



您可能感兴趣的文章、专题：

“煤矿智能化综述及关键技术”专题

“我国煤矿瓦斯防治理论技术研究进展”专题

“煤地质与煤结构”专题

“煤矿安全+智能化”专题

“深部煤层瓦斯精准抽采”专题

“中国煤科首席科学家”专题

“矿井冲击地压灾害防治技术及工程实践”专题

2020 年度《煤炭科学技术》“领跑者 5000”论文

2021《煤炭科学技术》封面文章

煤炭行业青年科学家论坛(2021 年)

机电与智能化



移动扫码阅读

杨梦,陈宁,范誉航.煤矿事故案例存储与检索研究[J].煤炭科学技术,2021,49(9):103-109. doi: 10.13199/j.cnki.cst.2021.09.015

YANG Meng, CHEN Ning, FAN Yuhang. Study on storage and retrieval of coal mine accident cases [J]. Coal Science and Technology, 2021, 49(9): 103-109. doi: 10.13199/j.cnki.cst.2021.09.015

煤矿事故案例存储与检索研究

杨梦¹,陈宁²,范誉航³

(1.中国矿业大学(北京)机电与信息工程学院,北京 100083;2.北京大学遥感与地理信息系统研究所,北京 100871;

3.北京邮电大学国际学院,北京 100876)

摘要:为解决事故案例非结构化、多源异构、难以共享的问题,提高事故案例在应急救援管理中的利用率,利用网络爬虫技术获取由各地监管部门发布在互联网上的大量实时事故案例,通过框架法构建数据结构以表示事故案例蕴含的知识,建立了一个通用、全面、共享的事故案例数据库;在事故案例数据库的基础上,初步提出了一种新的案例检索算法,利用搜索引擎中倒排索引技术实现对案例非结构化数据进行检索,同时结合传统案例相似度计算方式对结构化数据进行匹配,实现利用少量关键信息进行非结构化案例数据的高效筛选,可使系统依据指挥人员意愿结合非结构化数据和结构化数据,进行有侧重、有倾向的案例检索,以中国煤矿安全生产网为例对瓦斯、水灾、火灾事故案例进行自动爬取,实践结果表明,此案例检索流程及算法提高了案例检索的有效性和实用性。

关键词:网络爬虫;框架表示法;信息检索;倒排索引;案例相似度计算

中图分类号:TD77

文献标志码:A

文章编号:0253-2336(2021)09-0103-07

Study on storage and retrieval of coal mine accident cases

YANG Meng¹, CHEN Ning², FAN Yuhang³

(1.School of Mechanical Electronic and Information Engineering, China University of Mining and Technology-Beijing, Beijing 100083, China;

2.Institute of Remote Sensing and Geographic Information System, Peking University, Beijing 100871, China; 3.International school,

Beijing University of Post and Telecommunications, Beijing 100876, China)

Abstract: In order to solve the problem of unstructured, heterogeneous, and difficult to share accident cases, and to improve the utilization rate of accident cases in emergency rescue management, this paper uses network crawler technology to obtain a large number of real-time accident cases published on the Internet by local regulatory authorities. The framework method constructs a data structure to express the knowledge contained in accident cases, and establishes a universal, comprehensive and shared dynamic database of accident cases; on the basis of accident case database, a new case retrieval algorithm is initially proposed, using the index technology as search engine to retrieve unstructured case data. At the same time, the traditional method of case similarity calculation is used to match structured data, and realize the efficient screening of unstructured case data with a small amount of key information, so that the system can be based on the wishes of the commander combining unstructured data and structured data, a focused and inclination case search was carried out. Taking China Coal Mine Safety Production Network as an example to automatically crawl gas, flood, and fire accident cases. The practical results show that this case retrieval process and algorithm improve the effectiveness and practicability of case retrieval.

