prefix rdfs: <http://www.w3.org/2000/01/rdf-schema#> .

[rulecriterion: (?p :hastwo ?m), (?m :hasthree ?g), (?g :explore_name '勘察一般规定') -> (?p rdf:type :criterion)]
[ruledetail: (?p :hasone ?m) -> (?m :hastwo ?p)]
```

图 4 规则推理文件

Fig. 4 Reasoning document

3.2 自动审查平台设计

3.2.1 勘察报告合规性自动审查系统框架

勘察报告的自动审查系统主要涉及的技术流程包括构建勘察规范领域本体、基于本体的规范实体信息抽取、勘察规范关系数据库构建、语义检索和推理、报告信息提取、数据比对、输出结果等,具体框架见图 5。

将黄土地区高层建筑勘察规范知识库与勘察报告信息提取数据库,集成到系统性的平台中,实现一定程度上的勘察报告自动审查。本研究中系统的实现包括两大模块,分别为:使用 Protégé + Apache Jena + Fuseki + MySQL 来构建知识库的查询业务,使用 TF-IDF 结合正则表达式实现的报告信息提取模块,将这两大模块使用 PyQt5 + Anaconda3 开发环境进行融合,来实现文档的上传、处理、审查目标,为用户提供知识服务。

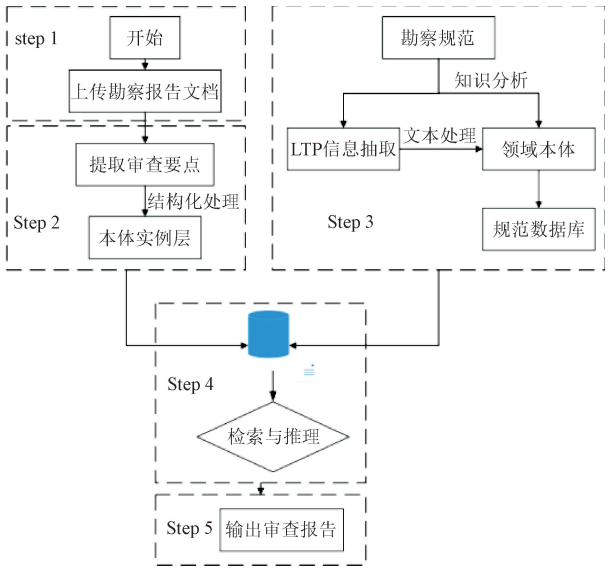

图 5 勘察报告自动审查框架

Fig. 5 Framework for automated review of survey reports

在系统实现时，首先点击上传文件按钮，将预处理的勘察报告文档上传，此时系统会首先进行文本处理；利用 TF-IDF 及正则表达式得出关系数据，将关系数据输入进检索系统，检索系统会将关系数据转化为查询语句，脚本语言会执行 SPARQL 查询从本体 Fuseki 服务器中获取 RDF 数据；将 RDF 数据同报告中得到的关系数据进行字符串比对，给出结论。

整个平台的内部实现流程主要包括如下四个模块。

1）文本预处理模块。在进行上传文档并确认后，平台会对文档进行预处理。首先将 word 文档转化为 txt 文本，去除图表、目录、公式、格式字符等信息。同时将勘察报告以序号、标题、内容格式以 csv 文档保存。

2）报告信息结构化模块。在将 word 文档处理成 csv 文档后，结构化模块会自动获取当前文件夹并读取，首先进行分词、停用词去除，随后利用加权 TF-IDF 算法计算 TF-IDF 值并将关键词进行输出，根据输出的关键词利用字符串搜索和正则表达式将关键词对应的描述信息进行结构化输出。

3）查询模块。将处理勘察报告所得到的结构化数据，利用正则表达式转化为 SPARQL 语句，执行 SPARQL 查询。通过查询语句，访问 Fuseki 服务实现对 RDF 数据的访问和结果输出。

4）结果输出模块。得到访问 RDF 数据的输出结果后，将这两部分结构化数据进行字符串和数据比对，得出审查信息、判定的结果以及给出依据的规范条文。

3.2.2　案例应用

以某大厦岩土工程勘察报告书为例进行应用，该项目位于西安市区，属于典型的湿陷性黄土地区。整个项目共包括 A、B、C 三栋建筑物，高度分别为 99.8 m、99.8 m、14.9 m，其中 A 座和 B 座工程重要性等级为一级，其余各建筑物工程重要性等级为三级，场地复杂程度等级为二级，地基复杂程度等级为二级，岩土工程勘察等级属甲级。报告书中所需审查的项目包括地基土力学性能室内试验、原位测试等。通过构建可视化的自动审查操作页面来实现对勘察报告的流程化处理。

在上传文档并执行确认命令后，系统会执行以上处理模块，将审查的结果以 csv 格式数据输出到预设文件路径。系统主要功能见图 6。

