基于大数据的建筑区域火灾事故因素分析

【摘 要】雇佣劳动部每年都会对施工现场发生的过失事故进行汇总，并发表“产业灾害现状”报告书。本报告汇总了特定行业的事故和死亡率、灾害分类和统计数据，但其有效性很低。这是由于它没有充分说明事故的直接原因或有关其因果关系的信息。然而，本研究利用近年来受到业界和学术界广泛关注的大数据方法，收集近十年来发生在建筑工地的火灾事故的互联网文章。此外，通过主成分分析，根据时间、地点、诱因和事故模式推导出季节特异性因子。在此基础上，针对常见因素，推导出直接火花和油雾。火灾事故是由于工作因素造成的，疏忽的监督和违反安全法规是造成火灾事故的原因，说明火灾事故是人为的。据调查，在火灾发生时，还会发生倒塌、掩埋、爆炸、窒息等次生事故。本研究中使用的大数据分析方法被认为是非常有效的，可以在未来成功地用于推断大量的文本数据。

引言

随着信息通信技术的发展，大数据受到越来越多的关注。大数据的处理和分析能力表明了其未来的竞争力，2012年世界经济论坛将大数据作为最显著的技术进步。根据这一趋势，在规模越来越大、复杂程度越来越高、专业性越来越强的现有建设现场，需要和产生大量的信息，开发用于管理和处理这些现场的相应系统受到了极大的关注。

韩国职业安全健康院(KOSHA)每年都会制定产业灾难原因调查报告。本报告分析了调查后对工作事故原因的调查结果和工作事故的过程。但是，统计数据只提供了事故的行业、规模、发生类型、原因事项等方面的统计，而没有提供每一种事故类型的详细原因。例如，数据没有提供火灾事故发生的时间、季节、地点和工作的细节。因此，利用大数据分析技术推导出统计中未提供的火灾事故的详细原因是十分恰当的。

本研究采用工业界和学术界普遍关注的大数据分析方法，以季节为单位，收集近十年火灾事故的新闻信息，利用主成分分析(PCA)推导火灾事故的季节特异性因子。

最近开发出了利用信息技术应用于施工现场的“泛在传感器网络(USN)”为基础的施工现场安全管理监视系统。此外，以学习者为中心的工具已被开发，以提供建筑安全教育在大学。然而，利用建筑工地产生的大量信息的研究还不多。因此，作为本研究核心要素的大数据方法是一种通过收集大量信息并利用现有统计分析发现新的事实、规则或可预测模式的方法。

大数据方法可以分为社交网络服务(SNS)收集数据时分析模式和趋势的“社会大数据”、分析地区和空间特定数据的“空间大数据”、分析事物之间的传感器数据的“物联网”。 针对建筑行业社会大数据的研究，Choi 等人开发了社交板块用于收集社交媒体数据，并根据灾害发生情况分析特定地区的位置披露信息，从而建立可靠的灾害响应系统。Bae等人也进行了爆炸检测实验，通过分析社会大数据中的句法模式来检测包括爆炸在内的人为灾难。

针对物联网与大数据相关的研究，澳大利亚一家建筑公司开发并实施了大数据系统，以防止矿山或建筑工地的疏忽事故。

通过对施工现场生成的数据进行实时分析，该系统为工人提供了风险预警。该系统分析的数据包括现有事故记录和日常任务报告、闭路电视图像以及各种危险因素的位置数据。此外，还使用了基于文本和视频分析的各种预测模型。

Shin等人通过分析火灾爆炸灾害的预防数据，确定了化工材料火灾爆炸危险的优先级，提出了中小型场地防灾的指导、监督和安全教育方法。Shin提出焊接、加热器和电力是目前引起火灾爆炸的主要因素，各因素占78%，并建议将输配电、机械设备和化工产品作为引起火灾爆炸的材料。Oh研究了7个主要行业的火灾因素，并针对公寓建筑项目制定了针对性的预防措施。为此，通过对灾害案例的分析，制定了对策，并根据工种、易燃物、易燃物等分析了工作进度、火灾危险因素，提出了预防措施。如前所述，火灾事故原因分析的研究有只分析火灾原因材料和各个工作过程的研究。然而，目前还没有一项研究分析火灾事故的各种火灾原因(如时间、季节和官方管理等)。因此，本研究旨在通过收集建筑现场积累的新闻数据，运用大数据分析方法，分析火灾事故的火灾原因。

