



大数据

智慧计算原理方法

朱定局 著



科学出版社

大数据智慧计算原理方法

朱定局 著

国家科技支撑计划课题(No. 2015BAH42F01)

国家社会科学基金重大项目(No. 14ZDB101)

国家自然科学基金(No. 61105133)

科学出版社

内 容 简 介

本书的内容均为原创成果。其原创性在于:提出并研究给出了大数据智慧计算原理与方法,具体又提出并研究给出了跳板大数据智慧计算原理与方法、耦合大数据智慧计算原理与方法、先验大数据智慧计算原理与方法、自适应大数据智慧计算原理与方法、增量大数据智慧计算原理与方法、自动大数据智慧计算原理与方法、分治大数据智慧计算原理与方法、冗余大数据智慧计算原理与方法。

本书可作为大学和院所相关专业的教学用书和研究用书,同时还可供政府部门和企事业单位参考。

图书在版编目(CIP)数据

大数据智慧计算原理方法/朱定局著. —北京:科学出版社,2015.2
ISBN 978-7-03-042863-9

I. ①大… II. ①朱… III. ①计算机网络-数据处理 IV. ①TP393

中国版本图书馆 CIP 数据核字(2014)第 305090 号

责任编辑:杨帅英 唐保军 / 责任校对:张小霞

责任印制:张 倩 / 封面设计:北京图阅盛世文化传媒有限公司

科学出版社出版

北京东黄城根北街16号

邮政编码:100717

<http://www.sciencep.com>

三河市骏杰印刷有限公司印刷

科学出版社发行 各地新华书店经销

*

2015年2月第 一 版 开本:787×1092 1/16

2015年2月第一次印刷 印张:12

字数:270 000

定价:68.00 元

(如有印装质量问题,我社负责调换)

前 言

人类发明了算盘、计算器、计算机的目的是将人类从脑力劳动中解放出来,所以计算机也叫电脑,表明人类希望计算机和人脑一样具备智慧。当今随着计算机芯片技术、超级计算技术、云计算技术、网络技术的日益发展,计算机的计算能力、存储能力、网络传输能力发展非常迅速,甚至超过了人脑的计算、存储和协同能力,但即使这样,计算机能代替人来处理日常业务吗?显然不能,如果能的话,那么我们就不用工作了,那么这本书的作者就有可能是个计算机了。目前为止,计算机还只是一个工具。例如,本书的写作都是在计算机上完成的。如果没有计算机,纸上写这本书,可能会花更多的时间,因为在计算机上修改、编辑、统计字数都变得非常方便。人类已经将诸如文字编辑、科学计算、社会计算等人类处理事情的方法融入了计算机系统中。计算机硬件是计算机的身体,计算机软件是计算机的灵魂。我在上大学的时候,计算机系统连我这个大学生学起来、用起来,都非常费劲。可现在,小孩子都能很快学会并熟练地在计算机上玩游戏、聊天。这是为什么?这是因为计算机系统的软件中融入了越来越多的人类智慧,计算机系统越来越智慧,进而使得计算机系统越来越容易被使用。我上大学时的计算机如果只有1岁孩子的智商,那么现在的计算机估计已经有5岁孩子的智商,自然就更容易打交道了。大数据的兴起给进一步提升计算机的智力提供了一个机会,本书的目的就是在大数据处理中融入更多的智慧,使得计算机系统的智商进一步提高。

大数据的大不但表现在量大,更表现在包罗万象上,就是会包含不同对象,但在现实中这些表面上不怎么相关的数据的数据源也是分离的,所以需要一种方法能把这些数据源关联起来。第2章正是针对处理海量大数据的需要,提出和研究了跳板大数据智慧计算原理与方法,以关联不同的数据源,使得不同的数据源之间可以互联互通,从而可以产生新的服务和新的数据。正是利用了跳板大数据智慧计算原理与方法,才使得云计算的调度与绿色能源的调度连接了起来(2.1节),从而提高绿色电力的利用率;才使得虚拟建模与物联网连接了起来(2.2节),从而通过虚拟模型可以操纵现实;才使得移动终端与超级计算机连接了起来(2.3节),从而通过移动终端可以方便使用超级计算机。

大数据的结构非常复杂,体现在数据内部和数据之间都存在着异构性。传统处理大数据的框架如hadoop,比较擅长于处理文本大数据,但在处理异构大数据时就有些力不从心,因为异构大数据的数据内部、数据之间的关系非常复杂,所以急需研究专门针对异构大数据的更为智慧的计算方法,来针对异构大数据的结构复杂性进行高效处理,得到预期的结果和效果。第3章正是针对处理异构大数据的需要,提出和研究了耦合大数据智慧计算原理与方法,使得异构大数据之间可以进行有效耦合。正是利用了耦合大数据智慧计算原理与方法,才使得分布式供电节点与分布式用电节点得到了耦合,从而提高了电网效率(3.1节);才使得不同的云系统间得到了耦合,从而进一步发挥云的优越性(3.2节);才使得结构化数据与非结构化数据库在云中得到了耦合,从而既易于数据查询又易

