城市轨道交通建设项目安全事故致因挖掘与重要度评估

许 娜,王文顺,王建平,李 解,黄若鹏

(中国矿业大学 力学与土木工程学院,江苏 徐州 221116)

摘 要:城市轨道交通建设项目安全事故时有发生,造成巨额经济损失、人员伤亡和负面社会影响。为将安全事故的经验和教训迁移到其它项目中,实现知识共享和重用,采用文本挖掘方法对221例城市轨道交通建设项目安全事故调查报告进行数据分析。首先对事故报告进行文本预处理,然后构建适用于城市轨道交通建设项目事故致因提取的专业词库,再基于词频选择特征值提取出事故致因。改进传统TF-IDF值,引入信息熵的概念评估事故致因的重要度,为城市轨道交通项目的安全风险预测和预控提供参考。

关键词:城市轨道交通;施工安全;事故致因;文本挖掘;信息熵

DOI:10.6049/kjjbydc.2018GC0099

中图分类号:U231 文献标识码:A 文章编号:1001-7348(2018)24-0134-05

0 引言

城市轨道交通是城市的生命线,具有建设周期长、参与方众多、与周围环境接口复杂、隐蔽工程多的特点,是典型的高风险复杂项目。由于城市轨道交通建设项目具有一次性、复杂性以及风险不确定性等特征,风险管理的经验和教训很难迁移到其它项目中实现知识共享和重用。为了吸取安全事故的经验和教训,很多学者聚焦于从安全事故中寻求致因,主要体现为3类:第一类是对事故数据进行统计学分析,如邓小鹏等[2]利用饼状图、柱状图等对126个地铁施工安全事故进行分析,揭示了事故发生的时间地点、事故类型等统计学规律;第二类是对事故致因进行统计学分析,如周洁静[5]运用鱼骨图从人、物、管理和环境4个方面分析了国内95起地铁施工事故的致因。许娜[6]统计了161起轨道交通项目安全事故,按照人、物、管理、环境4类事故致因对安全事故进行了分类;第三类是通过具体分析某一例事故,提取出事故致因。如Zhipeng Zhou等[4]详细分析了杭州地铁重大基坑坍塌事故的经过,提取出了该起事故的致因。以上第一、二类基于统计学的事故致因分析方法,从总体上揭露了城市轨道交通建设项目安全事故发生的规律,但主要是学者利用自身经验对事故致因进行判断和归纳,具有较强的主观性,而第三类个案研究虽可以深度剖析安全事故始末,但具有样本上的局限性。

文本挖掘是从大量文本数据中抽取有价值的信息和知识的计算机处理技术。文本挖掘技术具有从大量数据样本中提取特征值的优势,在建筑施工安全领域的应用主要为利用词频统计提取安全风险因素。如Esmaeili和Hallowell[7]分析了上千份建筑工程伤害事故记录,利用文本挖掘技术提取了22个安全风险因素;Tixier等[8]结合文本挖掘与自然语言分析技术,从建筑工程伤害事故记录中提取了安全风险因素。国外的分析对象为建筑工程伤害记录,且由于英文和中文表述的不同,在分析和词库构建上存在较大差异。国内学者李解等[9]收集了100份地铁施工安全事故报告,运用χ^2统计对特征项降维,得到了29个致险因素。由于特征项降维会损失部分致险因素,因此,本文在其研究的基础上,将数据源扩展到221份轨道交通建设项目施工安全事故报告,利用词频分析提炼出引发安全事故的致因因素,并引入信息熵评估事故致因的重要度,从大数据分析的视角为城市轨道交通建设项目的施工安全管理提供借鉴和参考。

1 文本挖掘方法和流程

1.1 文本挖掘流程

文本挖掘流程包括文本预处理、结构化数据、数据分析、结果可视化、知识发现等步骤,其分析过程如图1所示。

(1)事故报告收集与筛选:收集和筛选城市轨道交

收稿日期:2018-08-15
基金项目:住建部科技基金项目(2016-R3-036)
作者简介:许娜(1982—),女,江苏徐州人,中国矿业大学力学与土木工程学院博士研究生,研究方向为工程项目风险管理。

