文章编号：2096-1472(2019)-06-47-03 DOI:10.19644/j.cnki.issn2096-1472.2019.06.012

网络环境下的大数据采集和处理

耿贞伟，保 富
(云南电网有限责任公司信息中心，云南 昆明 650041)

摘 要：在信息媒体技术高速发展的今天，网络信息日新月异，大数据正在世界掀起一股狂热浪潮，成为人类社会的一大生产力。网络环境下的大数据技术不仅仅只是对信息的收集，同时也是对信息的分析和处理，整合信息背后所隐藏的价值。大数据技术将会为人类社会带来新的改革方向，促进人类社会的文明发展。文章主要通过探索网络环境下大数据发展的规律，以及网络环境下大数据存在的特点，针对大数据资源的有效应用途径，进一步研究其数据的采集及处理技术。

关键词：网络环境；大数据；采集和处理
中图分类号：TP399 **文献标识码**：A

Research on the Big Data Acquisition and Processing in Network Environment

GENG Zhenwei,BAO Fu
(*Information Center of Yunnan Power Grid Co.,Ltd.,Kunming* 650041,*China*)

Abstract: With the rapid development of information and media technology,network information is changing with each passing day.Big data is setting off a craze wave in the world and becoming a major productive force for human society. Big data technology in the network environment is not only used to collect information,but also analyze and process information,so as to integrate the potential value hidden in the information.Big data technology will bring new direction of reform for human society and promote the development of human civilization.This paper mainly explores the development laws and characteristics of big data in the network environment,and further studies the data acquisition and processing technology for the effective application of big data resources.

Keywords:network environment;big data;acquisition and processing

1 引言(Introduction)

大数据主要指超过数据系统收集和处理能力的数据，是网络环境中常见的数据信息。大数据对信息规模和传送能力的要求都比较高，由于与原数据系统不相匹配，大数据信息并不会得到系统很好的整理及处理。大数据对于当今时代的发展具有重要作用，能促进新生产力的发展。所以，在信息更迭速度加快的信息时代，为了获取大数据所隐藏的信息价值及促进生产力发展，相关技术人员应有所行动，采取行之有效地方法对之进行采集和处理。

2 网络环境下大数据的发展(Development of big data in network environment)

最早给出大数据时代已经到来定义的是咨询公司麦肯锡。该公司在社会研究调查报告中指出，大数据信息已经融入到了社会各行各业中，成为社会各领域中重要的生产力。而随着人们对海量数据的运用又预示着新信息与生产率增长时代的到来。大数据的发展是人类社会进步发展的标志，它预示着人类社会将朝向更文明、科技化、信息化方向进步发展。在现代社会中，大数据已经成为计算机网络媒体技术中的热门内容，引起了电子计算机行业的研究热潮，成为相关企业发展的主要动力因素。目前大数据在我国的发展和研究都还处于初级阶段，但大数据的市场价值已经显现出来，在很长一段时间内我国都会注重对大数据的研究与应用，大数据在我国市场中拥有可观的发展未来。

大数据的未来整体态势及发展趋势主要表现在几个方面：大数据与人类社会活动；大数据与人类文化艺术；大数据隐私安全；系统处理；应用关键技术；产业影响。在大数据的整体态势中，数据的规模将会逐渐扩张，数据信息成为引用资源，数据共享将会成为推动企业发展的核心内容。在大数据发展环境下，会催生出一批批新兴产业，伴随开创一些新兴职业，如数据分析师、数据工程师、数据分析师等。这些新兴职业能为各国社会提供更多的就业岗位，一方面也为国家解决了就业的民生问题。大数据下信息共享是必不可少的，伴随着又会带来信息隐私安全的问题，信息共享带来便利的同时也蕴藏了较大的信息隐私安全问题。数据信息资源化发展，大数据在各个国家及地区都会成为竞争中最有力的战略资源，成为新的竞争优势和新发展动力。

3 网络环境下大数据的分类及特点(Classification and characteristics of big data in network environment)