于数据分合(3.3节)。

大数据的价值在于其中所蕴含的知识,而大数据中的知识只有依靠智慧计算才能充分地发现和利用。因为大数据不同于传统的数据,大数据是未经采样和加工的全数据,因此其数据质量远远低于传统数据,而数据复杂性远远高于传统数据,这就对处理数据的计算方法提出了更高的要求。第4章正是针对处理全大数据的需要,提出和研究了先验大数据智慧计算原理与方法,使得先验结果可以降解全大数据处理的难度。正是利用了先验大数据智慧计算原理与方法,才使得事后台的仿真结果可以用于实时突发事件的仿真(4.1节);才使得对各作者的文学作品的统计可以用于鉴别文学作品的作者(4.2节)。

大数据中的知识是隐蔽的,需要利用一定的算法进行处理,才能够挖掘出其中隐藏的知识,这些知识往往能够帮助信息系统更好地适应环境因素和用户需求。第5章正是针对挖掘大数据隐蔽性知识的需要,提出和研究了自适应大数据智慧计算原理与方法,通过大数据感知环境因素和用户需求,从而更好地为用户提供贴心的服务。正是利用了自适应大数据智慧计算原理与方法,才使得云计算系统可以适应不同的网络环境、服务端环境、客户端环境,来调用不同的模块,从而使得云计算系统可用性更高(5.1节);才使得超级计算机可以根据任务对节点的具体需求,将任务调度到相应计算能力的节点,从而使得超级计算效率提高(5.2节);才使得广告可以根据网页内容进行插入,提高网页用户对广告的兴趣(5.3节)。

大数据是动态发展的,有的大数据与日俱增,有的大数据与时俱增。如果大数据增加之后,重新全部处理一遍,那我们注定迟早会被淹没在无限增长的大数据的海洋之中而无法自拔。第6章正是针对处理动态大数据的需要,提出和研究了增量大数据智慧计算原理与方法,来充分利用已有大数据的处理结果。正是利用了增量大数据智慧计算原理与方法,才使得数字城市的更新无法从头再来,减少了数字城市更新的成本(6.1节);才使得知识库能与时俱进,逐渐扩展知识、提高知识的准确度(6.2节);才使得进行更细粒度的比对时,无需重复比对粗粒度中已经匹配成功的视频段,从而减少了对比的工作量(6.3节)。

大数据最明显的特点就是数据量大,数据量越大,则所需的处理时间越长,所以加快数据处理的速度就显得非常重要。第7章正是针对处理海量大数据的需要,提出和研究了自动大数据智慧计算原理与方法,来消除数据处理过程中的人工干预。因为人工处理会大大降低数据处理的速度,所以数据处理的自动化是人类一直追求的目标,使得人类可以从脑力劳动中解放出来。正是利用了自动大数据智慧计算原理与方法,才使得数字城市可以从遥感影像中自动重建出来,而无需手工处理(7.1节);才使得多媒体可以自动地被合适地切分,而无需人工干预(7.2节);才使得某些机器人加入或离开巡逻队伍,巡逻队伍能够自动得到重新调配,而无法人为调整(7.3节)。但只是自动还不够,因为即使自动了,海量的大数据还是难以及时地处理出结果,所以我们需要采用分治的方法,分而治之,化大为小来加速大数据的处理速度,使得大数据可以在有效的时间内处理出结果。第8章正是针对处理海量大数据的需要,提出和研究了分治大数据智慧计算原理与方法,来充分利用并行计算和云计算的优势来加速大数据的处理。分治,就是分而治之,从而大事化小。正是利用了分治大数据智慧计算原理与方法,才使得视频可以分为很多视频段同时转码,从而加快转码的速度(8.1节);才使得多机器人的任务可以分发给很多云节点分别同时地处理,从而提高多机器人的处理能力(8.2节);才使得密码可以隐藏在各云数据

分块的分布中,从而提高云安全性(8.3节)。除了分治方法可以加速海量大数据的处理,第9章针对处理海量大数据的需要,提出和研究了冗余大数据智慧计算原理与方法,以空间换时间,来进一步加速海量大数据的处理速度,其中把程序也当作一种数据。正是利用了冗余大数据智慧计算原理与方法,才使得损失了微小的重叠边界存储,换来了大幅度的并行处理时网络通信量的降低,从而可以大幅度地提高并行处理速度(9.1节);才使得损失了不同版本同时存在的系统开销,换来了用户体验的大幅度提高(9.2节);才使得损失了各周期结果数据存储开销,换来了更高级别周期数据处理速度的大幅度提高(9.3节)。

本书的主要创新如下:

(1) 首次提出并研究给出了跳板大数据智慧计算原理与方法,并通过原创性例子进行说明,例子包括通过云计算调配绿色能源的原理与方法、虚拟建模物联云的原理与方法、通过移动终端访问超级计算机的原理与方法。

(2) 首次提出并研究给出了耦合大数据智慧计算原理与方法,并通过原创性例子进行说明,例子包括智能电网分布式耦合调度的原理与方法、多云服务调度的原理与方法、结构化与非结构化数据库融合的原理与方法。