通建设项目安全事故调查报告，这些报告中详细记录了事故发生的经过、事故的原因、责任划定等内容，是深入分析城市轨道交通项目安全事故致因的优质语料。

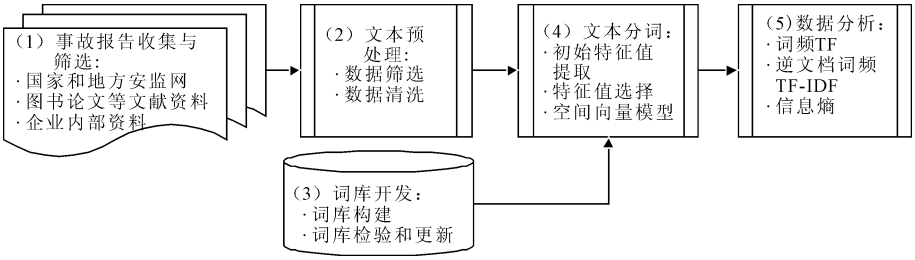

图 1　基于文本挖掘的事故致因识别过程

(2)文本预处理：对无意义数据、重复数据、缺陷数据等进行清洗，为挖掘结果提供高质量的、规范化的数据源。

(3)词库开发：由于城市轨道交通建设项目安全管理的标准化和规范化程度不够，事故致因的表述偏差较大，尚未形成具有明确语义的分词词典及语料库。因此，需要在词典匹配分词算法的基础上进行词库开发，包括词库构建、词库检验和更新两个步骤。

(4)文本分词：采用基于机械分词的术语抽取方法对安全事故报告进行分析，将事故报告中的非结构化数据转换为更加规范化的结构化数据。基于词频（TF）对初始特征值进行筛选，作为引发城市轨道交通建设项目安全事故的关键因素。

(5)数据分析：引用文档频率（DF）的概念计算事故致因引发安全事故的概率；改进传统 $TF\text{-}IDF$ 值，引入信息熵的概念评估事故致因的重要度。

1.2　事故致因挖掘

在初始特征值的基础上按照词频大小选择特征值，词频越大表示该词组对安全事故报告集的贡献越大。词频（TF）表示词组 T 在某个文档 D 中出现的频率，公式如下：

$$tf_{i,j}=\frac{n_{i,j}}{\sum_k n_{k,j}} \tag{1}$$

式中：$n_{i,j}$ 是词组在文件 D_i 中出现的频次，分母是所有文件中所有词组出现次数之和。本文关注的是在所有安全事故中蕴含事故致因的词组，公式(1)的分母对词组不具有区分性，因此，将 TF 值修正为：

$$TF_i=n_i \tag{2}$$

式中：分子 n_i 是词组在文件集中出现的总频次，TF_i 值越高，表示该词对文本集的贡献越大。将提取后的词组按照事故报告中的语义分析其代表的事故致因，用 S_i 表示，事故致因 S_i 的词频表示为 $TF(S_i)$。

1.3　事故致因的重要度评估

$TF\text{-}IDF$ 值用来评估一个词组对文本集中某一个文本文件的重要程度，词组在文本集中出现的次数越多，表示该词组的区分度越差，重要度越低，计算公式如下：

$$TF\text{-}IDF=tf_{i,j}\times idf_i=\frac{n_{i,j}}{\sum_k n_{k,j}}\times\log\frac{|D|}{DF_i} \tag{3}$$

式中：IDF 表示逆向文档频率。

$TF\text{-}IDF$ 值与词频呈正比，与文档频率呈反比。需要改进之处在于，对于词频，只考虑了词频数量绝对值多少，而没有考虑词频在文档中的分布，表征事故致因的词组在事故报告集中的分布越均匀，说明该事故致因越经常出现，应更加重视，所以，事故致因的重要度应与词频成正比，与词频在文档集中的分布成正比。