3.1 大数据的分类

3.1.1 以实时性特点划分，分为实时数据分析和离线数据分析

实时数据分析应用的领域主要是B2C产品，包括金融、互联网和移动领域内的B2C产品。这类产品往往会要求在限定时间内反馈上亿次的信息，以此来达到提高用户体验的目的。实时数据分析对处理体系的要求比离线数据分析的要高，而当前对大数据采集和处理的研究也主要针对的是实时数据信息。对于大多数对反馈时间要求不高的应用，如机器学习、推荐引擎计算、离线统计分析等，可以采用离线数据分析的方式，通过采集的数据将其转移到专用的分析平台。面对海量数据的冲击，传统ETL工具往往会出现崩溃的问题。ETL在格式转换上对资金的需求较高，所以在性能上根本无法满足对海量数据的采集和处理需求。从实时数据分析与离线数据分析来看，大数据在量上远高于其他物质流量，因此需要抓紧对大数据采集和处理体系的研究。

3.1.2 以数据量分析，分为内存级别、海量级别、BI级别

大数据中的内存级别是指不超过群体的内存最大值；海量级别是指对于BI产品和数据库都已经失效或者成本过高的数据量；BI是区别于内存较小来说，具有过大特征的数据量。这类数据一般可以放在传统BI产品和量身定做的BI数据中进行分析。无论是内存级别还是海量级别或BI级别的数据来说，都是以数据量为基础进行分析，同时也都满足大数据的概念。所以，在具体应用中需要对数据类型进行分类比较，以采取适当的采集及处理方法[1]。海量级别的大数据流量是当前发展的一个主要趋势，在这里所有的数据信息都会汇聚于此。而当前技术发展的要求是建立具体的海量数据分析系统，以确保信息特征的有效分类采集和处理。

3.2 大数据的特点

大数据在社会实践过程中表现了以下几个特点：海量性、易变性、多样性、高速性。社会各行各业都面临数据量范围增长的挑战。比如，在最近一次的IDC报告预测中曾说，在2020年，世界范围内的信息量将会扩大为原来的50倍。当前，大数据的增长规模仍然存在一个不确定的指标，单一数据集的范围可以从数十TB到数PB不等。1PB的数据需要两万台50GB硬盘的电脑才能进行处理，可见大数据的增长规模是非常巨大的。此外，在这个过程中还会催生出其他数据信息，数据量也会持续不断的增加，最终形成海量的数据信息流量。大数据一般都具有多层结构，各层结构都具有不同的特征表现，因此大数据也会有不同的类型和形式。相较于传统业务数据而言，大数据在该系统下会存在特征不明显、形状不规则等问题，传统数据处理软件是很难对其进行有效处理的。传统业务数据已经能生成符合自身发展的标准个数，能够被智能软件所识别。

目前，社会发展所面临的难题就是从海量数据中发觉其市场价值。社会大众普遍都认为，使用互联网进行搜索是造成数据增长的主要原因，改观点是不正确的[2]。数据的增长主要是新型多结构数据的原因，其中包括了网络日志、互联网搜索、手机通讯录等。数据的快速增长一直是处理大数据的一大难题，在这里人们需要考虑到更多方面的因素，以免出现数据采集和处理不恰当的情况。此外，传感器的安装也增加了数据量增长、范围扩大的可能性，推动数据多样性的发展；高速主要是指被创建和被移动的速度。在网络大环境下，流行的趋势是通过基于实现软件性能优化的高速电脑处理器和服务器。企业需要掌握快速创建数据、处理信息、反馈方法的方法，以此来对数据进行采集和处理。

4 网络环境下大数据的处理和分析(Processing and analysis of big data in network environment)

要将大数据作为一种资源，就必须对大数据进行相关处理和分析。大数据的处理和分析对海量数据处理，信息价值的提取都有较高的要求。大数据具有不同的类型及特征，所以目前，针对市场针对静态数据主要还是采取批处理的方法，对实时在线数据采取流处理。批处理方法首先需要将原始数据进行分类处理，以得出数据的具体含义，然后分配给不同的任务区进行处理，最终形成有效的处理措施。把计算推到数据是大数据处理的有效办法，对文本中构建良好的模型具有重要影响。于大数据的静态数据而言，比较适用于先储存后计算的处理系统。Dremel交互式数据分析系统是Google公司推出了一款实时数据处理系统，这款系统一经问世便引领电子计算机行业向实时数据处理迈进。实时数据的处理大致可分为流式处理和交互式处理两种模式，交互式的处理在具有灵活、直观、操作便捷的特点，处理的结果可以直接得到运用。交互式数据处理比较常用于服务器实时日志数据的采取上，交互式处理系统可以适当将网络上的PB级数处理时间缩短到最小，尽可能地对数据进行处理和分析并给出结果[3]。交互式处理是常见的大数据处理方式，在应用该技术的过程中需要更加注意对数据PB级数的处理，以提高大数据的处理效率。