本研究采用R统计分析程序，由信息收集、文本和数据挖掘两部分组成。在信息收集方面，利用R图书馆收集了SBS网上有关近10年施工现场事故的报道，并通过网络爬行建模，收集了需要的信息。另外，收集了2000年1月1日至2014年12月31日期间的网络新闻文章，并以“建筑工地过失事故”为关键词。

对于文本挖掘，在测量频率之前进行形态学分析，将互联网新闻文章的句子提取为词汇格式，排除后位置和后缀，将它们处理为一个有用的格式，然后进行数据挖掘。经常测量的词是包括表示施工现场发生的事故类型和事故原因的词。

在文本挖掘过程中，根据各种事故类型的出现频率，推断出近10年来建筑工地经常发生的6种疏忽事故(火灾、爆炸、倒塌、重型设备问题、坠落、窒息)。被认为影响这类事故经常出现的因素共有170个。然后将该集合命名为因子字典，用于收集事故特定过失报告时进行文本挖掘，有效地提取与过失事故相关的单词。然后对通过文本挖掘识别出的建筑工地频繁发生的火灾事故进行数据挖掘。经常出现的词是包括被认为是影响这类火灾事故的因素和妨碍分析的因素。在使用因子字典去除不适合本研究的因子词后，对过滤结果进行主成分分析(PCA)，推导出被认为影响火灾事故的因素。

2 理论考虑

2.1大数据系统

考虑到大数据分析的特点，有必要建立一个能够在短时间内实时处理和分析大量信息的系统。图1是政府机构和企业建立最频繁的基于Hadoop的大数据系统。Hadoop是一个开源框架，它支持分布式应用程序，这些应用程序在大规模计算机集群中运行，以进行大容量处理。

在Hadoop支持的功能中，MapReduce是一种利用多节点PC单元并行处理的方式收集大容量数据的功能。因为Hadoop是一个专门用于信息收集的程序，所以它需要一个单独的分析程序来分析信息。应用最广泛的程序是Hadoop HIVE和开源统计程序R，这些程序为不熟悉统计的用户提供了适合的聚类分析、模式分析、趋势分析和预测分析算法，可以方便地获得结果。

2.2 大数据分析程序R

R软件是美国私营研发机构贝尔实验室在S语言的基础上开发的一种编程语言。它是作为自由软件的一部分开发的，即一个大规模协作项目，作为一种用于应用统计计算和图形的编程语言。虽然R在过去没有得到广泛的应用，但近年来随着大数据的发展趋势，R作为一个综合R档案网络的软件包，在各个领域作为大数据分析工具被频繁使用。通过R包集成SNS和云数据，通过Wordcloud可视化方法，是分析和表达大数据的理想工具。

2.3 网络爬行

作为搜索引擎的基础，网络爬虫是一种跟踪互联网上的网络文档以收集必要信息的技术，是大多数互联网行业使用的核心技术，包括互联网搜索系统如Yahoo!电子商务产品搜索系统。网络爬虫是一种通过流通的网络服务器收集各种网站上发布的各种类型的信息的程序。该程序无需重复跟踪每个网站链接来获取信息，而是自动分析连接到每个URL的Web页面的内容。

要使用网页抓取，首先要选择一个包含所需信息的网站。由于网络爬行的基本概念是搜索选定的网站，它需要由易于访问的安全组成来实现。此外，由简单的超文本标记语言(HTML)结构组成，可以在较短的时间内获得高采集率的网络爬行结果。网络爬行的重要方面是确定网页搜索的范围，从中收集信息，以及确定信息收集的周期。