利用信息熵的概念表征词频在文档集中的分布，词频的概率分布越均匀，说明不确定性越大，信息熵也越大。假定事故致因 S_i 分布在 m 个事故报告中，频次表征为文档频率 DF_j^i 值，那么，S_i 在文档集中的概率分布 p_i 可以表示为：

$$p_i=\frac{DF_j^i}{\sum_{j=1}^{m}DF_j^i} \tag{4}$$

根据信息熵公式，事故致因 S_i 在事故报告中分布程度的信息熵 $H(S_i)$ 表示为：

$$H(S_i)=\sum p_i\log\frac{1}{p_i}=-\sum p_i\log p_i \tag{5}$$

综合词频和信息熵因素，将事故致因 S_i 的相对重要度 $I(S_i)$ 表示为：

$$I(S_i)=TF(S_i)\times H(S_i)=-n_i\times\sum p_i\log p_i \tag{6}$$

$I(S_i)$ 值越大的事故致因对安全事故的影响越重要。

2　数据收集和处理

2.1　事故报告收集和筛选

安全事故报告来源于国家及地方安全生产监督部门网站、在建地铁城市的质监站网站、已发表的论文和书籍、轨道交通公司单位内部资料和网络媒体。最终收集到 1999－2017 年国内发生的地铁施工安全事故报告 221 项，共涉及城市 27 个，占我国开通地铁城市（截止 2017 年底）的 80%。

2.2　文本预处理

由于安全事故调查报告篇幅较长，为了减少与风

险因素无关的词组对挖掘结果的影响，仅筛选调查报告中的“事故经过”“事故原因分析”这两部分内容作为文本挖掘的语料。同时，对报告中的错别字等进行修正。将所有文本资料统一集成到一个文本文件中，形成待挖掘的语料库。

2.3 词库开发

词库的构成包括过滤词词库、城市轨道交通建设项目安全风险专业词库、归并词表 3 类。

(1)过滤词词库：选取系统自带的《现代汉语虚词词典》。根据文本集的特点，在过滤词库中增加“地铁”“事故”“原因”等虽然出现频率高但无价值的词组。

(2)城市轨道交通建设项目事故致因专业词库：选用百度输入法、谷歌输入法中的《土木建筑》领域专业词典，逐条检查拟分词后形成的原始特征项，将一般词汇组合成具有特定含义的专业词组，整理形成城市轨道交通建设项目安全事故致因专业词库。

(3)归并词群表：文中包含大量的同义词，可靠的词库能有效分辨这些同义词，使文本对同一对象的描述尽可能趋同，降低数据节点的离散性，进一步聚焦文本挖掘结果，突出事故致因，例如“坍塌”“塌方”“塌”“倾塌”“塌陷”都可归一表达为“坍塌”，在此基础上结合人工识别不断改进词库。

(4)词库检验与更新：借鉴 Esmaeili 和 Hallowell[9] 对开发的词库进行检验和更新，流程如图 2 所示。

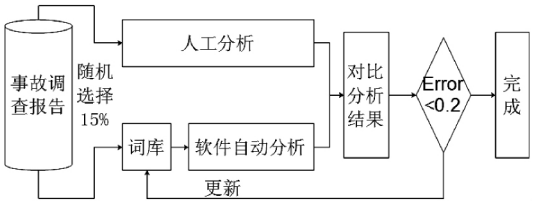

图 2 词库检验与更新流程

从 221 例事故报告中随机选取 188 份进行文本拟分词，剩余 33 份进行人工分析。仔细阅读每一份事故报告，记录与事故致因有关的关键词，列出关键词清单，最后将这份关键词清单与软件自动分析的结果进行对比，将人工分析出的新增关键词添加到词库中。重复上述过程，直到公式(7)所示的错误值小于 0.2。

$$Error = \frac{number\ \ of\ \ discrepancies}{total\ \ number\ \ of\ \ keywords\ \ from\ \ manual} \tag{7}$$

此项工作经历了 4 轮，意味着共更新了 4 次词库中的词语，最终把错误值控制在低于 0.2 的可接受范围内。每轮词库开发对应的错误值如表 1 所示。

2.4 文本分词

利用汉语词法分析系统(ICTCLAS 软件)进行文本分词，得到了 11 642 个初始特征值，根据公式(2)计算初始特征值的 TF 值，按照 80/20 法则，结合事故报告中对该词语语义的描述，提取 30 个特征值表征事故报告样本中的主要事故致因，如表 2 所示。

表 1 词库开发过程错误值

词库更新次数	错误值
1	0.58
2	0.32
3	0.21
4	0.16

表 2 城市轨道交通项目安全事故关键致因(部分)