Key words: web crawler; frame representation method; information retrieval; reverse index; similarity calculation

0 引言

煤炭是我国主要能源,煤炭开采过程易受多种灾害如瓦斯爆炸、水害、火灾等的威胁,随着开采强

度和深度的加大,灾害的威胁程度也相应增大,易发生群死群伤事故^[1]。煤矿事故的发生是一个复杂的过程,受各种因素综合影响,发生条件具有随机性、模糊性和不确定性,但大量的事故统计结果呈现

收稿日期:2021-02-25;责任编辑:赵瑞

基金项目:国家重点研发计划重点专项资助项目(2016YFC0801800)

作者简介:杨梦(1988—),女,山西晋中人,讲师,博士。E-mail: m.yang@cumt.edu.cn

出明显的规律性,事故的某些因素有一定的再现性。煤矿事故案例记录了事故处理和救援方式,包含了事故处理过程中的经验知识,通过分析煤矿事故案例,可以预防同类事故的再次发生,也可应急救援提供决策支持^[2]。

国内许多学者已经利用事故案例在应急处置和应急决策支持方面进行了研究和应用。胡冬红^[3]通过对大量的煤矿事故案例进行深层剖析,结合事故成因理论找出了煤矿事故的发生机理,建立了煤矿安全预警管理指标体系,并构建了安全预警管理系统。孙晋非^[4]研究在计算机的支持下如何对于突水预测预报做出决策,利用过去的案例或经验进行推理来求解新问题。祁运田^[5]在分析煤矿事故与隐患的关系后,结合井下灾害系统的特点提出开发煤矿安全决策支持系统,为煤矿安全决策和事故控制提供辅助决策信息。

目前在应急领域已经积累了大量的事故案例,但对以往事故案例的研究缺乏对案例系统建设的整体规划^[5],事故案例的使用率和利用率低,主要体现在:①传统的事例资料管理由本行业本单位进行调查和登记,经由出版社汇总发表在纸质版的出版物上,使得案例查阅困难,检索方式效率低,制约了对案例的检索利用;②现有的一些矿山安全信息系统提供了事故案例的存储和管理功能^[6-7],但主要侧重于信息的管理,案例多以非结构化的自由文本描述,案例蕴含的知识无法直接被计算机使用,难以深度梳理;③已有的一些应急辅助决策系统对案例中的知识进行了表示和利用^[8-9],但只侧重于局部信息的深入挖掘,解决了某一方面的问题,忽略了案例信息的整体利用和开发。为有效研究利用事故案例,需要开发一个系统化的信息支撑平台,即建立一个通用的、全面的、共享的事故案例数据库,并用合理的方式进行组织和表示知识,转换成可以被计算机存储、理解的形式,使系统不仅具有数据收集、存储和管理的功能,而且还能为事故隐患辨识、预警、危险性评价、灾害预测及应急预案决策等功能提供数据或知识支持。

随着互联网媒体的发展,已有很多大型发布煤矿信息的门户网站被企业、政府、学术机构及媒体等多方同时认可,网站收录了各类煤矿资料,为从事煤矿安全管理工作的人士及专业人员传播各类实用的煤矿信息。为利用互联网资源,笔者提出并设计了一个事故案例信息系统,利用网络爬虫技术搜集各安全监管部门发布在网上的多元的、大量的、实时的事故案例,然后采用框架表示法来表示事故案例蕴含的知识,解决事故案例非结构化、多源异构、难以

共享的问题。在对煤矿事故案例进行表示和存储的基础上,提出一种新的案例检索算法,使系统可依据指挥人员意愿结合非结构化数据和结构化数据,进行有侧重、有倾向的案例检索。

1 事故案例的获取、表示和存储

1.1 基于网络爬虫的案例获取

网络爬虫是通过模拟个体用户使用通用资源定位符(URL, Uniform Resource Locator)访问网页,并根据一定规则进行解析,得到具有实际意义数据的软件程序。Scrapy 框架是一套比较成熟的 Python 开源爬虫框架,是使用 Python 开发的快速、高层次的信息爬取框架,可以高效率地爬取 Web 页面并提取出用户关注的网页数据^[10]。