网络爬行的使用可以分为选择整个网站作为搜索范围的情况以及通过URL模式的输入，将搜索集中在一个特定的主题上，或者只搜索特定的URL。为了从收集的url中获得所需的信息，需要进行单独的解析。解析是分析HTML结构以实现有效编程的过程。最终收集到的数据被转换成文本文件并保存。为了实现网络爬行，本研究使用的的核心技术是R库。

2.4 文本挖掘

文本挖掘是指利用机器学习和统计领域的一种算法，从大量的文档中找到有用的模式的过程。由于数据挖掘方法应用于文本，故又称文本挖掘。然而，与现有的数据挖掘技术不同的是，与编程和控制语言相反，非标准化数据没有得到应用。

与文本挖掘相关的研究领域包括信息搜索、自然语言处理和信息抽取。文档数据挖掘方法包括一种基于给定关键字进行文档分配的分类方法和一种无需预先提供信息即可绑定相似文档的聚类方法。文档分类方法包括使用决策树、贝叶斯分类、最近邻分类器和支持向量机。最后，聚类方法包括层次聚类、K-means、SOM和EM。

2.5 数据挖掘

数据挖掘分析的复杂性很高，因为要处理的数据量非常大，其中典型数据的比例很高。在统计现有数据的基础上，大多数分析方法采用分析方法中使用的算法，通过改进它们来处理大量数据，进行大数据处理分析。实时处理大量数据的分析技术包括:文本挖掘、在提取非典型句子的含义时建立提取信息之间的假设;意见挖掘、声誉指数分析、区分用户对特定服务和产品的意见;社交网络分析，以口碑为中心考察用户;聚类分析，通过分析相似度高的目标群体与其他聚类对象之间的差异来推断新的用户群体。作为一个通过打包实现商业化的分析工具的例子Yahoo!，它的优点是能够通过分布式集群技术的应用快速、可靠地对标准化和非标准化数据进行分析。此外，谷歌的大查询和亚马逊的发电机已经开发。

2.6 主成分分析

PCA是一种降低数据维数的方法，同时在组成数据集时保持相关变量的变化。例如，它可以任意定义为由vector中找到的维度随机变量组成的向量，然后所有数据都可以表示为一个矩阵。PCA的第一个目标是找到在归一化下使X的离散度最大化的参数，即找到满足以下最大化问题的参数。

然后，结果成为第一个主分量。接下来，找到与第一个主分量无关的离散度最大化的向量。随后，当找到它作为下面的最大化问题的解时，它就成为第二个主分量。通过连续应用这种方法，可以找到后续问题的主分量，但它通常停留在寻找变量数量中的主成分上，因为PCA的目标是降低数据的维数。在主成分分析中，假定每个主成分都是一种无法观察到的潜在变量，然而假设这些主成分可以通过观察到的变量线性组合来表示。此外，主成分非常有用，可以考虑分别显示不同的统计维度，因为根据定义主成分之间没有相关性。

3 基于大数据分析的过失事故因素分析

3.1 大数据系统的发展

因此有必要开发一个系统，利用文本和数据挖掘以收集近十年来发生在建筑工地的过失事故的文章，有效地推断出疏忽事故的因素。因此，本系统采用开源统计分析程序R进行开发如图3所示。

R的优点是允许开发人员基于具有许多用户的社区和基于其开源特性支持的各种库轻松获得新的分析方法和技术。

3.2过失事故信息的收集

为了在Web上获取所需的信息，使用R库进行Web爬行建模如图4所示。为了确保有效的信息收集，选择了具有相对容易访问的HTML结构的SBS新闻网站。为了从SBS新闻网站收集过去10年里发生在施工现场的过失事故的新闻信息，使用了“施工现场”的关键词，从2000年1月1日到2014年12月31日，共收集了2.8263万篇网络新闻。为了有效地抓取网页，理解URL和HTML结构也是必要的。

