

中文引用格式:吴志江,刘梦瑶,马国丰. 基于自然语言处理的施工安全需求信息提取[J]. 中国安全科学学报,2025,35(6):51-59.

英文引用格式:WU Zhijiang, LIU Mengyao, MA Guofeng. Extracting construction safety requirement information using natural language processing[J]. China Safety Science Journal,2025,35(6):51-59.

基于自然语言处理的施工安全需求信息提取^{*}

吴志江¹讲师,刘梦瑶¹,马国丰²教授

(1 扬州大学 建筑科学与工程学院,江苏 扬州 225009;2 同济大学 经济与管理学院,上海 200092)

中图分类号:X928

文献标志码:A

DOI: 10.16265/j.cnki.issn1003-3033.2025.06.1237

基金项目:国家自然科学基金资助(72401250);中国博士后科学基金资助(2024M761993);扬州市自然科学基金资助(YZ2024165)。

【摘要】 为解决潜藏在项目文档中的施工安全需求信息因欠缺关联性、语义模糊而难以被发掘的问题,开发一种结合自然语言处理(NLP)技术的2阶段集成框架,用于项目文档分析并进行需求信息的分类提取。首先,结合NLP的多元技术获取待评估工程的安全目标,并基于主题模型建立关联模型以推荐合适的施工安全需求类型;然后,考虑3类要素的语义特征,分别采用关键词分析、情感分析以及依存关系分析对3类要素进行提取;最后,以2类建筑工程(民用和工业建筑)为例,开展施工安全需求的分类提取测试。结果表明:2阶段集成框架分别为民用建筑和工业建筑推荐到4种合适的需求类型,结合词性和词汇情感色彩能够有效提取到需求关键词和行为观点词,并且主体要素在补充建筑类型说明后的提取精度可达到88.6%;测试结果证实响应安全目标特征能够从繁杂的施工安全需求信息中推荐出适宜类型,而结合自然语言处理的需求信息分类提取可以避免主观偏好以提高信息提取精度。

【关键词】 自然语言处理(NLP); 施工安全需求; 需求信息; 项目文档; 需求类型

Extracting construction safety requirement information using natural language processing

WU Zhijiang¹, LIU Mengyao¹, MA Guofeng²

(1 College of Architectural Science and Engineering, Yangzhou University, Yangzhou Jiangsu 225009, China;

2 School of Economics and Management, Tongji University, Shanghai 200092, China)

Abstract: To solve the problem that construction safety requirement information hidden in project documents is hard to be discovered without relevance and semantic ambiguity, a two-stage integration framework combining NLP techniques was developed for project document analysis and classification and extraction of requirement information. First, the safety targets of the project to be evaluated were obtained by combining the multivariate techniques of NLP, and an association model was established based on the topic model to recommend the appropriate requirement types. Then, the semantic features of the three types of elements were considered, and keyword analysis, sentiment analysis, and dependency analysis were adopted to extract the three types of elements, respectively. Finally, two types of construction projects (civil and industrial) were used as case to test the type recommendation and extraction of

construction safety requirements. The results show that the two-stage integration framework recommends four appropriate requirement types for civil and industrial buildings respectively, and the combination of lexical properties and lexical sentiment can effectively extract the requirement keywords and behavior opinion words, and the extraction accuracy of the main elements can reach 88.6% after supplementing the description of building types. The test results confirm that responding to safety target features can recommend suitable types from the complicated requirement information, and the classification and extraction of requirement information combined with NLP avoids subjective preferences and improves the accuracy of information extraction.

Keywords: natural language processing (NLP); construction safety requirements; requirement information; project documentation; requirement types

0 引言

施工安全是土木工程领域广泛关注的重要问题,它直接关系到建设项目的正向评价和价值增值问题。当前,施工安全事故频繁发生,这些事故可能因多种因素的单一或组合影响而爆发^[1]。一些学者认为,不遵守既有要求所产生的执行偏差是事故发生的关键,该偏差被划定为事故因素衍生的先决条件^[2-3]。特别是,未能准确遵守有关施工安全的规则、标准以及要求,则容易导致安全管理的失败^[4-5]。因此,实施有效的安全管理有益于施工安全相关要求的准确执行,而准确获取这些信息则成为准确执行施工安全要求的关键前提。

针对施工安全需求的范围和特征,众多学者开展了大量的研究。如采矿项目中的规则、程序 and 规定已被定义为施工过程中应遵循的施工安全需求^[6]。还有一些学者将所有可能导致安全事故的人为和环境因素识别为施工安全需求信息的来源之一^[5,7]。考虑到安全相关要求源于安全利益相关者和建设项目本身,WU Zhijiang 等^[8]结合这 2 类主体的特点,将相应的偏好、意图、要求和规范定义为施工安全需求。此外,相关学者结合需求信息的结构特征明确了 3 类要素的内涵,即将信息发出者(指工程项目和安全相关者)确定为主体要素,将信息内容(指具体指令)确定为行为要素,将信息接受者(指作用对象)确定为对象要素^[4,8]。安全相关信息隐含在复杂的项目管理材料中,特别是直接记录管理过程的项目文档^[9]。这些文档是记录施工安全管理信息的重要载体,包括会议纪要、现场安全记录和施工计划^[10-11]等。考虑到项目文档中记录的信息庞大且庞杂,在提取安全信息之前需要进行文档分类和安全特征预处理^[12-13]。此外,需要消除无关信息的干扰,以获得准确的安全信息。