本文专注于煤矿事故案例的抓取,选用了聚焦网络爬虫^[11]利用 Scrapy 框架自动对互联网上官方公开事故案例数据进行爬取。首先利用 Python 模拟用户访问事故案例列表页面链接,获取事故案例列表,包括事故案例标题和 URL。针对每个事故案例 URL 利用爬虫获取 URL 对应网页内容,并按照规则提取事故案例的标题、发布时间、事故编号、来源网站、网页 URL 以及事故案例文本内容。同时采用代理方法,避免事故案例网站对爬虫 IP 地址封禁。最后将获取半结构化信息用 JSON 格式存储。

1.2 事故案例的表示和存储

爬虫获得的案例以半结构化文本的形式保存,包含一定的结构化属性信息,如事故发生的时间和地点等,同时也包含一定的非结构化信息,如事故的直接、间接原因,事故经过、后果等,需要进行合理的组织、管理和存储。框架表示法是以框架理论为基础的一种结构化案例表示方式,是一种存储以往经验和信息的通用数据结构^[12]。该理论从心理学观点出发,认为人们获取的新知识以一种类似框架的结构存储在人的大脑之中。当遇到新事物时,大脑就会找出一个类似的框架并依据新事物加以修改,从而形成新的认识^[13]。本节首先设计了基于框架表示法的事故案例数据结构,然后使用 MongoDB 数据库进行存储。

1.2.1 数据结构设计

框架表示法是将陈述性和过程性知识相结合的一种结构化知识表示方式^[14],既可用于存储事故的各属性值,也可表示事故原因、救援、处置过程等描述性知识。框架通常采用“侧面-槽-值”表示结构,也就是说,框架由描述事物的各个方面的若干槽组成,每个槽有若干侧面,每个侧面有若干值,或者还可以有

几个不同的槽值,通过框架嵌套结构,形成以框架为节点的树形结构。框架的另一个特性就是其继承性,即当子节点的某些槽值没有直接赋值时,可从其父节点继承。框架的属性结构和框架的继承性使框架知识的存储量比其他知识表示方法小,且对描述比较复杂的对象特别有效,知识表示结构清晰,直观明了。

在综合分析煤矿事故案例特点之后,将事故案例进行插槽式表示,单个案例事故利用一个框架进行表示,框架名称即该事故案例名称,瓦斯事故案例框架见表 1^[15]。

表 1 煤矿瓦斯事故框架

Table 1 Frame structure of gas accident cases in coal mine

槽 001		事故基本信息	
侧面	001001	事故时间	* 年 * 月 * 日
	001002	事故地点	* 省 * 市 * * * 煤矿
	001003	事故类型	类型描述
	001004	事故性质	责任事故类型
	001005	死亡人数	* 人
	001006	受伤人数	* 人
	001007	直接损失	* 元
	001008	间接损失	* 元
槽 002		企业信息	
侧面	002001	企业名称	* * * 公司
	002002	企业性质	私营/国营
槽 003		矿井信息	
侧面	003001	矿井级别	国有重点/地方重点/乡镇煤矿
	003002	证件情况	有/无
	003003	审批情况	通过/未通过
	003004	年产量	*
	003005	从业人数	* 人
	003006	瓦斯等级	低瓦斯/高瓦斯/煤与瓦斯突出矿
	003007	矿井开采方式	立井/斜井/平硐/综合 开拓
	003008	矿井通风方式	中央并列式/中央分列式/ 两翼对角式/分区对角式/ 区域式/混合式等
槽 004		事故区信息	
侧面	004001	事故位置	回采工作面等
	004002	具体位置	* * *
	004003	煤层厚度	*
	004004	平均倾角	*
	004005	煤层自燃倾向	容易自燃/自燃/ 无自燃倾向 煤层
	004006	煤尘爆炸倾向	有/无 爆炸倾向
	004007	走向长度	*
	004008	工作面长度	*
	004009	绝对瓦斯涌出量	*
	004010	透气性系数	*
	004011	坚固性系数	*
	004012	采煤工艺	综合机械化放顶煤/ 炮采放顶煤开采等
	004013	顶板控制	全部陷落法等