(3) 首次提出并研究给出了先验大数据智慧计算原理与方法,并通过原创性例子进行说明,并通过具体例子进行说明,例子包括仿真知识库下实时仿真的原理与方法、文学作品作者自动鉴别的原理与方法。

(4) 首次提出并研究给出了自适应大数据智慧计算原理与方法,并通过原创性例子进行说明,例子包括自适应云计算的原理与方法、超级计算机自适应调度的原理与方法、网页广告自适应插入的原理与方法。

(5) 首次提出并研究给出了增量大数据智慧计算原理与方法,并通过原创性例子进行说明,例子包括数字城市增量更新的原理与方法、知识库增量式更新的原理与方法、多粒度视频比对的原理与方法。

(6) 首次提出并研究给出了自动大数据智慧计算原理与方法,并通过原创性例子进行说明,例子包括数字城市全自动生成的原理与方法、多媒体自动切分并行的原理与方法、可扩展多机器人巡逻的原理与方法。

(7) 首次提出并研究给出了分治大数据智慧计算原理与方法,并通过原创性例子进行说明,例子包括视频分片并行转码的原理与方法、多机器人云系统的原理与方法、以云数据分布特征为密码的原理与方法。

(8) 首次提出并研究给出了冗余大数据智慧计算原理与方法,并通过原创性例子进行说明,例子包括重叠边界并行处理的原理与方法、云服务无缝升级的原理与方法、视频点播数据多级云处理的原理与方法。

由于作者水平有限,书中难免有不妥甚至错误之处,恳请读者批评指正。

朱定局

2015年1月2日于华南师范大学

目 录

前言

第 1 章 大数据智慧计算原理与方法	1
1.1 大数据的特性	1
1.2 大数据对智慧计算的需求	2
1.3 大数据智慧计算原理方法	5
第 2 章 跳板大数据智慧计算原理与方法	9
2.1 绿色能源	9
2.1.1 现有绿色能源技术的不足	9
2.1.2 通过云计算调配绿色能源的原理	10
2.1.3 通过云计算调配绿色能源的方法	11
2.2 虚拟建模	17
2.2.1 现有虚拟建模技术的不足	17
2.2.2 虚拟建模物联云的原理	17
2.2.3 虚拟建模物联云的方法	19
2.3 超级计算机访问	22
2.3.1 现有超级计算机访问技术的不足	22
2.3.2 通过移动终端访问超级计算机的原理	22
2.3.3 通过移动终端访问超级计算机的方法	23
第 3 章 耦合大数据智慧计算原理与方法	30
3.1 智能电网的调度	30
3.1.1 现有智能电网调度技术的不足	30
3.1.2 智能电网分布式耦合调度的原理	30
3.1.3 智能电网分布式耦合调度的方法	32
3.2 云计算服务的调度	41
3.2.1 现有云服务调度技术的不足	41
3.2.2 多云服务调度的原理	42
3.2.3 多云服务调度的方法	43
3.3 结构化与非结构化数据库	50
3.3.1 现有数据库技术的不足	50
3.3.2 结构化与非结构化数据库融合的原理	50
3.3.3 结构化与非结构化数据库融合的方法	52
第 4 章 先验大数据智慧计算原理与方法	56
4.1 实时仿真	56

4.1.1	现有实时仿真技术的不足	56
4.1.2	仿真知识库下实时仿真的原理	60
4.1.3	仿真知识库下实时仿真的方法	60
4.2	文学作品作者鉴别	69
4.2.1	现有作者鉴别技术的不足	69
4.2.2	文学作品作者自动鉴别的原理	69
4.2.3	文学作品作者自动鉴别的方法	71
第5章	自适应大数据智慧计算原理与方法	80
5.1	云计算	80
5.1.1	现有云计算技术的不足	80
5.1.2	自适应云计算的原理	80
5.1.3	自适应云计算的方法	81
5.2	超级计算机的调度	83
5.2.1	现有超级计算机调度技术的不足	84
5.2.2	超级计算机自适应调度的原理	84
5.2.3	超级计算机自适应调度的方法	85
5.3	网页广告的插入	91
5.3.1	现有网页广告插入技术的不足	91
5.3.2	网页广告自适应插入的原理	91
5.3.3	网页广告自适应插入的方法	92
第6章	增量大数据智慧计算原理与方法	96
6.1	数字城市的更新	96
6.1.1	现有数字城市更新技术的不足	96
6.1.2	数字城市增量更新的原理	98
6.1.3	数字城市增量更新的方法	100
6.2	知识库的更新	107
6.2.1	现有知识库更新技术的不足	107
6.2.2	知识库增量式更新的原理	108
6.2.3	知识库增量式更新的方法	109
6.3	视频比对	113
6.3.1	现有视频比对技术的不足	113
6.3.2	多粒度视频比对的原理	113
6.3.3	多粒度视频比对的方法	115
第7章	自动大数据智慧计算原理与方法	121
7.1	数字城市的生成	121
7.1.1	现有数字城市生成技术的不足	121
7.1.2	数字城市全自动生成的原理	122
7.1.3	数字城市全自动生成的方法	123