尽管过往研究关注到施工安全需求信息的重要性,但在实施有效安全管理之前,仍有 2 个关键问题需要解决。隐含在项目文档中的施工安全需求难以评估其真实价值,尤其是在安全信息与实际工程无关的情况下。另外,施工安全需求可能以口语化和碎片化的形式被记录在文档中,导致难以采用传统文本挖掘手段获取其准确信息。近年来,以自然语言处理(Natural Language Processing, NLP)为代表的文本处理技术被广泛采用,且该技术已被大量学者证明具有一定的先进性,特别是在复杂信息的挖掘方面。例如,HASSAN 等^[14]使用 NLP 中的语义分析技术挖掘合同文本的管理风险元素,并将其与卷积神经网络相结合进行特征聚类,获得了较为准确的合同文本信息。此外,NLP 中的多技术融合可用于解析项目文档中繁杂的信息^[15-16],为解决施工安全需求信息提取时带来的项目关联和模糊语义挑战提供思路。现有研究关注到施工安全需求因与项目文档的弱关联以及语义模糊的缺陷而导致信息价值未能被充分挖掘,而 NLP 的多技术融合则为解决该问题带来新途径。

鉴于此,文中提出 2 阶段集成框架用于项目文档分析,并利用 NLP 的多技术融合分类提取施工安全需求信息,以期得到 3 类要素契合语义特征的具体内容,从而组成为完整的施工安全需求。

1 需求信息提取的 2 阶段集成框架

2 阶段集成框架包括匹配安全目标的类型推荐和信息要素的精准提取。第 1 阶段,筛选出与安全相关的项目文档,再通过构建安全目标和需求类型相对应的关联模型对需求类型进行推荐。第 2 阶段,根据 3 类要素的特征,通过 NLP 中的多技术融合实现需求的精准提取,从而提取具体的需求信息。

为更加高效地从项目文档中获取需求信息,文

中借鉴文献[8]提出的分类框架划分需求类型,并在后续研究中采用该分类标准。

1.1 匹配安全目标的需求类型推荐

1.1.1 安全目标与施工安全需求的关系

基于工程项目的安全目标特征属性,将安全目标与需求之间映射关系的划分为2个过程,并明确安全目标—工程项目—安全需求的3层映射关系。

第1步,提取每个工程项目文档中的安全目标,建立工程项目与安全目标的映射关系,并引入隐含狄利克雷分布(Latent Dirichlet Allocation, LDA)主题模型进行关系建模。

第2步,依据项目文档中的安全目标所对应的需求类型,建立工程项目与需求类型之间的映射关系,即需求—目标(Requirement-Targets, RTs)标志。采用下式分别表示二者之间的关系:

$$\begin{cases} r_{pt} = \{ < r_{p,1}, r_{p,2}, \dots, r_{p,M} > \} \\ r_{pr} = \{ < r_{p,1}, r_{p,2}, \dots, r_{p,N} > \} \end{cases} \quad (1)$$

式中: r_{pt} 、 r_{pr} 分别为工程项目与安全目标、需求类型之间的关系; $r_{p,M}$ 、 $r_{p,N}$ 为工程项目与安全目标、需求类型的相关程度; M 和 N 为二者在安全目标短文中的数量。基于相关性的概念, $r_{p,M}$ 、 $r_{p,N} \in [0,1]$, 且其取值大小根据关联性程度进行度量。

根据映射关系的对应要求,建立安全目标和需求类型映射关系的关键在于抽取中间的工程项目层。由于部分工程项目所设立的安全目标较多,故其对应的施工安全需求应该是一个类型的组合。因此,通过项目文档中的项目规模条款搜寻尽可能相似的工程项目,来过滤掉相似度较小的工程项目,从而剔除不适宜的需求类型以提高推荐效率。

1.1.2 施工安全需求类型推荐

为推荐与安全目标相适宜的需求类型,按照话题聚类、建立关联模型以及推荐决策进行推荐。

1) 话题聚类。为从安全目标文本中提取安全特性,文中引入 LDA 模型从文本中抽取主题,该模型将每个安全目标文本都看成是词袋,且每个词语之间是独立个体。同时, LDA 模型假定“文本→话题→词语”的对应概率,使得每篇文本存在某种主题,并以特定概率分布的形式呈现。对于 LDA 模型而言,文档集 D 就是模型的输入,而每个工程项目安全目标描述的文本向量 $\mathbf{d}_{p-LDA} = (t_1, t_2, \dots, t_c)$ 为输出。其中, t_p 为对应工程项目 p 的安全目标主题,且 $p \in [1, c]$ 。将一定数量(m 个)的工程项目作为训练集并利用 LDA 模型获取安全目标主题。再

对安全目标主题进行向量化并构建特征矩阵。假定该训练集提取到 n 个安全目标主题,主题矩阵表示如下:

$$\mathbf{A} = \begin{bmatrix} t_1 T_1 & t_1 T_2 & \dots & t_1 T_n \\ t_2 T_1 & t_2 T_2 & \dots & t_2 T_n \\ \vdots & \vdots & & \vdots \\ t_m T_1 & t_m T_2 & \dots & t_m T_n \end{bmatrix} \quad (2)$$

式中: $\mathbf{A}$ 为同类别所有工程项目安全目标主题的特征矩阵; $t_p T_n$ 为工程项目 p 属于安全目标主题 T_n 的概率,且 $t_p T_n \in [0,1]$ 。另外, $\mathbf{A} = \langle t_p T_1, \dots, t_p T_n \rangle$ 指的是工程项目 p 与所有共享安全目标主题 t_p 的相关性,每1行的 $t_p T_n$ 概率之和为1。此外,特征向量 $\mathbf{A}_p$ 对应的概率值很小时;需求,可忽略其相关性。