电子设备的普及和发展，催生了海量的图像、影音数据信息，由于数据信息本身的特征能表示事物之间的相关关系，大数据融合的特征使信息范围的规模不断增长，达到了上百万或上亿个节点的大图数据库，这也对计算机的数据采集和处理技术提出了巨大的挑战。图数据需要选取适当的图分割和计算机模型来进行处理挑战，以实现对大数据的采集和处理效果。大数据在互联网领域、公共服务领域及安全防护领域内的运用以初见成效。目前据调查显示，在微博、微信中人们都会对根据文本、图片、联系方式等信息来推荐的内容倍感兴趣，提升了用户本身的体验效果。对大数据的处理是信息快速发展的要求，受大数据海量性、易变性、高速性等特点的影响，处理大数据需要更多行之有效地方法。大数据的分析系统需要依靠传感分析处理系统，对抓取的数据进行分块处理，发现每个文本信息所对应的特征，将信息归类到具体的板块。

5 网络环境下大数据的采集和筛选(Acquisition and screening of big data in network environment)

5.1 网络大数据采集的基本流程与关键技术

5.1.1 网络大数据采集的基本流程

数据采集的整体框架主要包括六大板块：网站页面、链接抽取、链接过滤、内容抽取、爬虫URL队列和数据。这些板块各自所带有的功能分别是：网站页面的功能是获取网站的网页内容；链接抽取是抽取网站页面中的链接地址；链接过滤是辨别该链接地址下的网页内容是否被抓取过；内容抽取是从网页内容中抽取需要的属性，然后提取所需内容值；URL队列是为爬虫提供需要抓取数据网站的url；数据主要包括Site Url，被抓取的数据网站中的url、Spider Url，已经抓取过的网站数据

Url和Spider Content，被抓取的网页内容[4]。

网络大数据的采集过程主要包括：将被抓取的网站Url信息写入Url Queue程序中；在URL队列中获取被抓取网站的信息；集中获取某一网站的网页内容；在网页内容中抽取链接地址；在数据库中读取被抓取网站的网页地址；过滤Url，对比抓取前后的Url信息；分类处理，将没有被抓取过的网页地址写入Spider Url数据库，已抓取过的放置对该网页地址的抓取；获取网页地址内容，抽取所需属性的内容值；将抽取的内容值写入数据库[5]。对大数据的处理必须具备条理清晰的处理流程和方法，这样才能井然有序地对海量数据信息进行处理。此外，在处理大数据时还要针对数据信息的特征来进一步来分析处理，以确保其精准度。

5.1.2 数据采集的关键技术

数据采集的关键是布隆过滤器。链接过滤技术的核心是判断当前链接是否在已被抓取过的链接里面。在进行网页大数据的采集时，可以增加布隆过滤器来完成对链接的筛选。运用布隆过滤器的想法是：当集合中加入一个新元素，可以通过X个散列函数实现对该元素的X个映射点数组，并将这些数组置为1。在检索的过程中我们只需要观察这些点是否为1就基本能判断该集合中有没有这个新元素了。若这些点数组中存在一个0，则被检测的元素一定不存在；若都是1，则被检测的元素很可能会存在于这些点数组中[6]。布隆过滤器在横向和纵向上都具有很大的优势：

(1)在简洁程度方面，布隆过滤器的简洁程度达到了最大值，该过滤器的储存空间和插入、查询时间都为常数。

(2)在相互关系上，散列的函数并没有连接关系，有利于各硬件的并行运行。

(3)在储存方面，布隆过滤器并不储存元素本身，这一点是布隆过滤器在保密性要求较高场合中的优势。布隆过滤器的运行方法是，已被抓取的url，以X个hash函数计算，得出X个值，再与一个bit数组的这X个位置的元素对应起来(元素值已被设为1)。根据上面所论述过的方法即可得出判断。