续表

槽 005		事故地质水文条件	
侧面	005001	事故水文 环境描述	(非结构化文本 描述 * * *)
	005002	地质构造复杂 程度描述	(非结构化文本 描述 * * *)
槽 006		事故过程	
侧面	006001	过程描述	(非结构化文本描述 * * *)
槽 007		事故原因	
侧面	007001	直接原因描述	(非结构化文本描述 * * *)
	007008	间接原因描述	(非结构化文本描述 * * *)
槽 008		救援措施	
侧面	008001	救援措施	(非结构化文本描述 * * *)
槽 009		防范措施	
侧面	009001	防范措施描述	(非结构化文本描述 * * *)

框架利用 9 个槽对事故案例进行详细描述,包括事故基本信息、企业信息、矿井信息、事故区信息、事故地质水文条件、事故过程、事故原因、救援措施以及防范措施。按照给定框架对原始事故案例数据相关信息进行提取,包括事故时间、事故地点等多个结构化数据以及事故过程、事故原因、救援过程等非结构化文本数据。经过提取后,将所提取信息依据煤矿事故框架对信息进行组织和表示。

1.2.2 案例存储

利用基于框架的知识表示法对瓦斯事故案例进行表示后,需要将表示好的案例数据进行有效存储。笔者选择 MongoDB 数据库进行事故案例数据的存储。MongoDB 数据库是 NoSQL 类型数据库中的 1 种,可以看作是基于文档类型的数据库,具有很强的灵活性和可扩充性,对复杂数据类型支持良好。在 MongoDB 中,1 个数据库(database)可以包含多个数据集合(collection),而 1 个数据集合(collection)则可以包含多个文档(document),1 个文档(document)又由多个域(field)组成。煤矿事故案例利用框架表示法进行表示后,其组织结构具有固定的格式,只需要在存储前将采集的数据转换成 MongoDB 中的文档,并加入相应的元数据信息,如案例 id、录入时间等。

2 案例检索及关键技术

案例检索是指依据输入信息从案例库中检索出相似案例的过程,案例检索算法是案例推理技术的核心,也是基于案例的决策支持系统的关键。案例检索通常转化为计算输入信息以及案例库中案例之

间的相似度问题。传统的案例检索基本过程包括^[15]:①确定目标案例的属性以及属性类型;②利用相似性度量算法计算目标案例的属性与案例库中案例属性间相似度;③输出相似案例。

煤矿事故案例的检索旨在快速查询与问题最匹配的历史案例,但随着事故的发展和衍变,决策人员的关注点也会发生变化,产生不同的检索倾向和检索需求;另一方面,事故案例是以半结构化方式存放的,有结构化的属性特征,也有非结构化的文本描述,传统的检索方式多是利用结构化的属性进行匹配,没有对非结构化信息的检索。笔者在传统案例检索算法基础上提出一种新的案例检索算法,案例检索算法流程可以分为检索准备阶段和实际检索过程 2 个阶段,如图 1 所示。

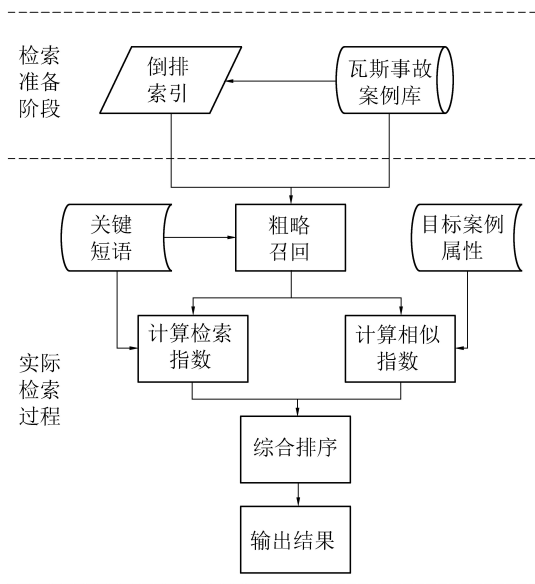


图 1 案例检索算法流程

Fig.1 Flow of case retrieval algorithm

2.1 检索准备阶段

在案例检索前,首先需要基于案例库建立用于支持关键词检索算法的倒排索引。倒排索引是一种特殊的索引技术,通常应用于信息检索领域。现代搜索引擎如谷歌、百度等均是基于倒排索引技术进行构建的。倒排索引核心思想是构造一种能够将词映射到文档的数据结构,从而方便通过少许关键词对复杂文档进行有效检索^[17]。通过以下步骤对事故案例建立倒排索引:

1) 收集煤矿领域专业术语词典,并基于该词典对事故案例文本数据进行分词,从而将文档映射为由多个词语构成的集合,其中分词采用开源软件库“结巴分词”。

2) 收集常见停止词,并对文档词语集合去除停止词。

3) 对各个文档对应的词语集合统计词频并记录。

4) 扫描所有文档的所有词,构建“词-文档”的映射结构,并进行持久化存储。

2.2 实际检索过程

实际检索过程是指系统依据当前案例的环境属性信息,以及决策者输入的检索关键短语信息触发一次检索行为的整个过程。在该过程中检索系统会进行以下步骤:

1) 粗略召回。粗略召回是指系统基于输入的搜索内容,利用倒排索引从案例库中召回全部相关案例的过程。搜索内容可以是关键词、短语或一段文字描述。检索时首先对搜索内容进行细致化分词,将其分解为 1 个或多个关键词,然后通过倒排索引检索出每个关键词所对应的全部文档,检索出的文档至少包含 1 个或多个关键词。

2) 计算检索指数。对粗略召回的案例依据上一步得到的搜索关键词计算对应的检索指数,检索指数越大,说明案例与输入信息越契合。对于粗略召回的结果,经过统计可以形成矩阵 A , 其中每 1 行表示 1 个事故案例,每 1 列表示对关键词进行分词后的 1 个关键词,矩阵中的每 1 个元素表示该案例包含该关键词的词频数目。

由于不同关键词的词频数量一般不同,为了平衡不同关键词带来的影响,需要对矩阵 A 进行式 (1) 的标准差标准化得到矩阵 B 。

$$B_{i,j} = \frac{A_{i,j} - \text{mean}(A_j)}{\text{std}(A_j)} \quad (1)$$

其中: $A_{i,j}$ 、 $B_{i,j}$ 分别为对应矩阵的第 i 行、第 j 列元素; A_j 为第 j 列数据; mean 为计算均值; std 为计算标准差,其目的是防止出现除法溢出现象,从而保证系统的稳定性。

为得到案例的检索指数,需要对不同关键词的数据进行线性加权组合,见式 (2)。

$$G_i = \sum_{j=0}^n W_j B_{i,j} \quad (2)$$

其中: G_i 为案例 i 的初步检索指数; W_j 为第 j 个关键词对应的权重值。第 j 个关键词 s_j 的权重 W_j 计算公式为

$$W_j = f(s_j) = \begin{cases} 0, & s_j \text{ 为停止词} \\ 2^{\text{len}(s_j)-1}, & s_j \text{ 为非停止词} \end{cases} \quad (3)$$

其中, len 为计算字符串长度函数。对于非停止词类关键词,其权重会随着关键词长度变大呈指数增长,所匹配的关键词长度越长,该关键词所对应的权重越大。在得到各个案例的初步检索指数 G 后,

在案例维度上对 G 进行线性变换,使之线性映射到 $[0,10]$ 区间,并得出最终的检索指数 \hat{G} 。

3) 计算相似指数。计算粗略召回的案例与当前发生事故的相似指数,相似指数越大,表示检索案例与当前事故在给定属性范围内相似性越高。基于煤矿事故特点以及当前获取到的事故案例所包含的有限信息,提取 14 个属性特征作为案件相似度计算条件,14 个属性特征具体见表 2。