2) 建立关联模型。基于工程项目和安全目标、需求类型之间的关系,为每个工程项目建立安全目标和需求类型的关联模型。将需求类型表示为特征向量,所建立的关联模型如下:

$$TR = \begin{bmatrix} r_{p,1} & r_{p,2} & \dots & r_{p,x} \\ r_{2,1} & r_{2,2} & \dots & r_{2,x} \\ \vdots & \vdots & & \vdots \\ r_{m,1} & r_{m,2} & \dots & r_{m,T} \end{bmatrix} \quad (3)$$

式中: $r_{p,x}$ 为工程项目 p 所需要的施工安全需求类型; x 为部位—活动—事故(Part-Activity-Incident, PAI)分类体系中所定义的施工安全需求类型总数。显然,当 $r_{p,x} = 1$ 时,意味着工程项目 i 需要这个需求类型 $r_{p,x}$,而当 $r_{p,x} = 0$ 时,则意味着该需求类型 $r_{p,x}$ 不属于工程项目 p 。

3) 推荐决策。为了保证待评估工程的需求类型有效性,可以从安全目标特征相似度的视角进行分析。引入余弦相似度来计算案例工程项目与待评估工程项目的之间的安全目标特征相似度,其计算规则如下:

$$\begin{aligned} \text{sim}(i, j) &= \text{con}(\theta) = \\ &= \frac{\sum_{n=1}^N t_i p_n \times t_j p_n}{\sqrt{\sum_{n=1}^N (t_i p_n)^2 \sum_{n=1}^N (t_j p_n)^2}} \end{aligned} \quad (4)$$

式中 i, j 分别为案例工程项目和待评估工程项目。通过安全目标特征相似度的计算,可获得安全目标相似的工程项目集合 m^* 。考虑到部分噪声的影响,必要时可基于阈值排除相似度较小的工程项目,从而获得高相似度的工程项目集合 m^* 。

随之,引入协同过滤思想推导对应安全目标的

需求类型。需求类型的推荐标准参考相似工程项目需求类型的概率值,计算公式如下:

$$P(i, R_{p_{f_i}}) = \frac{\sum_{k \in M^*} \text{sim}(i, k) P(k, R_{p_{f_k}})}{\sum_{k \in M^*} \text{sim}(i, k)} \quad (5)$$

式中: P 为某种需求类型出现在工程项目中的概率值; k 为相似案例工程项目; $R_{p_{f_i}}, R_{p_{f_k}}$ 为关联模型的计算值。根据计算规则可知: 当 $P(k, R_{p_{f_k}}) = 1$ 时, 表示工程项目 i 需要施工安全需求类型 R_{pr} 。反之, 当 $P(k, R_{p_{f_k}}) = 0$ 时, 则表示不需要。 $\text{sim}(i, k)$ 为工程项目 i 和 k 之间的相似度。计算出待评估工程与所有候选需求类型的概率值后, 依据大小对概率值进行排序, 从而为待评估工程推荐适宜需求类型。

1.2 施工安全需求信息要素的精准提取

1.2.1 提取对象要素

依据3要素的核心内涵, 对象要素的词性通常为名词, 并且该要素在同一项目的需求信息中较为明确, 故作为最显著指标。由此, 将需求信息中以名词为主的对象要素词汇定义为特征词汇, 并利用文档关键词提取方法进行要素提取。

词频-逆文本频率指数(Term Frequency-Inverse Document Frequency, TF-IDF)算法因其可量化且具有便捷性而得到了广泛应用。对于某个特定词汇, 其TF-IDF表示为:

$$\text{TF-IDF}(w) = \frac{w^d}{w_T^d} \times \log\left(\frac{S}{S_w} + 0.01\right) \quad (6)$$

式中: w^d 为词汇 w 在文档 d 中的词汇数; w_T^d 为文档 d 的总词汇数; S 和 S_w 分别为文档集的总文档数和含词汇 w 的文档数。由此, 可以计算得出各候选关键词的综合权重, 并依据权重进行优先级排序。同时, 结合语义分析判断剔除与需求无关的词汇仅保留相关联的关键词。然后, 再选取前若干个词汇作为项目文档最终提取的关键词。

1.2.2 提取行为要素

在一条完整的需求信息中, 行为要素不仅承担衔接主体要素和对象要素的作用, 还因其蕴含的语义情感而成为表达需求信息真实内涵的关键。譬如, “项目经理要求降低伤亡率”和“砌筑工人降低接缝标准”这2条信息, 其中“降低”在2种语义环境下的表现为截然不同的情感, 前者是正向情感(即被安全所需要的), 而后者则是负向情感(即被安全所摒弃的)。由于行为要素对应的词汇在不同

语境下具有差异性, 故可将其引申为行为观点词进行分析。考虑到项目文档在组词过程中存在激活效应, 故可利用该效应提取与需求关键词相关联的行为观点词。继而, 引入词激活力模型提取行为观点词。由模型中可知, 对于任意一对词 w_a 和词 w_b (词 w_a 为激活源, 词 w_b 为激活目标), 则词 w_a 到词 w_b 的激活力(waf_{ab})计算规则如下:

$$\text{waf}_{ab} = (f_{w_{ab}}/f_{w_a}) \cdot (f_{w_{ab}}/f_{w_b})/l_{ab}^2 \quad (7)$$

式中: f_{w_a} 和 f_{w_b} 分别对应词的词频; $f_{w_{ab}}$ 为2词的共线频率, 即在文本网络中词 w_a 在词 w_b 之前出现的次数; l_{ab} 为2词在所有共线情形下的平均距离, 且词相邻时的距离规定为1。