7.2 多媒体的并行处理	128
7.2.1 现有多媒体并行技术的不足	128
7.2.2 多媒体自动切分并行的原理	128
7.2.3 多媒体自动切分并行的方法	131
7.3 多机器人巡逻	137
7.3.1 现有多机器人巡逻技术的不足	137
7.3.2 可扩展多机器人巡逻的原理	137
7.3.3 可扩展多机器人巡逻的方法	139
第8章 分治大数据智慧计算原理与方法	145
8.1 视频转码	145
8.1.1 现有视频转码技术的不足	145
8.1.2 视频分片并行转码的原理	145
8.1.3 视频分片并行转码的方法	146
8.2 多机器人系统	148
8.2.1 现有多机器人系统技术的不足	148
8.2.2 多机器人云系统的原理	149
8.2.3 多机器人云系统的方法	151
8.3 云安全	157
8.3.1 现有云安全技术的不足	157
8.3.2 以云数据分布特征为密码的原理	158
8.3.3 以云数据分布特征为密码的方法	159
第9章 冗余大数据智慧计算原理与方法	166
9.1 并行处理	166
9.1.1 现有并行处理技术的不足	166
9.1.2 重叠边界并行处理的原理	166
9.1.3 重叠边界并行处理的方法	167
9.2 云服务的升级	170
9.2.1 现有云服务升级技术的不足	170
9.2.2 云服务无缝升级的原理	170
9.2.3 云服务无缝升级的方法	171
9.3 视频点播数据处理	173
9.3.1 现有视频点播数据处理技术的不足	173
9.3.2 视频点播数据多级云处理的原理	174
9.3.3 视频点播数据多级云处理的方法	175
参考文献	179
后记	180

第 1 章 大数据智慧计算原理与方法

1.1 大数据的特性

大数据有六大特性:数据量大(数量)、包罗万象(对象类型)、结构复杂(格式类型)、全数据(真实性)、知识隐蔽(隐蔽性)和动态发展(时态),如图 1.1 所示。其中,数据量大主要是从数量的角度来对大数据进行观察的结果,其中数据的数量非常庞大,体现在需要大量的存储空间,需要大量的计算资源对其中数据进行处理;包罗万象主要是从对象类型的角度来对大数据进行观察的结果,其中对象的类型非常多,在传统数据中不相关的对象类型也有可能被包罗进同一个大数据;结构复杂主要是从格式类型的角度来对大数据进行观察的结果,大数据中可以包含有大量异构的格式,这在传统数据中也是罕见的;全数据主要是从数据真实性的角度来对大数据进行观察的结果,全息数据、高维数据、中间数据等大大提高了数据的真实性;知识隐蔽是指无法通过普通的搜索、统计等手段获得其中知识。动态发展主要是从时态的角度来对大数据进行观察的结果,大数据和传统数据一样都会随时间变化,但由于大数据的量,其增量也大,使得这随时间变化导致的数据量的剧增更不能忽视,而是要专门研究一种解决的方法。

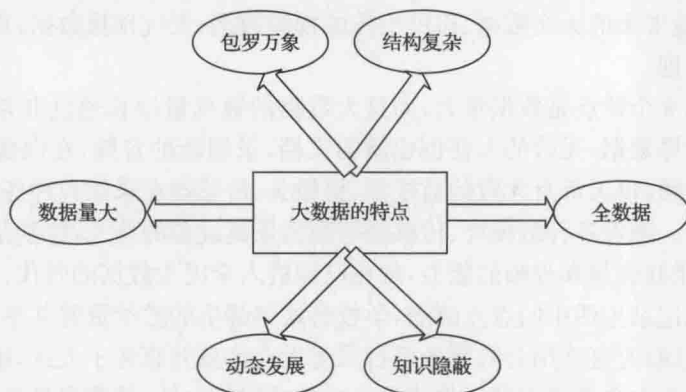


图 1.1 大数据的特点

大数据的第 1 个特点是包罗万象。在传统数据时期,看起来毫不相关的数据,在大数据时代也会统一起来考虑,所以就导致很多数据原本看起来风马牛不相及,但现在我们要将这些数据关联起来考虑。但在现实中这些表面上不怎么相关的数据的数据源也是分离的,所以需要一种方法能把这些数据源关联起来。一旦这些原本分离的数据源被关联起来,将会产生新的数据,从而又能进一步促进大数据的发展。

大数据的第 2 个特点是结构复杂。大数据相对于特定类型的数据,如数字、文本、图

像、声音、视频这些单一的数据类型而言是结构最为复杂的一种数据。大数据是一种集视频、图像、声音、文字、数字于一体的大数据,也是一种集时间维、空间维、本性维于一体的大数据。正是因为其组成成分的复杂性,以及不同维之间的关联性,使得其大数据的特性更为突出。不但大数据内部的结构复杂,而且由于大数据拆分、加工、重组、挖掘等技术的应用,使得不同大数据之间也存在着千丝万缕的联系,形成了大数据之间结构的复杂性。

