大数据采集技术的实现

Realization of Big Data Acquisition Technology

杨代全

Daiquan Yang

贵州省数据信息室

中国·贵州 贵阳 550002

Guizhou Provincial Data Information Room,

Guiyang, Guizhou, 550002, China

【摘要】21 世纪是信息的时代,大数据技术的发展对经济、军事领域都造成了很大的影响,特别是信息的采集和处理已经成为大数据能否有效利用的前提和基础。因此,针对大数据采集的有效路径进行研究,实现高效的大数据采集技术显得至关重要。

【Abstract】The 21st century is the era of information, and the development of big data technology has had a great impact on the economic and military fields, especially the collection and processing of information has become the premise and basis for the effective use of big data. Therefore, it is very important to study the effective path of large data acquisition and to realize the efficient large data acquisition technology.

【关键词】大数据采集技术;高效;网络大数据采集

【Keywords】big data acquisition technology; high efficiency; network big data acquisition

【DOI】10.36012/etr.v1i3.422

1 大数据技术的发展

大数据的未来发展和应用趋势主要体现在民用大数据、军事大数据发展等领域,如何实现大数据信息高效采集与处理,基于任务进行数据融合,对有用信息进行抽取、清洗与挖掘等关键技术,同时保证大数据的隐私和安全,这些都是大数据未来发展的重点方向。在大数据的不断发展之中,数据规模将逐步扩大,而对数据的采集和处理将显得至关重要。在大数据开发的背景下,将产生一大批民用、军用领域新兴行业,并增设数据信息方面相应的部门,同时,创建出大数据采集、大数据分析、大数据应用开发、大数据系统研发等新兴岗位,这些新兴部门和岗位的出现让大数据应用更加专业化,使民用大数据、军事大数据预测分析、辅助决策等方面更加精准高效。随着数据和信息资源的发展,大数据将成为国家和地区最具竞争力的战略资源,将成为新的竞争优势和新的发展动力。

2 大数据采集

要将大数据作为一种资源,就必须对大数据进行相关的采集工作。随着信息技术的发展,大数据的信息采集渠道大大拓宽,采集手段逐步增多,可从网络、空间、地面、地下、实时传感器获取多维度、多角度、多类型等远近期海量数据。数据的内容不仅包括数字、字符等格式化信息,也包括文本、图形、图像、声音、视频等非格式化信息。只有采集到足够多的有用信

息,并进行分析处理,才能实现这些信息的高效利用,因此,必须重视对大数据的采集工作。由于大数据具有不同的类型和特征,在具体的大数据采集工程中必须按照实际情况采用不同的采集方法,以提高大数据采集的效率。

3 大数据采集技术的实现

3.1 传感器采集技术

传感器采集技术在民用、军事方面的应用都非常广泛,通过在设备上安装有线、无线传感器,实现对设备的全天候实时监测。传感器通过测试对象的物理特性,通常包括体积、温湿度、电压、压力等物理信息,在采集这些信息之后,将数值转换成计算机可以准确识别的信号,然后加载到数字终端中执行导入并完成数据的采集工作,通过数值的变化判断测试对象的状态变化,通过专家分析库进行状态分析。随着技术不断的发展进步,传感器也实现了有线、无线、可视化的转变,随着 5G 等技术的应用,采集传输速率将会不断提高,传感器的海量数据的采集也能实时进行。

3.2 系统日志采集技术

数据源系统可以从系统日志文件生成数据,以实时记录各种数据源操作,如用户访问行为和实时网络流量监视以及一些网络服务器注册的金融软件股票等。许多互联网公司都有自己的大数据收集工具,主要用于收集系统日志,如 Hadoop 的 Chukwa、Flume de Cloudera、Facebook 的 Scribe 等,

所有这些工具都使用分布式架构来满足每秒数百兆的数据采集和传输要求^[4]。

3.3 爬虫

爬虫是一种为搜索引擎下载和存储网页的程序,是搜索引擎和 Web 缓存的主要数据收集方法,通过网络抓取工具或网站 API 从网站获取数据信息。爬虫从网页中提取非结构化数据,将其存储为统一的本地数据文件并以结构化方式存储;支持文件、音频、视频、其他文件或附件的信息采集,并且附件和文本可以自动关联;数据类型非常复杂,生成数据的方法和路径也在增加,数据的表示自然更加多样化。而该技术可以实现从简单的文本、图像和视频到更复杂的预留和抽象价值信息的采集。

3.4 过滤器

高效的数据采集工作的关键是 Bloom 过滤器。链接过滤技术的核心是确定当前链接是否在已经被网络抓取的链接中。从网站收集大数据时,可以添加 Bloom 过滤器以完成链接过滤。使用 Bloom 过滤器的方法是,当一个新项添加到集合中时,X 映射点可用于实现项的 X 映射点数组,并且将这些矩阵设置为 1。在此过程中搜索,只需要查看这些点是否为 1,基本上可以判断集合中是否有任何新元素。如果这些点集中存在 0,则检测到的元素不得存在;如果它们是 1,则检测到的元素可能存在于这些点集中。Bloom 过滤器在水平和垂直方向都具有很大的优势:①在大数据采集的简便性方面,Bloom 过滤器的简便性已经达到最大值,并且过滤器的存储空间和插入和查询时间是恒定的;②在关系映射之中,散列函数没有连接关系,这导致每个硬件的并行操作,有效地提高了采集工作的效率;③在存储方面,Bloom 过滤器不存储元素本身,这是 Bloom 过滤器在高机密性要求的情况下的优势。Bloom 过滤器使用已经提取的 url 执行,由 X 哈希函数计算,得到 X 值,然

后将其与位数组的 X 位置元素相关联(元素的值为已经建立 1)试验可以按照上面讨论的方法进行。

3.5 网络大数据采集

基于网络的大数据采集的一般框架包括 6 个主要部分:网页、链接提取、链接过滤、内容提取、跟踪 URL 队列和数据。这些部分的功能是:网站页面的功能是获取网站的网站内容;链接提取是提取网站页面上的链接地址;链接的过滤是为了区分链接地址下的网页内容是否已被提取;内容提取是从网页内容中提取所需的属性,然后提取所需的内容值;URL 队列是为抓取工具提供捕获数据网站所需的 URL;数据主要包括站点的 URL,已获取数据网站的 URL、Spider URL,已获取网站点数据的 URL 和 Spider Content,已获取网页的内容。

网络大数据采集的实现包括写入 Url Queue 程序中跟踪的网站的 URL 信息;获取 URL 队列中跟踪的网站信息;收集特定网站的网页内容;提取网页链接地址的内容;阅读数据库中跟踪的网站的网址;过滤 URL,比较抓取前后的 URL 信息;对已获取的 Spider Url 数据库中尚未获取的 Web 地址进行排序和处理放置网页地址获取,获取网页地址内容,提取内容值将所需属性写入数据库中。

4 结语

大数据的采集技术复杂多样,因此,必须加强信息计算方法与信息挖掘技术的研究,从根本上提高大数据采集工作的效率和质量,针对不同的数据结构类型,采取相应的采集方法进行数据采集,才能实现高效的大数据采集工作,推动大数据工作的全面发展。

参考文献

[1]殷学丽.如何安全高效的进行大数据计算机信息处理[J].中国新技术新产品,2018(15):20-21.