根据词激活力模型判断词之间的相关性, 即词在语义上的关联程度, 可判断词汇的可替换性。进而, 亲和度概念可用于表示词汇的关联程度, 并提出改进的亲和度计算规则^[17]。

$$A_{ab}^{\text{waf}} = \frac{1}{|K_{ab}| \sum_{k \in K_{ab}} \text{OR}(\text{waf}_{ka}, \text{waf}_{kb})} \cdot \frac{1}{|L_{ab}| \sum_{l \in L_{ab}} \text{OR}(\text{waf}_{al}, \text{waf}_{bl})} \quad (8)$$

式中: K_{ab} 和 L_{ab} 分别为能够激活 w_a 和 w_b 节点集合和能够被 w_a 和 w_b 激活的节点集合; OR为任意2个前后节点元素值的最小最大值情况, 其可使用二者的重叠率函数计算得到。在计算亲和度之后, 可获得与需求关键词相关联的原始行为观点词, 继而得到行为要素。

1.2.3 提取主体要素

在获取到行为观点词后, 可借助其与主体要素的依存关系来提取对应主体要素的词汇。为此, 引入NLP中的依存关系概念, 分析需求信息中词汇之间关系。假设 $D = \{d_1, d_2, \dots, d_m\}$ 是1组可能的依存关系类型, 依存关系 $d \in D$, 用于描述句子中2个词之间的依存关系。在众多常见的依存关系算法中, 图算法因其能够对多语句中词语关系的精准判断而被广泛使用^[18]。鉴于项目文档词汇语义的复杂性, 选择图算法进行依存关系分析是合适的。该算法中, 对于给定的句子 x 和分析模型的参数 θ , 在所有可能的依存图集合 G 中, 找到概率最大的依存图 G^* 。将候选句子的句法树拆解为多条依存弧, 把每条弧都用特征向量表示, 即 $f(i, j)$, 而每个特征向量也有对应的权值 W , 可根据特征向量和权值对每条弧进行打分, 分数累加后即为整棵句法树的分数, 表示为 $\text{score}(x, y)$ 。

$$\text{score}(x,y)=\sum_{i,j\in y}W\cdot f(i,j)\tag{9}$$

接下来,根据最大生成树即可得到最优结果。即搜寻得分最高的候选句法树,具体算法如下:

$$y^*=\operatorname{argmax}_{y\in G}\text{score}(x,y)$$

随之,可根据依存关系得分找出相应的标签以判断行为要素依存的主体要素。

2 施工安全需求信息提取案例研究

考虑到当前民用建筑和工业建筑占国内建筑工程总量的 50% 以上,是普遍存在且较为典型的建筑工程类型^[19],故选择建筑工程作为研究案例。而且,基于常规的工程项目规模、施工工艺、管理模式等方面选择典型案例,并且发生的安全事故也为常见事故。同时,依据项目特征采集施工类文档和安全规范类文档。其中,施工文档主要包括合同、施工组织设计、会议纪要等,而安全规范文档则包括项目对应的规范、标准等。最终,总共收集到 91 份项目文档,其中民用建筑 49 份,工业建筑 42 份。此外,此外,结合安全目标特征收集 209 份事故案例文档(包括民用建筑 137 份和工业建筑 72 份)用于安全目标特征的比较分析。为便于区分,结合安全目标特征分别编码为 SF1-SF137 和 SF1-SF72。

2.1 施工安全需求类型推荐结果

计算案例工程与待评估工程(民用建筑和工业建筑)的之间的安全目标特征相似度并结合对应需求类型的概率值确定潜在的需求类型。

1) 话题聚类结果。依据 OKINE 等^[20]提出的安全相关指标可知:安全目标分为伤亡、事故、规范及环境 4 个类别,因此,可确定文本的话题类别为 Topic 1—Topic 4。在 Python 3.9 中完成安全目标话题的聚类分析,通过文本分词、文本向量化以及 LDA 模型拟合等步骤测算出各个案例事故对应的安全目标话题概率。同时,计算各个特征词在安全目标话题中的权重,并依据特征词权重大于 0.65 的标准确定每个话题排名第 3 的特征词。最终结果记录在表 1 中,安全目标话题概率反映事故案例中安全目标短文属于某个话题的概率,可根据最大概率值确定安全话题特征。同时,特征词的分布情况进一步揭示了安全目标短文的内容特质,也为后续的需求信息提取提供线索。

2) 需求类型的概率值。依据概率的分布情况计算案例工程与待评估工程之间的安全目标特征相似度。基于安全目标主题特征矩阵,再应用特征相

表 1 4 个安全目标话题的特征词分布

Table 1 Distribution of characteristic words for four safety target topics

话题类别	特征词
Topic 1	[(‘工伤率’, 10.112 402 230 665 116), (‘重伤率’, 9.231 819 926 015 34), (‘负伤率’, 5.434 860 252 924 8)]
Topic 2	[(‘无重大’, 10.721 290 614 334 907), (‘火灾’, 8.812 313 409 433 295), (‘死亡事故’, 5.885 643 344 813 133)]
Topic 3	[(‘持证’, 9.956 477 710 760 14), (‘特种’, 8.051 253 999 312 595), (‘标准’, 7.185 467 314 077 14)]
Topic 4	[(‘扬尘’, 8.728 864 294 334 907), (‘文明’, 7.413 523 409 442 115), (‘噪声’, 5.885 643 333 863 372)]