大数据的第3个特点是全数据。大数据是对现实世界的记录和复制,不丢弃貌似无关紧要的信息,而传统数据是人类或程序对客观世界的记录,其中有人类对客观世界的抽象及取舍。全数据在数据的维度上没有任何损失,所以大数据的挖掘价值更大。当然,这也是相对的,因为大数据的分辨率是有限的,而且大部分大数据也不是全息,因此大数据也无法完全地复制现实,所以说大数据只是相对的全数据,但随着采集技术、存储技术的发展,大数据的分辨率会越来越高,而且会朝着全维、全息的方向发展,所以大数据的全数据性会越来越高。

大数据的第4个特点是知识隐蔽。大数据之所以被学术界和企业界甚至政府所重视,就是因为其中可以挖掘出大量的知识,但这些知识不是显而易见的,而是隐蔽的,需要采用专门的算法才可以分析出来。而且不存在一种放之四海而皆准的万能算法可以分析挖掘出所有大数据中蕴含的知识,必须要有针对性地在特定场景下对特定大数据进行特定的分析和挖掘。

大数据的第5个特点是动态发展。大数据不是死的数据,不是一成不变的数据,不是一旦处理完毕就一劳永逸的数据。因为我们的世界是动态的,整个世界最原始的状态就是一个连续不断的、将会延续无数亿年的大数据流。大数据在日新月异,甚至每秒都在剧增,如天文望远镜采集的天文数据、市民上传的视频数据、天气预报数据,都是在不断增加和发展中的大数据。

大数据的第6个特点是数据量大,而且大数据的数据量增长速度非常快。每天有无数的计算机在计算数据,无数的人在创建新的文档,录制新的音频,在向优酷等在线视频网站上传各种视频,每天都有无数的监控器、摄像头、传感器在采集各种各样的实时的、非实时的数据……。随着各种触摸屏、传感器等数据采集设备的普及,特别是智能手机也具备了输入信息、录制音频和视频的能力,现在已经进入全民大数据的时代。人们喜欢用文字、音频、视频来记录生活中的点点滴滴,学校喜欢将师生的教学资源共享,用视频来记录老师上课的实况,政府喜欢用公共服务平台采集大众信息并服务于大众,用视频来监控社会的动态。一个文本文件或图像文件或声音文件或视频文件,其数据量有大有小,小的有兆级,大的有吉级,超大的有太级,如高清视频就能达到太级。这些大数据含有非常丰富的信息,要占用大量的存储空间。同时,随着互联网特别是移动互联网的发展,越来越多的数据被上传到网上进行分享、拆分、加工、重组,从而使得从无数的原始数据中又衍生出更多数据量更大的目标数据,这又进一步增加了大数据的数据量。

1.2 大数据对智慧计算的需求

由于大数据具有6大特性,即数据量大(数量)、包罗万象(对象类型)、结构复杂(格式类

型)、全数据(真实性)、知识隐蔽(隐蔽性)和动态发展(时态),这6大特性是传统数据所不具备或不明显具备的,但现有计算原理和方法一般都是针对传统数据来进行研究的,所以用现有计算原理方法处理大数据时就会显示出局限性,这就对大数据计算原理方法产生了创新的需求,本书提出并研究了能满足这种需求的8种大数据智慧计算原理方法,如图1.2所示。