5.2 网络大数据筛选的基本流程与关键技术

5.2.1 网络大数据筛选的基本流程

网络大数据筛选的整体框架和处理过程主要包括四个板块：分词、排重、整合、数据。四个板块的功能主要是：分词，是对抓取到的网页进行切词处理；排重，是对网页内容进行多重排重检查；整合，是对不同源信息进行格式层面的整合；数据，主要指Spider Data和Dp Data两方面的数据。网络大数据处理的基本流程主要有六个步骤：将抓取的网页进行分词；将分词处理的结果写入数据库；将抓取网页进行排重检查；将排重结果写入数据库；更具分词和排重结果，对数据进行整合；将整合结果写入数据库。

5.2.2 网络大数据筛选的关键技术

网络大数据筛选的关键技术之一是排重技术，而排重技术则需要依靠Simhash算法来实现高效排重。Simhash算法是实现对海量数据进行文本排重的算法，simhash算法在一定程度上避免了文本比较的复杂方式，于余弦角、Jaccard相似系数而言，效率得到了有效提高。运用Simhash算法对抓取的网页内容进行排重，可以有效扩大数据容量，提供更快速的数据处理速度。运用Simhash算法的想法是：输入一个为K的维向量V，如文本的特征向量，但需要保证每个特征都具有一定的权重。输出一个N位的二进制签名X[7]。

(1)初始化一个N维向量Q为0，N位的二进制签名S为0。

(2)对向量K中的所有特征使用传统的Hash算法算出一个N位的散列值H。判断式子为1≤i≤N，如果H中的第i位数是1则Q的第i个元素减去该特征的权重。

(3)如果Q的第i位元素大于0，则S的第i位数为1反之则为0。

(4)返回签名S。在Simhash算法得出每篇文档的签名之后，需要再计算两个签名间的海明距离。根据实践经验，64位的Simhash海明距离在3以内可以认为相似度较高。

数据筛选的又一关键技术便是整合技术，整合就是把抓取来的网页同各企业间建立连接关系。对于企业来说，可以使用一组关键词对企业进行描述，同样，经过处理的网页内容也可以使用一组关键词来进行描述。所以，整合就成了两组关键词之间的信息整合。网页内容的分词结果通常表现出两个特点：分词结果数量大；大部分分词对网页内容来说没有价值。所以，需要对网页内容的分词进行简化处理，使用常频词汇来描述网页内容[8]。简化之后，关键词信息的整合效率得到了很大的提升，精确度也得到了保障。在关键词的信息整合之后，抓取来的网页内容就与企业之间建立了联系，就能知道某一企业的相关数据信息了。

6 结论(Conclusion)

网络环境下大数据的信息类型是丰富多样的，需要采集和处理的数据量庞大，对信息数据展现的要求较高，并且会着重关注数据的高效性和可用性。大数据是促进社会生产力新发展的关键因素，是引起人类社会变革的重要条件。所以，在网络环境中要抓紧对大数据采集和处理技术的研究，以提高大数据的高效性和可用性，促进社会生产力的发展。

参考文献(References)

[1] 丁俊,郑辉.大数据时代下的动态可配置数据采集系统的研究与设计[J].计算机应用与软件,2018(3):75−79.
[2] 陈廉芳.大数据环境下图书馆用户小数据的采集、分析与应用[J].国家图书馆学刊,2016,25(3):069−074.
[3] 亓峰,唐晓璇,邢宁哲,等.未来大数据环境下的配用电通信网虚拟网络架构及应用[J].数据采集与处理,2015(3):511−518.
[4] 鞠娜.移动互联网的大数据处理关键技术[J].信息与电脑(理论版),2015(23):38.
[5] 张继栋,许鹏.大数据采集系统的设计与应用[J].大观,2017(11):40.
[6] 李明.大数据环境下无线终端智能通信信号采集系统设计[J].电子设计工程,2017,25(16):94−98.
[7] 林文辉.基于Hadoop的海量网络数据处理平台的关键技术研究[D].北京邮电大学,2014:29−36.
[8] 杨刚,杨凯.大数据关键处理技术综述[J].计算机与数字工程,2016,44(4):694−699.

作者简介：

耿贞伟(1973−)，男，硕士，工程师.研究领域：云计算信息技术.
保　富(1993−)，男，硕士，助理工程师.研究领域：云计算信息技术.