似度计算规则计算得到特征相似度,计算结果如图 1 和图 2 所示。根据计算结果可知:案例工程与待评估工程的特征相似度存在显著差异,这也表明对应不同安全目标的案例工程在相似度方面存在偏差。为了排除相似度较小的案例工程对于需求类型推荐的负面影响,选取相似度阈值在 0.9 以上的案例工程构成高相似度的案例工程集合,并作为需求类型推荐的案例工程。其中,民用建筑分别选择 SF22、SF25、SF71、SF128、SF133,而工业建筑选择 SF1、SF11、SF37。

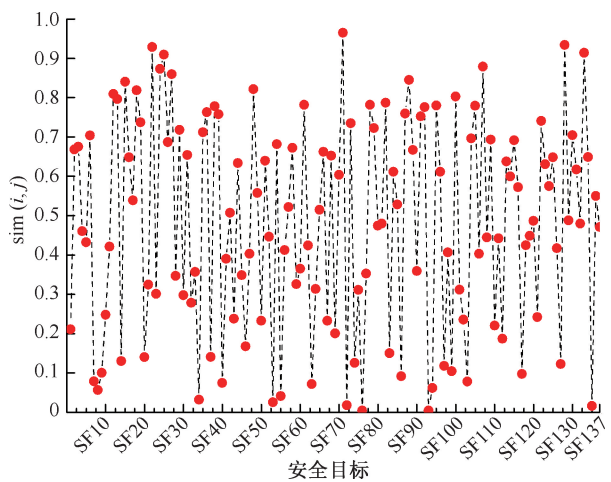


图 1 民用建筑的安全目标特征相似度

Fig. 1 Similarity of safety target characteristics in civil building projects

依据匹配模型的计算结果确定待评估工程与所有候选施工安全需求类型的概率值。随之,分别计算待评估工程与需求类型的概率值,并结合匹配模型中的计算值,判断概率值为 0 或 1 的情况。最终,

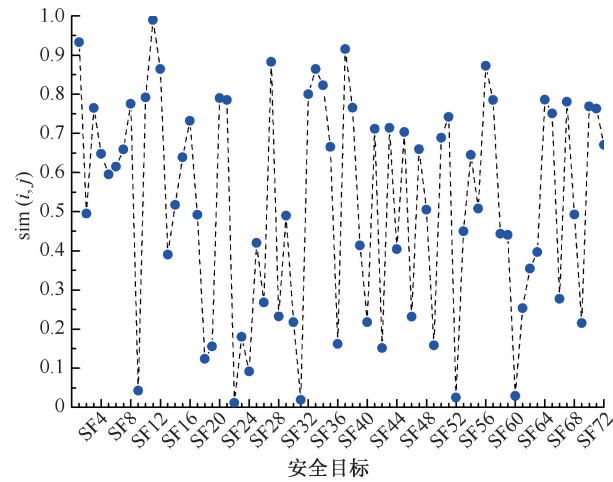


图2 工业建筑的安全目标特征相似度
Fig. 2 Similarity of safety target characteristics
in industrial building projects

各个需求类型的计算结果如图3和图4所示。对于民用建筑,概率值的结果显示需求类型P3A14I8(工程部位为“地面与楼面”,施工工序为“模板支护”,事故类型为“被掉落的物体击中”)的概率值为1,这表明需求类型完全契合该工程项目。相应地,安全目标短文所记录的信息也显示案例工程项目(SF25)曾发生过模板支撑坍塌的事故,从而进一步证实该需求类型与案例工程的适配性。而对于工业建筑,需求类型P2A7I8(工程部位为“主体”,施工工序为“吊装”,事故类型为“被掉落的物体击中”)的概率值为1,即需求类型与该工业建筑完全契合。为了充分挖掘需求信息,依据均值原则将概率值大于或等于0.5的需求类型纳入到对应建筑工程的分析对象。至此,分别确定4种民用建筑的需求类型(P1A1I2、P7A17I3、P2A12I6、P5A12I6)和4条工业建筑的需求类型(P2A8I6、P2A12I8、P2A12I6、P2A7I8),为划定施工安全需求信息提取范围提供依据。

2.2 施工安全需求信息要素提取结果

1) 对象要素的提取结果。结合需求类型的推荐结果,分别确定37份施工类文档、4份安全规范类文档以及28份事故案例类文档作为民用建筑的需求信息提取文档,此外工业建筑对应的项目文档分别为27份施工类文档、2份安全规范类文档以及14份事故案例类文档。而且,2类建筑工程的施工类文档和安全规范类文档均对应所推荐的需求类型进行内容筛选。

在构建完209份事故案例文档的训练集后,运用TF-IDF算法对民用建筑和工业建筑所选取的项

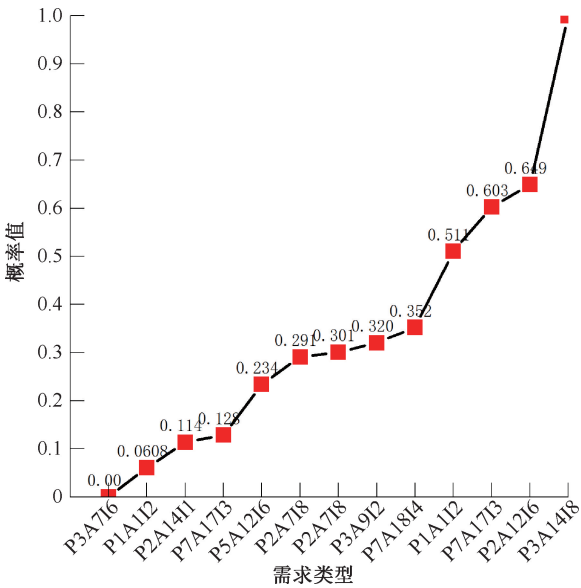


图3 民用建筑的需求类型概率值
Fig. 3 Probability values of requirement
types for civil buildings