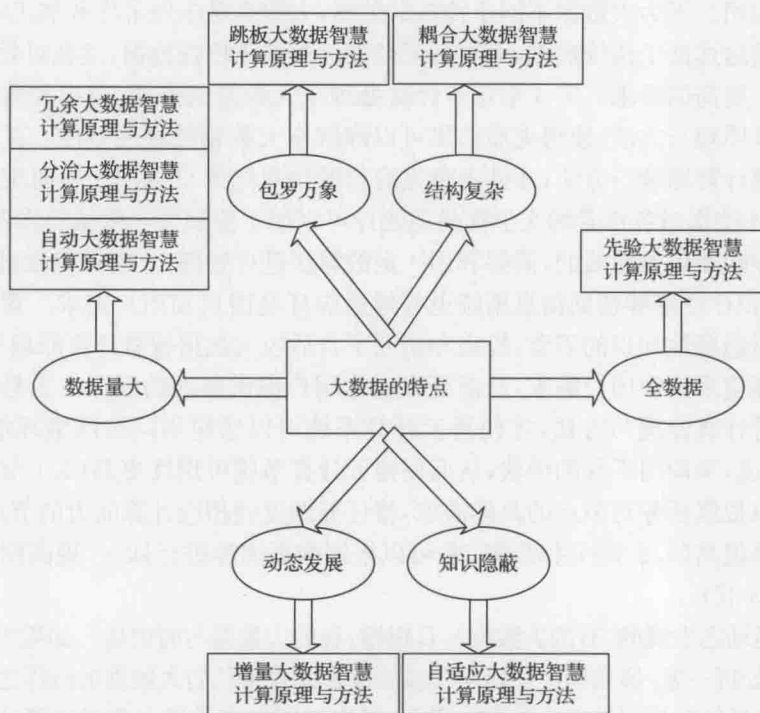


图 1.2 大数据对智慧计算的需求

大数据的大不但表现在量大,更表现在包罗万象,就是会包含不同对象,但在现实中这些表面上不怎么相关的数据的数据源也是分离的,所以需要一种方法能把这些数据源关联起来。第2章正是针对处理海量大数据的需要,提出和研究了跳板大数据智慧计算原理与方法,以关联不同的数据源,使得不同的数据源之间可以互联互通,从而可以产生新的服务和新的数据。正是利用了跳板大数据智慧计算原理与方法,才使得云计算的调度与绿色能源的调度连接了起来(2.1节),从而提高绿色电力的利用率;才使得虚拟建模与物联网连接了起来(2.2节),从而通过虚拟模型可以操纵现实;才使得移动终端与超级计算机连接了起来(2.3节),从而通过移动终端可以方便使用超级计算机。

大数据的结构非常复杂,体现在数据内部和数据之间都存在着异构性。传统处理大数据的框架如hadoop,比较擅长于处理文本大数据,但在处理异构大数据时就有些力不从心,因为异构大数据的数据内部、数据之间的关系非常复杂,所以急需研究专门针对异构大数据的更为智慧的计算方法,来针对异构大数据的结构复杂性进行高效的处理,得到预期的结果和效果。第3章正是针对处理异构大数据的需要,提出和研究了耦合大数据智慧计算原理与方法,使得异构大数据之间可以进行有效耦合。正是利用了耦合大数据

智慧计算原理与方法,才使得分布式供电节点与分布式用电节点得到了耦合,从而提高了电网效率(3.1节);才使得不同的云系统间得到了耦合,从而进一步发挥云的优越性(3.2节);才使得结构化数据与非结构化数据库在云中得到了耦合,从而既易于数据查询又易于数据分合(3.3节)。

大数据的价值在于其中所蕴含的知识,而大数据中的知识只有依靠智慧计算才能被充分发现和利用。因为大数据不同于传统的数据,大数据是未经采样和加工的全数据,因此其数据质量远远低于传统数据,而数据复杂性远远高于传统数据,这就对处理数据的计算方法提出了更高的要求。第4章正是针对处理全大数据的需要,提出和研究了先验大数据智慧计算原理与方法,使得先验结果可以降解全大数据处理的难度。正是利用了先验大数据智慧计算原理与方法,才使得事先后台的仿真结果可以用于实时突发事件的仿真(4.1节);才使得对各作者的文学作品的统计可以用于鉴别文学作品的作者(4.2节)。

大数据中的知识是隐蔽的,需要利用一定的算法进行处理,才能够挖掘出其中隐藏的知识,这些知识往往能够帮助信息系统更好地适应环境因素和用户需求。第5章正是针对挖掘大数据隐蔽性知识的需要,提出和研究了自适应大数据智慧计算原理与方法,通过大数据感知环境因素和用户需求,从而更好地为用户提供贴心的服务。正是利用了自适应大数据智慧计算原理与方法,才使得云计算系统可以适应不同的网络环境、服务端环境、客户端环境,来调用不同的模块,从而使得云计算系统可用性更高(5.1节);才使得超级计算机可以根据任务对节点的具体需求,将任务调度到相应计算能力的节点,从而使得超级计算效率提高(5.2节);才使得广告可以根据网页内容进行插入,提高网页用户对广告的兴趣(5.3节)。

大数据是动态发展的,有的大数据与日俱增,有的大数据与时俱增。如果大数据增加之后,重新全部处理一遍,那我们注定迟早会被淹没在无限增长的大数据的海洋之中而无法自拔。第6章正是针对处理动态大数据的需要,提出和研究了增量大数据智慧计算原理与方法,来充分利用已有大数据的处理结果。正是利用了增量大数据智慧计算原理与方法,才使得数字城市的更新无需从头再来,减少了数字城市更新的成本(6.1节);才使得知识库能与时俱进,逐渐扩展知识、提高知识的准确度(6.2节);才使得进行更细粒度的比对时,无需重复比对粗粒度中已经匹配成功的视频段,从而减少了对比的工作量(6.3节)。

大数据最明显的特点就是数据量大,数据量越大,则所需的处理时间越长,所以加快数据处理的速度就显得非常重要。第7章正是针对处理海量大数据的需要,提出和研究了自动大数据智慧计算原理与方法,来消除数据处理过程中的人工干预。因为人工处理会大大降低数据处理的速度,所以数据处理的自动化是人类一直追求的目标,可以使人类从脑力劳动中解放出来。正是利用了自动大数据智慧计算原理与方法,才使得数字城市可以从遥感影像中自动重建出来,而无需手工处理(7.1节);才使得多媒体可以自动地被合适地切分,而无需人工干预(7.2节);才使得某些机器人加入或离开巡逻队伍,巡逻队伍能够自动得到重新调配,而无法人为调整(7.3节)。但只是自动还不够,因为即使自动了,海量的大数据还是难以及时处理出结果,所以我们需要采用分治的方法,分而治之,化大为小来加速大数据的处理速度,使得大数据可以在有效的时间内处理出结果。第8章正是针对处理海量大数据的需要,提出和研究了分治大数据智慧计算原理与方法,充分利