序号	原词组	词频(TF)	事故致因	事故报告中的语义描述
S_1	支护体系	326	支护体系失稳	由于未超前支护，或已支护但存在不足，发生的支护(围护)体系失稳破坏，如掌子面爆破后封闭不及时、未及时进行支护等
S_2	管理	319	现场管理混乱	指总承包单位的现场安全监管不到位，包括现场安全管理工作不到位、管理薄弱、安全管理人员不足、施工作业无管理人员监管、查出安全隐患后未能及时纠正等
S_3	违章作业	282	违章施工作业	包括施工人员不按施工方案、规章制度、标准规范等要求擅自作业或简化工序流程。如 2016 年 2 月重庆某地铁线路在拆除支护结构贝雷梁过程中，盲目切割贝雷梁间接受力杆件，造成贝雷梁瞬间失稳倾倒

2.5 数据分析

根据公式(2)—(6)计算出事故致因 S_i 的词频 TF 值、文档频率 DF 值、引发安全事故的概率 RF 值、TF-IDF 值、信息熵 $H(Si)$ 值、重要度 $I(S_i)$ 值，如表 3 所示。

3 结果分析

3.1 事故致因因素分析

从表 3 可以看出，“S_1 支护体系失稳”、“S_5 复杂的地质条件”、“S_7 雨污水管道破损”、“S_{12} 不明地下水文条件”、“S_{15} 燃气管道破损”、“S_{18} 降雨”、“S_{26} 勘察或补勘不足”等均与城市轨道建设项目地下施工环境的特殊性有关。由于基坑作业和隧道区间的掘进等均与复杂的地质条件和施工环境密切相关，而且一旦发生安全事故多为坍塌事故类型，常造成大面积人员伤亡和巨额经济损失。

“S_3 违章施工作业”、“S_9 施工技术欠缺”、“S_{10} 设备设施故障或操作不当”、“S_{13} 吊车起重不当”、“S_{14} 安全培训不足”、“S_{19} 信息沟通滞后”、“S_{20} 安全意识不足”、“S_{28} 施工质量缺陷”、“S_{30} 材料堆放不当”体现了城市轨道交通建设项目施工工人在技术水平上的不足。我国的建筑工人多数来源于农民工，流动性较强，普遍安全意识不足，业务素质不高，也没有经过系统和专业的培训。他们一般只在开工前经历过短暂的培训，这些培训并不能显著提高其安全意识和技能水平。

“S_2 现场管理混乱”、“S_4 安全检查不足”、“S_8 施工

监测数据滞后”、“S_{16}施工方案不当”、“S_{17}补救措施不足”、“S_{21}应急预案不当”、“S_{22}施工组织协调不力”、“S_{23}未按设计要求施工”、“S_{24}安全防护不足”、“S_{25}安全交底不充分”、“S_{27}违章指挥”体现了城市轨道交通建设项目在安全管理的具体措施和落实方面还存在很多盲点。由于施工工序多、组织较为复杂，其较一般的建筑工程项目在管理控制方面还有诸多不足。

表 3　安全风险因素及其特征指标

序号	事故致因	TF	DF	TF-IDF	$H(S_i)$	$I(S_i)$
S_1	支护体系失稳	326	77	148.62	1.28	417.6
S_2	现场管理混乱	319	79	142.01	1.31	418.7
S_3	违章施工作业	282	81	122.78	1.44	406.8
S_4	安全检查不足	184	74	87.87	1.50	275.6
S_5	复杂的地质条件	160	77	73.05	1.49	238.4
S_6	安全制度及其落实不足	138	48	91.74	1.11	154.0
S_7	雨污水管道破损	129	61	72.28	1.45	186.8
S_8	施工监测数据滞后	120	55	72.07	1.43	170.8
S_9	施工技术欠缺	111	52	69.84	1.33	146.6
S_{10}	设备设施故障或操作不当	105	52	66.35	1.40	147.4
S_{11}	监理失职	105	29	91.86	0.96	100.4
S_{12}	不明地下水文条件	103	61	57.82	1.43	147.6
S_{13}	吊车起重不当	92	26	85.92	1.07	98.7
S_{14}	安全培训不足	92	46	62.73	1.29	119.0
S_{15}	燃气管道破损	88	20	91.74	0.92	81.2
S_{16}	施工方案不当	85	44	59.21	1.29	109.0
S_{17}	补救措施不足	85	37	65.92	1.21	102.3
S_{18}	降雨	79	42	56.82	1.30	103.2
S_{19}	信息沟通滞后	70	44	48.92	1.33	93.1
S_{20}	安全意识不足	68	39	51.58	1.27	86.7
S_{21}	应急预案不当	64	42	46.25	1.31	84.3
S_{22}	施工组织协调不力	64	39	48.79	1.19	76.7
S_{23}	未按设计要求施工	61	33	50.07	1.16	70.3
S_{24}	安全防护不足	55	28	49.90	1.11	61.5
S_{25}	安全交底不充分	52	26	48.11	1.07	55.2
S_{26}	勘察或补勘不足	50	31	42.20	1.15	57.0
S_{27}	违章指挥	42	22	42.36	0.98	41.6
S_{28}	施工质量缺陷	41	20	42.05	0.89	36.2
S_{29}	设计缺陷	26	11	33.54	0.72	18.5
S_{30}	材料堆放不当	22	15	25.99	0.84	18.6