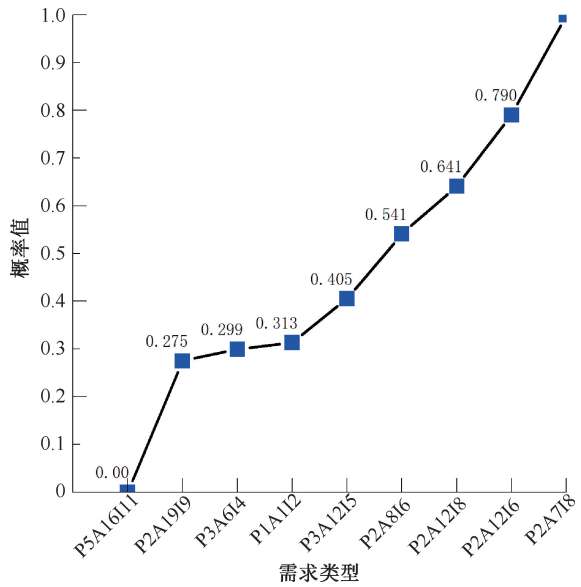


图4 工业建筑的需求类型概率值
Fig. 4 Probability values of requirement
types for industrial buildings

目文档进行测试,其中2篇文档的关键词分析结果见表2。对于民用建筑,所选取的项目文档为《钢模施工专项施工方案》,记录钢模施工从钢模选择、安装、拆除等全过程的技术要点。由表2可知:“钢模”、“精度”及“垂直度”均为新增关键词,且“钢模”的权重最大。这也表明TF-IDF算法在考虑词性、词位置和词跨度能够获取到更符合真实内涵的关键词,提高对象要素提取效果。工业建筑所选取的项目文档为《钢结构吊装安全管理会议纪要》,该

纪要详细记录钢结构吊装过程中的注意事项、意外情况及应急响应措施等内容。提取后新增“指挥”和“信号”2 个新关键词,舍弃篇幅较多但无明显实际价值的关键词(如“持证”、“合格”)。该算法能够更精准地捕捉到项目文档重点关注的内容,所提取的需求关键词更具代表性。

表 2 关键词提取结果

Table 2 Results of keyword retrieval						
案例	关键词					
民用	关键词	钢模	质量	检查	精度	垂直度
	权重	0.407 5	0.312 9	0.304 6	0.217 1	0.203 4
工业	关键词	吊装	起重机	钢丝绳	指挥	信号
	权重	0.430 5	0.302 4	0.301 1	0.231 9	0.120 5

2) 行为要素的提取结果。基于交互信息 (Mutual Information, MI) 值判断各需求关键词对应的观点词是否为种子观点词,并将其作为进一步提取需求行为观点词的依据。表 3 记录上述 2 个项目文档中部分需求关键词对应的种子观点词,并根据 MI 值筛选并剔除一些不相关的词汇。以需求关键词“钢模”为例,其在项目文档中对应的种子观点词为“安全”(MI = 0.634)和“可靠”(MI = 0.418),二者作为其他需求行为观点词提取的基准。

表 3 关键词对应的种子观点词提取结果

Table 3 Extraction results of seed opinion word for matching keywords			
案例	关键词	种子观点词	标准化后 MI 值
民用建筑	钢模	安全	0.634
		可靠	0.418
	精度	控制	0.754
		把控	0.661
工业建筑	吊装	稳定	0.821
		严禁	0.612
	指挥	明确	0.619
		清晰	0.533

为筛选候选观点词,采用亲和度计算公式计算“安全”、“控制”、“稳定”及“明确”4 个种子观点词与候选观点词之间的亲和度,结果见表 4。

对于民用建筑,《钢模施工专项施工方案》文档中“安全”“控制”2 个种子观点词的亲和度阈值分别设定为 0.237 和 0.107。对于工业建筑,《钢结构吊装安全管理会议纪要》文档对应“稳定”“明确”2 个种子观点词的亲和度阈值确定为 0.117 和 0.248。由此,通过提取算法可得到对应的观点词。

3) 主体要素的提取结果。为了适应中文文档的语句特征,利用 Python 3.9 并基于主流的深度学习

表 4 观点词的提取结果

Table 4 Extraction results of opinion word		
案例	种子观点词	候选观点词(亲和度)
民用建筑	安全	稳定(0.512)、足够(0.382)、光洁(0.343)、平整(0.297)、合格(0.281)、优质(0.274)、合理(0.237)
	控制	提高(0.447)、改进(0.402)、适宜(0.341)、达标(0.107)
工业建筑	稳定	严控(0.653)、防止(0.543)、注意(0.519)、平衡(0.422)、不得(0.353)、严谨(0.344)、必须(0.218)、切记(0.117)
	明确	有效(0.551)、特殊(0.421)、清楚(0.339)、准确(0.248)

习框架 Pytorch 1.13 搭建模型,引入科大讯飞公司开发的 IFlytek Voice Cloud 设计。通过依存关系提取到的主体要素(主语形式)分别用 True、False、Miss 表示。其中,True 表示正确提取的主语数量,False 表示错误提取的主语数量,而 Miss 则表示未提取的主语数量。针对民用建筑和工业建筑,分别提取到 313 和 238 个观点词,并基于依存关系提取各观点词相关联的主语,提取结果见表 5。对于施工安全需求信息中缺乏主体要素的情况,则人工补充主语后再进行依存关系分析。

表 5 主体要素的提取结果

Table 5 Extraction results of entity elements					
案例	状态	总共	True	False	Miss
民用建筑	正常	279	247	32	51
	异常	34	25	9	17
工业建筑	正常	201	188	13	25
	异常	37	30	7	14