表 2 事故发生环境的属性特征

Table 2 Attribute characteristics of accident environment

编号	名称	变量类型/预设权值
004001	事故位置	无序类别型/4
003006	矿井瓦斯等级	有序类别型/4
004003	煤层厚度	数值型/4
004004	平均倾角	数值型/4
004005	煤层自燃倾向	有序类别型/4
003004	年产量	数值型/3
003008	矿井通风方式	无序类别型/3
004011	坚固性系数	数值型/3
004010	透气性系数	数值型/3
004007	走向长度	数值型/2
004008	工作面长度	数值型/2
004006	煤尘爆炸倾向	布尔型/1
003001	矿井级别	有序类别型/1
003007	矿井开采方式	无序类别型/1

针对不同类型的属性特征,设计对应的相似度计算方式,并通过线性映射保证每类属性的相似度保持在 $[0,10]$,然后采用线性加权方式将不同属性相似度进行归总,从而得到相似指数。

无序类别属性及布尔类型属性,如事故位置、矿井通风方式、矿井开采方式、煤尘爆炸倾向,采用式(4)进行计算,其中 a, b 表示不同案例的属性, $d(a, b)$ 表示 a 和 b 属性间的间距:

$$d(a, b) = \begin{cases} 10 & a = b \\ 0 & a \neq b \end{cases} \quad (4)$$

有序类别属性,如矿井瓦斯等级、煤层自燃倾向、煤矿级别,采用式(5)计算间距。

$$t(a, b) = |\text{rank}(a) - \text{rank}(b)| \quad (5)$$

数值型属性分为 2 种方式计算其对应相似度,一种是可转化为有序类别型的数值型属性,如煤层厚度、平均倾角。对于其他数值型属性,通过计算其与案例间位置距离来计算其相对其他候选案例与目标案例间的相似度。具体计算时,首先利用式(6)计算距离 t ,其中 l 为由召回案例和目标案例属性值构成的有序序列, $\text{iloc}(a, l)$ 为计算 a 在有序序列 l

中的位置。

$$t(a, b) = |\text{iloc}(a, l) - \text{iloc}(b, l)| \quad (6)$$

在得到每个属性的相似度后,通过式(7)计算所有属性相似度的线性加权和,从而得到各个案例的初步相似指数 D_i ,其中 Q_j 为预先设定的第 j 个属性对应的权值,该权值在设定时需要考虑不同属性间相对重要程度,在咨询专家意见后给出了属性特征的预设权值,见表 2,在使用过程中权值可以根据实际效果进行调整。 $d_{i,j}$ 为第 i 个案例中的第 j 个属性值。然后在案例维度上对 D_i 进行线性变换,使之映射到 $[0,10]$ 区间,得出最终的相似指数 \hat{D} 。

$$D_i = \sum_{j=0}^n Q_j d_{i,j} \quad (7)$$

4) 综合排序。步骤 2) 获得的检索指数体现了决策者主动检索需求,步骤 3) 获得的相似指数表达了案例与当前事故的发生环境相似程度,通过对检索指数 \hat{G} 和最终相似指数 \hat{D} 计算线性加权和求得综合指数 I ,其表达式见式(8),其中 w_g 和 w_d 分别为检索指数与相似指数的权值,本文当前均设定为 0.5。综合指数较大的案例相比于候选案例,更加满足与目标案例的相似性以及决策者主动检索的倾向性。

$$I = w_g \hat{G} + w_d \hat{D} \quad (8)$$

5) 输出结果:召回的案例通过利用综合指数从大到小进行排序,并将结果进行输出。

3 事故案例检索实践

基于 Scrapy 框架,利用 Python 语言实现网络爬虫,以中国煤矿安全生产网为例对瓦斯、水灾、火灾事故案例进行自动爬取,然后将获取的事故案例按照框架法表示,在清洗去重后共有 1 220 个事故案例存储在 MongoDB 数据库中,其中 1 个具体案例的详细信息如图 2 所示。

在用户输入检索内容时,结合当前事故环境特征,调用案例检索模块功能,执行以下步骤:

1) 获取请求中的检索短语(query),同时依据请求时的状态,自动获取当前环境状态相关属性信息。

2) 对 query 短语进行分词,对短语进行尽可能细粒度地切分,同时对停止词进行标记。

3) 对分词结果通过倒排索引进行查询,查询结果为所有包含任何分词的案例编号。

4) 利用步骤 3) 中查询到的案例编号从 MongoDB 数据库中获取对应的案例内容。

5) 对候选案例数据依照案例检索算法进行综

湖南省邵阳市新邵县七四煤业有限责任公司煤与瓦斯突出事故			
事故基本信息			
事故时间	2007-07-09 00:00:00	事故地点	湖南省邵阳市新邵县七四煤业有限责任公司
事故类型	煤与瓦斯突出	事故性质	责任事故
死亡人数	2	受伤人数	0
直接损失	60	间接损失	
企业信息			
企业名称	七四煤业有限责任公司	企业性质	股份制企业
矿井信息			
煤矿级别	地方重点煤矿		
证件情况	是	审批情况	是
实际生产能力	5		
从业人员数量	160	煤矿瓦斯等级	瓦斯突出矿井
矿井开采方式	综合开拓	矿井通风方式	
事故区信息			
事故位置	回采工作面	具体位置	
煤层厚度	2.4	平均倾角	20
煤层自燃倾向	无自燃倾向	煤尘爆炸倾向	无爆炸倾向
走向长度		工作面长度	
绝对瓦斯涌出量			
坚固系数		透气性系数	
采煤工艺		顶板管理	全部陷落法

图2 事故案例的展示

Fig.2 Display of accident case

合排序。

6)对排序后的案例数据进行可视化展示,如突出显示关键词,获取相似属性信息等。

案例检索功能页面被划分为3个区域,分别为搜索区、当前事故环境信息区以及检索结果区。在搜索区内,用户可以输入搜索内容。当前事故环境信息区显示了当前发生事故或灾害的工作面以及对应矿井信息。检索出的案例展示在检索结果区,每个条目表示1个案例,多个案例按照综合指数从大到小顺次显示。每个检索结果的显示包括案例标题、案例内容、案例元信息以及案例与目标案例的相似标签。检索结果标题包括案例发生时间、地点以及事故类型,案例内容以用户检索为基础,快速截取包含用户检索信息的部分内容。案例元信息则主要指案例的排序指标,即检索指数、相似指数以及综合指数的大小,方便用户快速分析该案例可能的使用效果。通过标题下方的相似标签展示案例在不同属性维度上与目标案例的契合程度,方便用户进行快速筛选。

4 结 论

1)为了能深入分析包含在煤矿事故案例中的经验信息,提高案例的使用率和利用率,首先构建了一个通用、全面、结构化的事故案例数据库,通过基于Scrapy框架的Python网络爬虫动态爬取由各地监管部门发布在互联网上的海量的、多元的、实时的煤矿事故案例,然后将获取的案例利用框架表示法进行组织和表示,既可以存储事故的结构属性特征,

也可表示事故原因、救援、处置过程等描述性知识。

2)在事故案例数据库基础上,初步提出一种新的案例检索流程及算法,利用搜索引擎技术实现对案例非结构化信息的检索并结合传统的基于结构化属性特征的案例匹配结果,实现利用少量关键信息进行非结构化案例数据的高效筛选,可以输出符合搜索意愿并与当前事故发生环境相似的案例。但目前算法模型对输出结果尚未进行验证或评价,在下一步研究中会引入检索效果评价方法,来改进检索算法,提高检索结果的可信度。

参考文献(References):