(a) SPARQL查询　(b) RDF推理　(c) 中审查结果输出

图 6　系统部分功能界面

Fig. 6　System partial function interface

从图 6(c)可以看到，审查结果设置成四列，分别为：id、input、result、detail。第一列为审查项目的排序；第二列为报告中需要审查的点，其中有属性值的实体使用比对其属性值是否一致给出结论，对于没有属性值的，将对应条文的内容进行输出，实现辅助审查的目的；第三列为判定结果；第四列为报告信息对应的条文，在这项输出中将实体对应的不同规范中有关的条文都进行列出，条文名以规范名称缩写加章节码表示，为使用者提供知识参考。

输出结果表明，该方法在审查已构建的 7 项审查要点所涵盖的强制性条文标准范围内做到较全面的覆盖，能够实现对需要判定的要点的审查结果输出，完成钻孔点及间距、贯入试验、不扰动土样 3 项属性值约束审查。对于没有属性值约束的审查内容，不给出审查的结果，而是将对应条文的内容进行输出。输出的条文为后续审查提供依据，实现了辅助审查的目的，印证了方法的有效性。

4　结　论

本文提出了利用本体、自然语言处理等技术构建岩土勘察领域知识模型与合规性推理方法。构建了湿陷性黄土地区岩土勘察领域本体，实现了湿陷性黄土地区岩土勘察领域知识组织与表达。改进了 TF-IDF 算法结合正则表达式处理勘察报告文本，实现了勘察报告审查要点内容的提取。构建了勘察报告审查平台，结合本体和推理规则实现了基于知

识库的检索和推理。

本研究在一定程度上实现了湿陷性黄土地区岩土勘察报告的合规性自动审查,但还存在一定的局限性。

1) 本文构建领域本体基于对规范的知识分析和专家经验指导,手动构建的,分析能力有限,智能化水平低。后续可结合新规范针对本体的自动化构建展开研究。

2) 本文采用预设权重的关键词提取算法获取勘察报告审查要点,样本数量小,对语义信息理解有局限性。后续可以研究基于事件抽取的勘察报告审查要点信息抽取方法。

3) 勘察报告涉及的审查要点众多,本文仅完成小部分文本审查内容的提取。后续结合新技术可对勘察报告的内容进行完善,对图表、目录、公式、格式以及完整性进行深入分析。

参考文献:

[1] 林佳瑞,周育丞,郑哲,等. 自动审图及智能审图研究与应用综述[J]. 工程力学,2023,40(7):25-38.
LIN Jiarui, ZHOU Yucheng, ZHENG Zhe, et al. Research and application of intelligent design review[J]. Engineering Mechanics, 2023, 40(7): 25-38.

[2] 陈健,盛谦,陈国良,等. 岩土工程数字孪生技术研究进展[J]. 华中科技大学学报(自然科学版),2022,50(8):79-88.
CHEN Jian, SHENG Qian, CHEN Guoliang, et al. Research progress in digital twin technology for geotechnical engineering[J]. Journal of Huazhong University of Science and Technology (Natural Science Edition), 2022, 50(8): 79-88.

[3] EASTMAN C, LEE J M, JEONG Y S, et al. Automatic rule-based checking of building designs[J]. Automation in Construction, 2009, 18(8): 1011-1033.