用并行计算和云计算的优势来加速大数据的处理。分治,就是分而治之,从而大事化小。正是利用了分治大数据智慧计算原理与方法,才使得视频可以分为很多视频段同时转码,从而加快转码的速度(8.1节);才使得多机器人的任务可以分发给很多云节点分别同时地处理,从而提高多机器人的处理能力(8.2节);才使得密码可以隐藏在各云数据分块的分布中,从而提高云安全性(8.3节)。除了分治方法可以加速海量大数据的处理,第9章针对处理海量大数据的需要,提出和研究了冗余大数据智慧计算原理与方法,以空间换时间,来进一步加速海量大数据的处理速度,其中把程序也当作一种数据。正是利用了冗余大数据智慧计算原理与方法,才使得损失了微小的重叠边界存储,换来了大幅度的并行处理时网络通信量的降低,从而可以大幅度地提高并行处理速度(9.1节);才使得损失了不同版本同时存在的系统开销,换来了用户体验的大幅度提高(9.2节);才使得损失了各周期结果数据存储开销,换来了更高级别周期数据处理速度的大幅度提高(9.3节)。

1.3 大数据智慧计算原理方法

本书提出并研究给出了8种大数据智慧计算原理方法(图1.3):跳板大数据智慧计算原理与方法、耦合大数据智慧计算原理与方法、先验大数据智慧计算原理与方法、自适应大数据智慧计算原理与方法、增量大数据智慧计算原理与方法、自动大数据智慧计算原理与方法、分治大数据智慧计算原理与方法、冗余大数据智慧计算原理与方法。

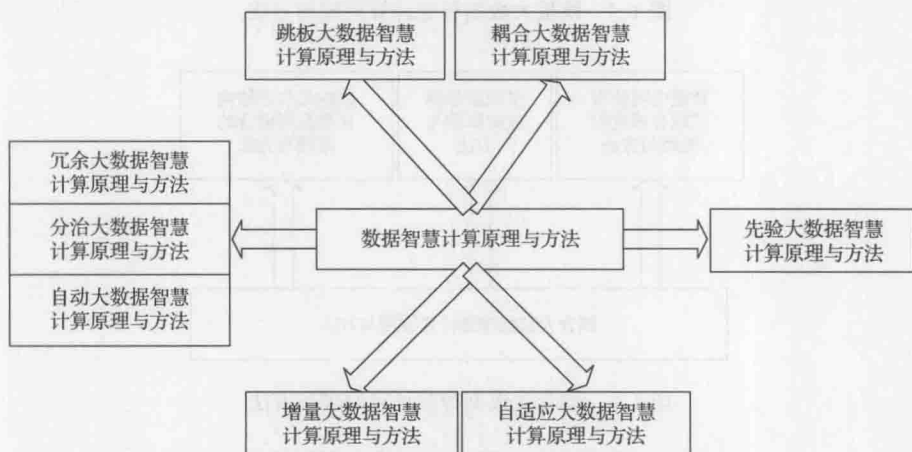


图 1.3 大数据智慧计算原理方法

每种大数据智慧计算原理与方法都通过原创性例子来进行说明。跳板大数据智慧计算原理与方法,例子包括通过云计算调配绿色能源的原理与方法、虚拟建模物联云的原理与方法、通过移动终端访问超级计算机的原理与方法,如图1.4所示。提出并研究给出了耦合大数据智慧计算原理与方法,例子包括智能电网分布式耦合调度的原理与方法、多云服务调度的原理与方法、结构化与非结构化数据库融合的原理与方法,如图1.5所示。提出并研究给出了先验大数据智慧计算原理与方法,例子包括仿真知识库下实时仿真的原理与方法、文学作品作者自动鉴别的原理与方法,如图1.6所示。提出并研究给出了自适应大数据智慧计

算原理与方法,例子包括自适应云计算的原理与方法、超级计算机自适应调度的原理与方法、网页广告自适应插入的原理与方法,如图 1.7 所示。提出并研究给出了增量大数据智慧计算原理与方法,例子包括数字城市增量更新的原理与方法、知识库增量式更新的原理与方法、多粒度视频比对的原理与方法,如图 1.8 所示。提出并研究给出了自动大数据智慧计算原理与方法,例子包括数字城市全自动生成的原理与方法、多媒体自动切分并行的原理与方法、可扩展多机器人巡逻的原理与方法,如图 1.9 所示。提出并研究给出了分治大数据智慧计算原理与方法,例子包括视频分片并行转码的原理与方法、多机器人云系统的原理与方法、以云数据分布特征为密码的原理与方法,如图 1.10 所示。提出并研究给出了冗余大数据智慧计算原理与方法,例子包括重叠边界并行处理的原理与方法、云服务无缝升级的原理与方法、视频点播数据多级云处理的原理与方法,如图 1.11 所示。