根据提取结果可知:民用建筑和工业建筑平均错误提取比例分别为 18.9%和 12.7%,显示通过依存关系分析能够较好地提取到与行为观点词相关联的主语。此外,异常数据集中所提取到的主语错误率要明显高于正常数据集,如民用建筑中异常数据集中的错误提取数所占比为 26.5%,而正常数据集为 11.5%。这也显示缺乏主语的异常数据集对于依存关系分析的干扰较大。

为了客观评估依存关系对于主体要素提取的效果,采用精准率、召回率以及 F 值进行结果评价。其中,精准率侧重于提取正确主语数量与整个主语提取数的比值,而召回率则采用提取正确提取主语数与所有标注主语数的比值。同时,将 2 种建筑工程中遗漏的主语补充到主语集合中,并再次开展依

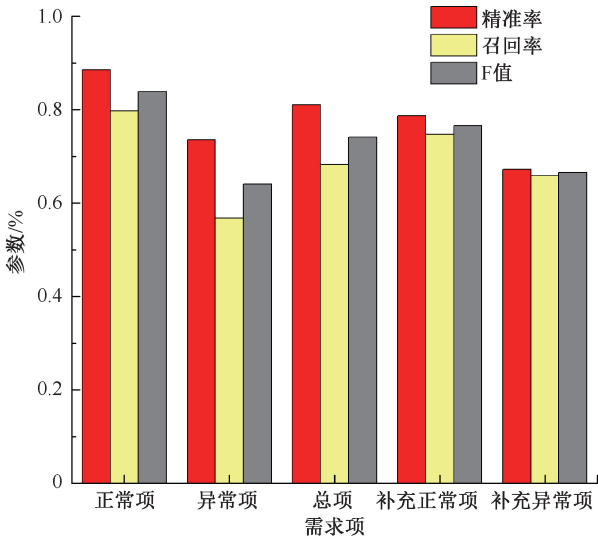


图 5 民用建筑的主体要素提取结果

Fig. 5 Results of entity element extraction (civil)

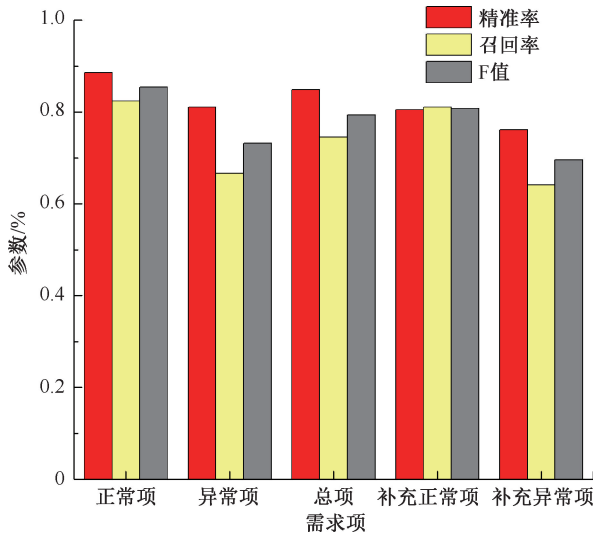


图 6 工业建筑的主体要素提取结果

Fig. 6 Results of entity element extraction (industrial)

存关系的测试,以深入检验关联效果。最终的测试结果描绘在图 5 和图 6 中。从中可知正常数据集补充建筑类型说明后(88.5%和 88.6%)的主语提取效果要明显好于异常数据集(73.5%和 81.1%),并且召回率和 F 值的表现相较于正常数据集也有所

下降。民用建筑因项目文档中所提取的观点词较多,且异常数据集中主体要素的语义存在较大差异,使得其基于依存关系的召回率相对较差(56.8%)。另外,在补充缺失主语后的提取测试中的提取效果都有不同程度的下降,这也反映出缺乏语义关联对于依存关系的建立产生阻碍。

至此,通过案例测试完成 3 要素的提取,并且证实基于语义特征的需求提取方法能够精准地提取出相应内容,并组成完整的需求信息。此外,测试结果也发现一个有趣的现象,即所提取出的施工安全需求涵盖了 2 类建筑施工中所发生的安全事故。例如,民用建筑所提取到的最可能对象要素为“钢模”(0.407 5),而工业建筑的最可能的对象要素为“吊装”(0.430 5)。而这 2 类建筑安全事故分布在钢模施工和吊装施工中,与所提取到的对象要素相一致。根据安全事故的致因机理可知,当施工安全需求未满足会出现安全风险因素,则有可能导致安全事故。

3 结 论

1) 基于施工安全目标所建立的关联模型能够从繁杂的原始需求信息中挖掘到与待评估工程直接关联的施工安全需求类型。而且,上述类型均响应对待评估工程的安全管理重点事项,甚至一些类型背后映射着安全风险因素。

2) 通过 NLP 的多技术融合可以突破模糊语义的限制以精准提取 3 类要素,并组合成完整的需求信息。同时,从项目文档中提取施工安全需求,可避免个人经验对于安全事项的偏好,且结合语义的文本处理技术能够提升提取效率。

3) 结合各要素特征的提取技术不仅可以根据词性提取到需求关键词,而且语义情感的嵌入也成为行为观点词提取的关键。根据依存关系进行要素提取是一种有效尝试,其错误提取比例甚至降低到 12.7%,而补充建筑类型说明后的提取精度可达到 88.6%,证实结合要素提取的可靠性。