SBS新闻的URL结构由SBS域、页码、搜索日期条件和搜索词代码组成。通过简单地修改搜索条件和关键字代码，不需要连接到网站，就可以根据修改的关键字收集互联网文章的搜索结果。为了在数年内获得基于相同关键词的互联网文章搜索结果，也可以进行基于年份的网页爬行。为了收集本研究中使用的信息，我们将过去十年中发生在建筑工地的疏忽事故的新闻文章按年份进行分类。

采集率高的原因是网络爬行的一个重要方面，如图5所示，没有年度分类的序列结构网络爬行的采集率为27%，而有年度分类的平行结构网络爬行的采集率为98%。基于此，本研究采用平行结构的网络爬行模型，共收集18868篇网络新闻文章中的18490篇，收集率为98%。此外，文章根据出版年份保存为文本文件。

3.3 文本挖掘

3.3.1 事故类型

通过网络抓取收集的新闻文章是由句子组成的。利用R库和数量特征，将文章分为词，得到词的出现频率。参考Kim的研究，建筑工地发生的事故类型分为倒塌、重型设备问题、窒息、坠落、爆炸和火灾。剩下的部分被认为是施工现场的过失事故的原因。表1显示了提到的单词的频率,事故类型提到增加周期的频率接近当前日期,这被认为是由于被记录的信息量的增加通过媒体和电信技术的进步,而不是增加事故发生的频率。

在表1中，近十年建筑工地最常发生的过失事故为火灾事故，被提及1077次，占总数的33.11%。然而，在就业和劳动部的工业灾害状况报告中，坠落事故占总数的72%，而火灾事故只占1%。造成这种反差的原因可能是，坠落虽然经常发生，但没有像火灾那样受到媒体的关注，而火灾通常规模更大，造成的损失更大。因此，本研究认为火灾是有效收集信息和分析各种因素的合适对象。

3.3.2 因子字典

根据文本挖掘可以获得的信息类型以及施工现场发生的过失事故类型，得出了造成此类事故的因素。一篇新闻文章的特点在于它真实地记录了事故的真实情况。新闻文章是目前开发因子词典的重要数据库，因为它们包含了导致事故因素的类型和频率。“因子字典”是对建筑工地发生的事故有直接/间接影响的高频率因素的集合。

这有助于我们在分析施工现场的过失事故时，制定这些因素的标准和范围。表2显示了因子字典中因子的分类。自然分类分为季节和天气的兴趣地点。时间分类分为星期和发生的时间。空间分类分为发生地和区位。工人行为分类分为工作、任务和物理变化。场地条件分类分为气源、类型、动态和产状。重型装备分类分为重型装备的类型。其余的被归类为其他因素。

3.4 收集火警事故资料

通过前期的文本挖掘，发现火灾事故是施工现场最常见的过失事故。

因此，本研究旨在透过网路抓取新闻报导，借由主成分分析来推断过去十年建筑工地发生火灾事故的各种因素。考虑到即使是同一类型的事故，引发事故的因素也随着季节的不同而不同，所以在网上收集了有关火灾事故的新闻报道。将“冬季1号”和“冬季2号”单独归类为搜索条件的原因是，“冬季1号”是融化季节，“冬季2号”是冰冻季节，因此造成事故的因素会有所不同。

此外，输入如表3所示的URL结构，使用特定年份的火灾事故关键字，搜索互联网文章12019篇，通过平行结构模型，共收集到10978篇，占91.3%，如图6所示。造成这一比率的原因是，本研究中使用的网络爬行模型固有地有2%的收集遗漏率，从而导致累积遗漏率。

3.5 火灾事故文本挖掘

通过网络爬行获得的10798篇关于建筑工地的网络新闻文章，再次按照季节进行汇总，并处理成适合数据挖掘的文字。利用R库及其特征，通过形态学分析将由句子组成的网络新闻文章分类成词，并计数测量火灾事故因素的频率。由于火灾事故因素包括那些直接和/或间接涉及造成火灾事故和阻碍分析的因素，因此使用因子字典进行过滤。

