基于 GIS 的敏感词网络数据分析

伍爽
（武警警官学院　四川　成都　610213）

摘要：随着社会经济的发展和进步，城市规模的不断扩大，许多矛盾相继到来，各种突发事件不断发生。应对突发事件，预见其发生至关重要。本文拟从网络数据中的敏感词数据入手，每一个敏感词网络数据都对应着一个主机的 IP 地址，建立敏感词网络数据库和与之相关的检测技术。检测到这一类敏感词时，根据其对应的 IP 地址映射出实际地址，用地理信息系统将其在地图中标注出来，达到预警的作用。

关键词：突发事件；敏感词检测；GIS

随着大数据时代的发展，网络数据的变化也一定程度的反映出事件的发展。相关研究表明，网络数据与事件是有关联的。利用这一点，找出其中的敏感词数据信息，而每一个敏感词数据都关联着其主机的 IP 地址,可以通过主机的 IP 地址用 GIS 将其映射到实际地图上。通过对这些敏感词数据的研究，找出数据异常，为发现和处置突发事件奠定基础。

1. 软件架构

本文设计主要是以 Arcgis Engine 组件为平台，运用了敏感词数据检测和点数据分析，在地理信息系统的基础上，结合对敏感词网络数据分析的需求，来实现对突发事件的预警和帮助处置效果。本设计主要分为两大块，一是敏感词数据的检测，二是点模式分析。

（1）敏感词数据检测

在平常未发生突发事件时，可以通过收集日常的网络数据与敏感词数据库进行比对，确定敏感词数据多的区域，对其进行重点的观察。在发生突发事件后，通过收集敏感词数据信息，立即获取数据集中区域，根据数据汇聚情况作出决策。在处置的同时，也可以收集数据，通过数据的波动趋势，可以预测事件的走向，为下一步指挥和决策提供便利[1]。

（2）点模式分析

敏感词数据监测中得到的主机 IP 地址地理位置信息，只是分散一个个点的数据，不便于分析趋势，本文采用 GIS 中点模式的分析方法。分析点的聚集、分散以及密集程度。分析每个点与其他各个点之间的距离，找出其中最短的距离，这个距离被称作为最邻近距离，通常用d_{min}来表示。1954 年 Clark 和 Evans 提出平均最邻近距离（$\bar{d}_{min}$），公式表达为：

$$\bar{d}_{min} = \frac{1}{n}\sum_{i=1}^{n} d_{min}(s_i) \quad (1\text{-}1)$$

平均最邻近距离的提出对研究点的分布情况有着很重要的作用，Clark 和 Evans 提出平均最邻近距离（$\bar{d}_{min}$）的同时，也提出对点分布情况很有用的一个概念—最邻近指数 R，公式表达为$R = \frac{d_{min}}{D}$，其中分母 D 是理论的平均最邻近距离，其计算公式为$D = \frac{1}{2\sqrt{p}}$，其中的 p 是指所有点所在区域的密度，计算公式为:点数/面积。当 R>1,点为均匀分布,R=1,点为随即分布,R<1,点为聚集分布[2]。根据 R 的大小就能判断当前敏感词发展的态势，据此设计软件的流程如图 1-1：

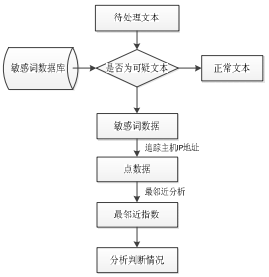

图 1-1 软件总体架构　图 1-2 最邻近指数分析

2. 敏感词数据库

2.1 敏感词数据库建立

敏感词数据库的建立，能够更快的识别和检测敏感词数据中的关键字词。由于数据模式相对简单，本文使用的是 SQLite 数据库软件[4]，将敏感词以 string 的形式存储在数据库中。本文建立的敏感词数据库几乎都是中国汉字，而在计算机中，汉字的存储方式，是将汉字转换为 Unicode 码。而 Unicod 编码都是使用的字符串来表示，在 SQLite 数据库中使用 STRING 字符串作为敏感词数据的存储类型，会大大减少处理工作。

2.2 敏感词匹配

敏感词匹配主要通过 DFA 算法来实现。DFA 算法主要是指从一个状态通过一系列事件转换到另一个状态。应用在敏感词数据检测，则为通过检测，使敏感词数据从初始状态（第一个字）转换到终止状态（最后一个字）。

整个敏感词数据检测可以通过 Python 代码结合 DFA 算法实现，将文本输入后，先检测文字是否包含敏感词数据库里的数据，若是存在，则用“*”将敏感词数据的文本代替，然后输出结果。

2.3 纯真 IP 数据库

研究敏感词数据，根据敏感词对应的 IP 信息来确定这些敏感词数据的地理位置信息，本文拟使用纯真 IP 数据库[3]，它收集包含三大通讯公司和其他一些公司的 ISP 的最新的 IP 地址数据信息。当前版本的纯真 IP 数据库里面的记录到达了五十二万多条。通过网络数据分析原理得到的主机 IP 地址，可以将其与纯真 IP 数据库里的 IP 地址数据进行对比，从而得到该主机 IP 地址所在的位置信息以及对其负责任的

（下转第 33 页）

突出，机械制造难以实现与时俱进的更新和发展。在工艺设计方法的引进和实施中，部分设计人员如不能可靠的接受新的技术和管理手段，必然导致工艺设计的理念存在落后问题，导致港口机械制造能力在未来发展的竞争中，存在落后问题。