“S_6安全制度及其落实不足”体现了当前城市轨道交通建设项目施工企业在安全管理的基层建设上还有不足，例如安全职责不清、安全管理体系不健全等，说明我国城市轨道交通建设项目的安全管理还比较粗放，这既是安全管理的短板，同时也是造成安全事故的高风险因素。

“S_{11}监理失职”、“S_{26}勘察或补勘不足”、“S_{29}设计缺陷”体现了监理方、设计方、勘察方的失误对施工安全的影响，虽然在事故报告中提及较少，但对事故预控起到至关重要的作用。工程勘察受到经费和时间等条件的限制，使设计师不能完全了解工程地质情况和水文情况，设计时存在设计深度不够、参数选择错误和未结合现场实际工况等情况。设计和施工的分离是导致安全事故发生的重要原因，通过设计减少安全隐患一直是施工安全管理研究的重要方向。Gambatese[10]提出安全设计的概念，指出应将设计单位纳入到安全管理体系当中，使其在设计时即考虑施工过程中可能出现的安全隐患。因此，在城市轨道交通建设项目安全管理中应全面考虑勘察、设计和施工相关风险因素。

3.2　事故致因的重要度分析

对表 3 中事故致因的重要度进行标准化处理，得到柱状图，其中，S_1－S_5的重要度排在前列，这与概率分析的结论基本一致。在这 5 个致因中，虽然引发安全事故的概率基本相当，但由于“S_1支护体系失稳”、“S_2现场管理混乱”、“S_3违章施工作业”在一个事故报告中被提及的次数较高，因此，认为其重要度明显高于S_4、S_5。

由于 $DF_9=DF_{10}=52$，$TF_{10}=TF11=105$，因此，选取S_9、S_{10}、S_{11}的数据进行对比分析。尽管S_{10}和S_{11}的词频相等，但S_{10}的文档频率更高，说明S_{10}引发安全事故的概率更高，因而其重要度更高。对于S_9和S_{10}，其文档频率相等且 $TF_9>TF_{10}$，从 TF-IDF 值判断S_9更重要，但信息熵 $H(S_{10})=1.40>H(S_9)=1.33$，说明$S_{10}$在事故报告中的分布较为均匀，即在多个事故报告中被多次提及，而S_9可能在一个事故报告中提及次数较多，其它事故报告中提及次数较少，因此综合词频和信息熵来看，S_{10}的重要度略大于S_9。以上数据较好地验证了相较于传统 TF-IDF 值，以 $TF(Si)\times H(Si)$度量事故致因的重要度更具优越性。

3.3　低频致因因素分析

有调查研究指出，来自建设方的工期压力[11]是造成施工程序缩减（例如混凝土养护时间不够就拆模）、施工方案不完善等的重要因素，但事故报告中很少提及建设方及政府主管部门因素。在收集的报告中仅 2 例提到了工期压力，因此，在特征项提取过程中进行了删减。其不在分析范畴内，但该事故致因仍值得关注。

4　结语

本文面向城市轨道交通建设项目安全事故调查报告，构建了从事故调查报告中提取事故致因的流程，提出了基于词频的事故致因筛选方法，基于词频－信息熵的重要度评估方法，得出主要结论如下：