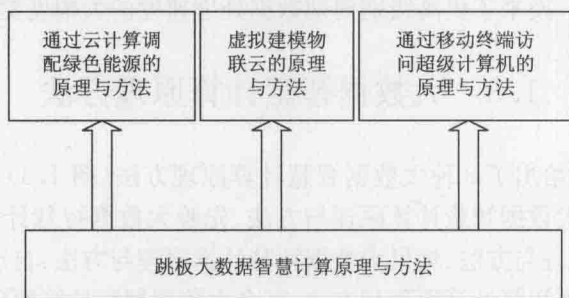


图 1.4 跳板大数据智慧计算原理与方法

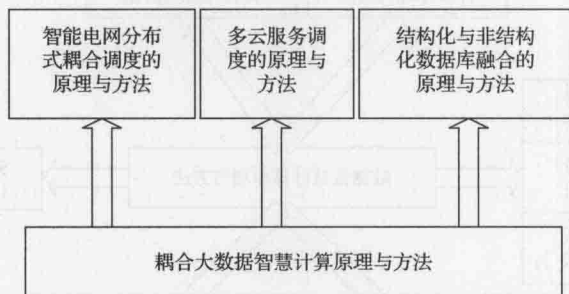


图 1.5 耦合大数据智慧计算原理与方法

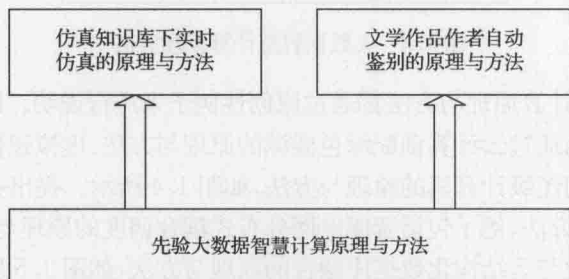


图 1.6 先验大数据智慧计算原理与方法

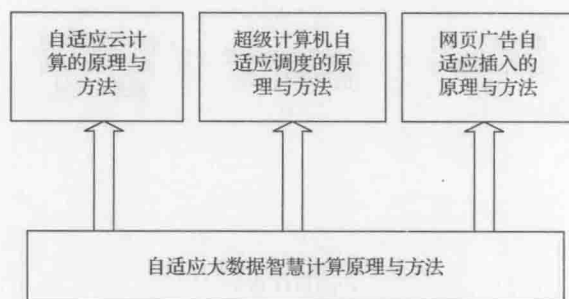


图 1.7 自适应大数据智慧计算原理与方法

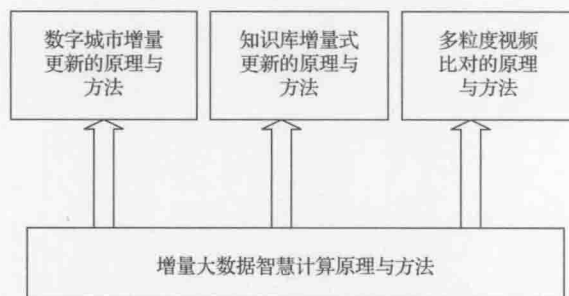


图 1.8 增量大数据智慧计算原理与方法

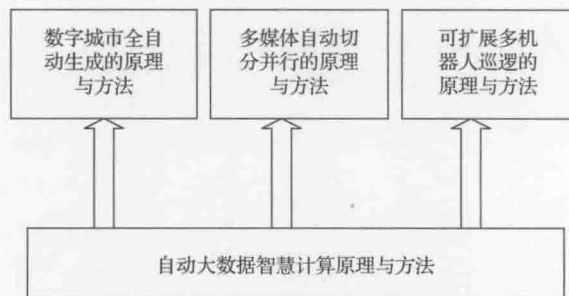


图 1.9 自动大数据智慧计算原理与方法

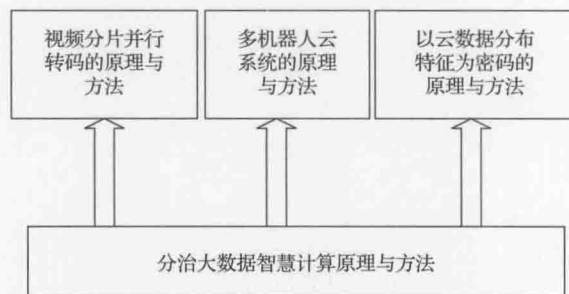


图 1.10 分治大数据智慧计算原理与方法

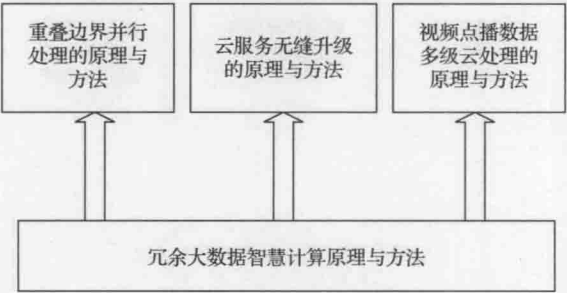


图 1.11 冗余大数据智慧计算原理与方法