参 考 文 献

- [1] LIN Jin, CAI Yuhang, LI Qinyuan. Development of safety training in construction: literature review, scientometric analysis, and meta-analysis [J]. Journal of Management in Engineering, 2023, 39(6): DOI: 10.1061/JMENA.MEENG-5445.
- [2] KULINAN A S, PARK M, AUNG P P W, et al. Advancing construction site workforce safety monitoring through BIM and computer vision integration [J]. Automation in Construction, 2024, 158: DOI: 10.1016/j.autcon.2023.105227.
- [3] AHMADISHEYKHSARMAST S, AMINBAKHSH S, SONMEZ R, et al. A transformative solution for construction safety:

- blockchain-based system for accident information management [J]. *Journal of Industrial Information Integration*, 2023, 35: DOI: 10.1016/j.jii.2023.100491.
- [4] WANG Xiyu, EL-GOHARY N. Deep learning-based relation extraction and knowledge graph-based representation of construction safety requirements [J]. *Automation in Construction*, 2023, 147: DOI:10.1016/i.autcon.2022.104696.
- [5] SANTOS A J R, SANTOS S P, AMADO C A F, et al. Labor inspectorates' efficiency and effectiveness assessment as a learning path to improve work-related accident prevention [J]. *Annals of Operations Research*, 2020, 288(2): 609–651.
- [6] AHAMED M F N, MARIAPPAN M. A study to determine human-related errors at the level of top management, safety supervisors & workers during the implementation of safety practices in the construction industry [J]. *Safety Science*, 2023, 162: DOI: 10.1016/j.ssci.2023.106081.
- [7] CHOI B, AHN S, LEE S. Construction workers' group norms and personal standards regarding safety behavior: social identity theory perspective [J]. *Journal of Management in Engineering*, 2017, 33(4): DOI:10.1061/(ASCE)ME.1943-5479.0000511.
- [8] WU Zhijiang, MA Guofeng. NLP-based approach for automated safety requirements information retrieval from project documents [J]. *Expert Systems with Applications*, 2024, 239: DOI: 10.1016/i.eswa.2023.122401.
- [9] 徐进, 朱孝文. 基于信息抽取的项目记忆获取方法研究 [J]. *情报理论与实践*, 2020, 43(8): 167–172.
- XU Jin, ZHU Xiaowen. Research on item memory acquisition method based on information extraction [J]. *Information Studies: Theory & Application*, 2020, 43(8): 167–72.
- [10] ZHANG Jiansong, EL-GOHARY N M. Semantic NLP-based information extraction from construction regulatory documents for automated compliance checking [J]. *Journal of Computing in Civil Engineering*, 2016, 30(2): DOI:10.1061/(ASCE)CP.1943-5487.0000346.
- [11] WANG Hanhsiang. Semi-automated identification of construction safety requirements using ontological and document modeling techniques [J]. *Canadian Journal of Civil Engineering*, 2015, 42(10): 756–765.
- [12] 成全, 张双宝. 基于深度学习的特征增强式安全事故文本实体识别模型研究 [J]. *中国安全生产科学技术*, 2024, 20(6): 58–66.
- CHENG Quan, ZHANG Shuangbao. Deep learning-based feature-enhanced textual entity recognition model for safety accidents [J]. *Journal of Safety Science and Technology*, 2024, 20(6): 58–66.
- [13] 刘均昊, 苏义坤, 张智博, 等. 面向建筑施工安全的标准知识支持研究 [J]. *中国安全生产科学技术*, 2024, 20(7): 156–162.
- LIU Junhao, SU Yikun, ZHANG Zhibo, et al. Research on standards knowledge support for building construction safety [J]. *Journal of Safety Science and Technology*, 2024, 20(7): 156–62.
- [14] HASSAN U F, LE Tuyen, LYU Xuan. Addressing legal and contractual matters in construction using natural language processing: a critical review [J]. *Journal of Construction Engineering and Management*, 2021, 147(9): DOI: 10.1061/(ASCE)CO.1943-7862.0002122.
- [15] ZHU Xiaofeng, LI Haijiang, SU Tengxiang. Autonomous complex knowledge mining and graph representation through natural language processing and transfer learning [J]. *Automation in Construction*, 2023, 155: DOI: 10.1016/j.autcon.2023.105074.
- [16] ZOU Yang, KIVINIEMI A, JONES S W. Retrieving similar cases for construction project risk management using natural language processing techniques [J]. *Automation in Construction*, 2017, 80: 66–76.
- [17] GUO Jun, CHEN Guang, XU Weiran. Word activation forces: distinctive statistics revealing word associations [J]. *Wireless Personal Communications*, 2012, 66(3): 511–521.
- [18] JIANG Yukun, GAO Xin, SU Wenxin, et al. Systematic knowledge management of construction safety standards based on knowledge graphs: a case study in China [J]. *International Journal of Environmental Research and Public Health*, 2021, 18(20): DOI: 10.3390/ijerph182010692.
- [19] LI Clyde Zhengdao, TAM V W Y, ZHOU Meizhuan, et al. Quantifying the coupling coordination effect between the prefabricated building industry and its external comprehensive environment in China [J]. *Journal of Cleaner Production*, 2024, 434: DOI: 10.1016/j.jclepro.2023.140238.
- [20] OKINE E A, ZAREI E, ROGGOW B J. Exploring the intellectual insights in aviation safety research: a systematic literature and bibliometric review [J]. *Safety Science*, 2024, 170: DOI: 10.1016/j.ssci.2023.106354.

作者简介: 吴志江 (1992—),男,江西鄱阳人,博士,讲师,硕士生导师,主要从事人机交互安全与协同方面的研究。E-mail: zjwu@yzu.edu.cn。