据此，将火灾事故因素保存到适合编程的CSV文件中，并利用Excel内置函数根据火灾事故发生频率推断出可以代表火灾事故直接因素和/或间接因素的词。在过滤原则上，将施工现场过失事故造成因素的集合与火灾事故因素词之间的对应词视为火灾事故造成因素。此外，在抽样充分性检验的Kaiser-MeyerOlkin测度中，剔除了未计入词和累积频率低的词，得到了有利值与所收集的数据有显著的相关性。消去准则是在将高频词按顺序排列后，将累积频率小于90%的词剔除。表4至表8总结了通过文本挖掘推导出的因素。

3.6 火灾事故数据挖掘

在火灾事故文本挖掘得到的数据中，包含了季节性火灾事故成因因素和发生频率。由于认为高频因素不足以令人信服地认为是火灾事故的原因，使用PCA来推导季节特定火灾事故的主成分。因此，在进行主成分分析之前，需要验证数据的有效性。在本研究中，我们使用最小标准为0.6的样本适应度测度来验证所收集数据的有效性。然后旋转欧几里得坐标轴，看清楚数据和提取的主成分之间的关系。应用了Varimax旋转，它是一种易于分析和避免多重共线性问题的正交旋转。选择主成分的方法有多种，但以kaiser的高值1或更高和累积色散比为标准。在本研究中，各季节的主成分结果如表4 ~ 8所示，分别只显示主成分的显著因子。此外，累积离散率大于80%，说明季节性火灾更具解释性.

1~ 2月施工期间火灾事故的首要因素是黎明、上午和下午，气体、有毒气体和火焰也是火灾发生的主要因素。据调查火灾发生在建筑工地，烧毁了重型设备、设施和机器。第二主成分的时间因子为早晨、干燥和零度以下的天气，主要发生在公寓。第三个因素是建筑工地或塔楼设施消防设施不足。

3 - 5月施工期间火灾安全事故第一主体的时间因素为上午和下午，在控制不顺利的情况下，在施工现场和地铁工作时，由气体或火焰引起的火灾。第二主要成分是对施工现场进行非法、不合理的改造。第三个主要因素是重建场地设施的老化.

6 - 8月施工现场火灾安全事故第一个主要组成部分的时间因素是上午，发现火灾是由于工作场所的气体、火焰和机械引起的。第二种主要成分在地铁施工现场是非法的，第三种主要成分在下午出现在公寓施工现场.

9、10月份的第一个主要原因是屋顶和高层建筑的非法施工、不合理的施工造成的损坏或倒塌、煤气等。第二主构件在施工现场出现裂缝和缺陷，第三主构件的时间因素出现在上午。

11、12月施工现场火灾事故的第一个主要组成部分是塔楼施工现场、夹层板、干燥的天气、香烟、煤气、撞击。第二个主要组成部分是高层建筑和地铁的不稳定控制。第三个主要组成部分似乎是非法行为和风的领域。

每个季节的主要特征总结如下:在所有季节中，发现火灾是由于气体和火焰之间的接触而发生的。正在进行施工的春季、夏季、秋季等地，由于施工或现场控制不足发生了火灾。

另外，在开发现场的消防和设施管理也不完善，因此引发了火灾。结果发现，由于监理人员缺乏监督，工人违反安全规则，施工现场发生了火灾事故。

4 结论

本研究采用工业界和学术界普遍关注的大数据分析方法，以季节为单位，收集近十年火灾事故的新闻信息，利用主成分分析(PCA)推导火灾事故的季节特异性因子。常见因素为火灾中直接气体和火焰，季节性因素为冬季低温和干燥天气。研究发现，工作因素主要是由于主管人员的疏忽和违反安全法规造成的。此外，还发现火灾和设施管理不足，导致火灾发生。认为有效利用基于本研究的大数据分析方法，可以从多个角度分析过失事故因素。只是本研究的局限性在于无法了解火灾事故的密集信息，也无法通过事故统计来验证网络爬行的有效性