5. 港口机械制造技术的发展趋势

5.1 信息数字化的发展

随着信息数字化技术的不断发展，其在各个行业领域的技术应用中，都表现出较高的信息化数字应用价值，很多工作的开展要能做到在技术应用过程中，获得较高的信息化管理能力，获得港口机械制造过程中的存在信息，全面保障信息数字化应用价值的提升，指导有关港口机械制造技术的实施和调整。在信息数字化技术的发展推广中，首先应该能提高技术和管理人员的认知理念，对港口机械制造各个技术应用环节和具体流程内容有效分解，深入研究各种技术应用的基本需求和数字化管理理念，提升港口机械制造优势，获得较高的技术应用调整参考价值。信息数字化技术的应用，能在港口机械制造技术的实施过程中，获得突出的制造技术环节流程管控优势，一般在设计工作开展中，获得各种港口机械制造技术评价管控力，以信息化整合技术、大数据分析技术作为信息技术获取和分析的切入点，不断提高信息技术的应用价值，对港口机械制造技术的应用效果可靠检测，以信息技术指标资源，获得信息的评价反馈能力，提升信息技术的实施价值。一般在未来的信息化数字化技术应用中，为提升信息资源的整合和共享分析能力，需要做到在信息技术的传输过程中，搭建内部信息分析交流平台，集中、专业的做好信息资源数据的分析处理，提升管理指导能力，是吸纳管理资源的合理配置，全面保障管理资源的应用效率提升，管理优势提高。应能在信息数字技术的应用中获得较高的港口机械制造技术管理协调能力，提升管理发展的科学性，全面获得管理上的联系互动改善优势，提高技术应用价值。

4.2 自动化水平提升

在港口机械制造技术的实施中，需要满足其作业内容的需要，在工作开展中，能提高机械设备的可靠应用提升优势，以自动化集成技术，作为港口机械制造流程的主要实施方式，提升港口机械制造的实际工作效率，同时能可靠控制港口机械制造作业质量，提高机械制造的各个流程环节顺利开展，提升自动化水平，获得突出的港口机械制造过程管控能力，以自动化技术的应用，控制生产成本，提高生产控制优势。而在自动化技术的应用过程中，应能可靠保障认知理念的转变，经济在自动化技术的应用中，及时更新有关技术手段，能做好自动化改进的资金和技术人才管控能力，做好人才储备工作，提高人才的价值水平，全面获得自动化水平提升建设效果。机械企业自动化设备的引进，需要做好可靠的港口机械制造调研，需要提高技术的应用价值，获得技术实施的可靠管控优势，调整参数配置，提升实际应用优势。

4.3 节能环保型发展

港口机械制造技术的应用要能积极响应国家的有关的生态环保政策，在技术的发展中，只有符合国家有关政策法律的规范要求，才能保障港口机械制造技术可靠实施，同时有关技术获得群众的支持，避免对当地居民的生活起到不良的环境影响。绿色港口机械制造技术的实施也是时代发展的必然趋势和要求，只有充分保障具备较高的节能环保管控优势，才能保障在技术应用层面上，符合可持续化发展要求，提高节能环保优势，推动高效节约型技术的引入和实施，港口机械制造过程中的环境问题隐患。

6. 结语

港口机械制造技术的应用发展，应能以提升技术的应用效率和制造质量为导向，不断提升港口机械制造技术应用的总体竞争力，科学布局。优化管理和设计工艺流程，提高自动化技术应用水平，构建环保型技术应用模式。

参考文献：

[1]万江.港口机械制造工艺发展现状与未来发展趋势[J].中国高新区，2017（20）.

[2]张桂霞.论机械制造工艺发展现状与未来发展趋势[J].电子制作，2014（2）：266.

[3]全彦军，韩建普.机械制造工艺的发展现状及未来发展趋势[J].自然科学：全文版：00224.

[4]朱骞彬.微型机械加工技术发展现状和趋势及其关键技术[J].精密制造与自动化，2002（2）：9-11.

[5]汪亚非.港口装卸机械现代设计方法的应用与发展[J].物流技术，2001（6）：5-7.

（上接第 31 页）

代理商，最后获得相关部门许可，通过代理商追踪其数据包的来源，从而确定好主机 IP 地址的 XY 坐标位置[3]。

3. 最邻近指数分析

本文利用 arcgis 集成了最邻近指数分析功能，输入多个 ip 数据对应的物理位置坐标后，可对其最邻近指数进行分析如图 1-2 所示：每个小点代表的是一个 ip 地址，根据点与点之间的距离关系能得出最邻近距离和最邻近指数，若最邻近指数小于 1 则认为出现点聚集情况，考虑有突发事件的可能[5]。

4. 总结

本文设计了一个敏感词识别，将敏感词对应的 ip 地址映射到物理地址并分析其聚集情况的软件，在分析突发事件和社会舆论时有一定作用。

参考文献：

[1]伍玉英. 基于网页文本的敏感信息检测系统研究[D]. 重庆大学,2014

[2]汤国安，杨昕著. 地理信息系统空间分析实验教程（第二版）[M]. 科学出版社，2019.10.

[3]刘智敏,赵虹.基于 GIS 的群体性事件网络数据分析[J].湖南警察学院学报,2017,29(03):78-83.

[4]岑冬梅. 基于 SQLite 的空间数据库存储技术的研究与实现[D]. 武汉科技大学,2009.

[5]夏松,林荣蓉,刘勘.网络谣言敏感词库的构建研究——以新浪微博谣言为例[J]. 知识管理论坛,2019,4(05):267-275.