（1）基于词频分析，从 221 例事故调查报告中挖掘出了 30 个事故致因，经过与相关标准规范、期刊文献比对，基本涵盖了城市轨道交通建设项目安全风险因素的全貌，说明基于词频的事故致因筛选方法具有较

好的适用性。

(2)引入信息熵描述事故致因在事故报告中的分布情况,改进传统的 *TF-IDF* 值,提出基于词频—信息熵的重要度评估方法,经数据验证,该方法能够综合词频大小、词频分布两个指标综合考量事故致因重要度。

本文研究数据来源是事故调查报告,因此,报告内容的真实性和全面性是能否准确挖掘事故致因的关键。此外,对于未引发安全事故的隐性风险因素,仍然需要专家调查等主观数据收集方法进行补充。

参考文献:

[1] 新华网.中华人民共和国国民经济和社会发展第十三个五年规划纲要[EB/OL].http://www.china.com.cn/,2016-03-18.

[2] 邓小鹏,李启明,周志鹏.地铁施工安全事故规律性的统计分析[J].统计与决策,2010(9):87-89.

[3] 李启明,王盼盼,邓小鹏,等.地铁盾构坍塌事故中施工人员安全能力分析[J].灾害学,2010,25(4):73-77.

[4] ZHIPENG ZHOU, JAVIER IRIZARRY. Integrated framework of modified accident energy release model and network theory to explore the full complexity of the Hangzhou subway construction collapse [J]. Journal of Construction Engineering and Management, 2016, 32(5): 05016013.

[5] 周洁静.地铁施工项目风险评价研究[D].大连:大连理工大学,2009.

[6] 许娜.轨道交通项目安全事故发生趋势和诱因分析[J].华侨大学学报:自然版,2016,37(5):558-563.

[7] ESMAEILI B, HALLOWELL M, RAJAGOPALAN B. Attribute-based safety risk assessment II: predicting safety outcomes using generalized linear models[J]. Journal of Construction Engineering and Management, 2017, 141(8): 15-22.

[8] TIXIER J, HALLOWELL M, RAJAGOPALAN B. Automated content analysis for construction safety: a natural language processing system to extract precursors and outcomes from unstructured injury reports [J]. Automation in Construction, 2016(62): 45-56.

[9] 李解,王建平,许娜,等.基于文本挖掘的地铁施工安全风险事故致险因素分析[J].隧道建设,2017,37(2):160-166.

[10] 徐吉辉,梁颖,亓尧.装备研制中的技术成熟度评价方法研究[J].科技管理研究,2016,36(2):66-70.

[11] ESMAEILI B, HALLOWELL M. Attribute-based safety risk assessment I: analysis at the fundamental level[J]. Journal of Construction Engineering and Management, 2015, 141(8): 15-21.

[12] GAMBATESE J, BEHM M, RAJENDRAN S. Design's role in construction accident causality and prevention: perspectives from an expert panel[J]. Safety Science, 2008, 46(4): 675-691.

(责任编辑:万贤贤)

Using Text Mining to Extract Safety Accident Causes and to Assess Importance on Urban Rail Transit Construction Project

Xu Na, Wang Wenshun, Wang Jianping, Li Jie, Huang Ruopeng

(School of Mechanics and Civil Engineering, China University of Mining & Technology, Xuzhou 221116, China)

Abstract: The safety accidents of urban rail transit construction projects occur from time to time, causing huge economic losses, casualties and negative social impacts. In order to transfer the experience and lessons of safety accidents to other projects for knowledge sharing and reuse, this paper uses text mining methods to analyze the data of 221 safety accident investigation reports for urban rail transit construction projects. Firstly, the text of the accident report is pre-processed, and then the professional thesaurus that is suitable for the accident cause extraction of the urban rail transit construction project is constructed. Then the eigenvalues are selected based on term frequency to extract the cause of the accident. To improve the traditional TF-IDF values, the concept of information entropy was introduced to assess the importance of accident causes and provide a reference for the safety risk prediction and pre-control for urban rail transit projects.

Key Words: Urban Rail Transit; Construction Safety; Accident Causes; Text Mining; Information Entropy