- [1] 丁百川. 我国煤矿主要灾害事故特点及防治对策[J]. 煤炭科学技术, 2017, 45(5):109-114.
DING Baichuan. Characteristics and prevention countermeasures of major coal mine disasters in China[J]. Coal Science and Technology, 2017, 45(5):109-114.
- [2] 刘 洋, 林柏泉, 李贤忠, 等. 煤矿事故案例的安全职能分析及数据库系统设计[J]. 工矿自动化, 2012, 38(1):29-32.
LIU Yang, LIN Baiquan, LI Xianzhong, et al. Analysis of security function of coal mine accident cases and design of its database system[J]. Industry and Mine Automation, 2012, 38(1):29-32.
- [3] 李泉金. 煤矿安全事故成因分析及预警管理[J]. 工程技术:引文版, 2016(8):49-57.
LI Quanjin. Analysis on coal mine safety accident causes and forewarning management Research [J]. Engineering Technology: Citation Version, 2016(8):49-57.
- [4] 孙晋非. 突水预测预报决策支持系统关键技术研究[D]. 徐州: 中国矿业大学, 2012.
SUN Jinfei. Research on the key technologies of forecasting decision support system of mine water inrush[D]. Xuzhou: China University of Mining and Technology, 2012.
- [5] 祁运田. 煤矿安全决策支持系统的设计与实现[D]. 淮南: 安徽理工大学, 2009.
QI Yuntian. Design and implementation of coal mine safety decision support system[D]. Huainan: Anhui University of Science And Technology, 2009.
- [6] 田水承, 许永刚, 宫 雯, 等. 基于GIS的煤矿瓦斯事故应急救援系统研究[J]. 陕西煤炭, 2012, 31(4):29-31.
TIAN Shuicheng, XU Yonggang, GONG Wen, et al. Research on emergency rescue system of coal mine gas accident based on GIS [J]. Shaanxi Coal, 2012, 31(4):29-31.
- [7] 张宏丽, 乔钢柱, 曾建潮. 煤矿安全事故数据库系统的设计与开发[J]. 太原科技大学学报, 2015(4):288-293.
ZHANG Hongli, QIAO Gangzhu, ZENG Jianchao, et al. Design and development of coal mine safety accident database system[J]. Journal of Taiyuan University of Science and Technology, 2015 (4):288-293.
- [8] 孙晋非. 突水预测预报决策支持系统关键技术研究[D]. 徐州: 中国矿业大学, 2012.
SUN Jinfei. Research on key technologies of decision support

- system for water inrush forecasting and forecasting[D]. Xuzhou: China University of Mining and Technology, 2012.
- [9] 李艳妮. 基于 WebGIS 的煤矿应急救援系统设计与实现[D]. 邯郸:河北工程大学, 2015.
- LI Yanni. Design and implementation of coal mine emergency rescue system based on web GIS[D]. Handan: Hebei University of Engineering, 2015.
- [10] 安子建. 基于 Scrapy 框架的网络爬虫实现与数据抓取分析[D]. 长春:吉林大学, 2017.
- AN Zijian. Implementation and data crawling analysis of web crawler based on Scrapy framework[D]. Changchun: Jilin University, 2017.
- [11] 杨定中, 赵 刚, 王 泰. 网络爬虫在 Web 信息搜索与数据挖掘中应用[J]. 计算机工程与设计, 2009, 30(24):5658-5662.
- YANG Dingzhong, ZHAO Gang, WANG Tai. Application of WebCrawler in information search and data mining[J]. Computer Engineering and Design, 2009, 30(24):5658-5662.
- [12] MINSKYM. A framework for representing knowledge[J]. Readings in Cognitive Science, 1988, 20(3):156-189.
- [13] 翟丹妮, 黄卫东. 应急案例的框架表示方法研究[J]. 计算机技术与发展, 2011, 21(7):33-38.
- ZHAI Danni, HUANG Weidong. Frame-based representation of cases in emergency management[J]. Computer Technology and Development, 2011, 21(7):33-38.
- [14] 谷 岩. 基于框架和案例推理的应急预案表示和优选方法的研究设计[J]. 计算机科学, 2012, 39(6):163-165.
- GU Yan. Design of emergency preplan expression and optimization method based on frame and case-based reasoning[J]. Computer Science, 2012, 39(6):163-165.
- [15] 张丽圆. 基于案例推理的煤矿瓦斯应急决策研究与应用[D]. 徐州:中国矿业大学, 2014.
- ZHANG Liyuan. Research and application of case-based reasoning in coal mine gas emergency decision-making[D]. Xuzhou: China University of Mining and Technology, 2014.
- [16] YANG J W. An Index structure of semi-structure data set for similarity search[J]. Chinese Journal of Computers, 2002, 25(11):1219-1226.