[4] 钟波涛,胡云忠,骆汉宾. 工程建设标准本体建模与质量监控应用研究[J]. 土木工程学报,2013,46(8):136-142.
ZHONG Botao, HU Yunzhong, LUO Hanbin. Study on building code ontological modeling and construction quality checking application[J]. China Civil Engineering Journal, 2013, 46(8): 136-142.

[5] 胡云忠. 基于本体的建筑施工质量规范知识建模与应用研究[D]. 武汉:华中科技大学,2013.
HU Yunzhong. Ontology-based knowledge modeling and application research on building construction quality specifications[D]. Wuhan: Huazhong University of Science and Technology, 2013.

[6] XU Xin, CAI Hubo. Ontology and rule-based natural language processing approach for interpreting textual regulations on underground utility infrastructure[J]. Advanced Engineering Informatics, 2021, 48: 101288.

[7] MOON S, LEE G, CHI S. Automated system for construction specification review using natural language processing[J]. Advanced Engineering Informatics, 2022, 51: 101495.

[8] 周逸苇,王广斌,曹冬平. 基于BIM的自动合规性审查研究综述[J]. 土木工程学报,2024(4):102-110.
ZHOU Yiwei, WANG Guangbin, CAO Dongping. A review on BIM-based automated compliance checking[J]. China Civil Engineering Journal: 2024(4): 102-110.

[9] 陈培智,史健勇,姜柳. 基于BIM和本体的建筑抗震性能评估方法研究[J]. 土木工程学报,2020,53(9):52-59,67.
CHEN Peizhi, SHI Jianyong, JIANG Liu. Research on seismic performance assessment method for buildings based on BIM and ontology[J]. China Civil Engineering Journal, 2020, 53(9): 52-59, 67.

[10] 张吉松,于泽涵,李海江. 基于语义网的BIM结构模型合规性审查方法[J]. 图学学报,2023,44(2):368-379.
ZHANG Jisong, YU Zehan, LI Haijiang. Compliance checking approach for BIM structural model under semantic web[J]. Journal of Graphics, 2023, 44(2): 368-379.

[11] 甘晨. 基于IFC和本体的建筑施工图合规性审查研究[D]. 武汉:华中科技大学,2019.
GAN Chen. Research on compliance review of building construction drawings based on IFC and ontology[D]. Wuhan: Huazhong University of Science and Technology, 2019.

[12] 陈远,张雨,康虹. 基于知识管理的BIM模型建筑设计合规性自动检查系统研究[J]. 图学学报,2020,41(3):490-499.
CHEN Yuan, ZHANG Yu, KANG Hong. Research on knowledge-based BIM for automated compliance checking system in architectural design[J]. Journal of Graphics, 2020, 41(0): 490-499.

[13] 刘平,尚永涛,许家铭,等. 基于BIM技术的地铁车站安全设计自动审查[J]. 铁道标准设计,2024(12):156-161.
LIU Ping, SHANG Yongtao, Xu Jiaming, et al. Automatic safety design review of metro station based on BIM technology[J]. Railway Standard Design: 2024(12): 156-161.

[14] YANG Mingsong, ZHAO Qin, ZHU Lei, et al. Semi-automatic representation of design code based on knowledge graph for automated compliance checking[J]. Computers in Industry, 2023, 150: 103945.

[15] 孙澄宇,柯勋. 建筑设计中BIM模型的自动规范检查方法研究[J]. 建筑科学,2016,32(4):140-145.
SUN Chengyu, KE Xun. Method of automatic design code checking for BIM models[J]. Building Science, 2016, 32(4): 140-145.

[16] 杨明松. 面向ACC的工程设计规范和设计模型的表达与存储研究[D]. 西安:西安理工大学,2019.
YANG Mingsong. Research on representation and storage of engineering design code and design information modelfor ACC[D]. Xi'an University of Technology, 2019.

[17] CHE Wanxiang, FENG Yunlong, QIN Libo, et al. N-LTP: An open-source neural language technology platform for Chinese[M]//Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing: System. Demonstrations. Online and Punta Cana: Association for Computational Linguistics, 2021: 42-49.

(责任编辑 王绪迪)