基于粒计算的大数据处理

徐 计[1),2),3)] 王国胤[3)] 于 洪[2)]

[1)](西南交通大学信息科学与技术学院 成都 610031)

[2)](重庆邮电大学计算智能重庆市重点实验室 重庆 400065)

[3)](中国科学院重庆绿色智能技术研究院 电子信息技术研究所 重庆 401122)

摘 要 在大数据时代,如何充分挖掘出蕴藏于数据资源中的价值正在成为各国 IT 业界、学术界和政府共同关注的焦点.使用云计算平台分布式地存储和分析大数据已经成为共识并且得到了广泛应用,但这并没有完全解决大数据的 3V 特性带来的问题.全面应对大数据的挑战需要来自存储技术、下一代网络、处理器、计算模型等各个领域的创新.粒计算是在求解问题过程中使用"粒"的理论、方法、技术和工具的集合,适用于近似求解有不确定性和层次结构的问题.该文综述了大数据处理的研究现状,分析了当前大数据处理研究存在的局限性;根据运用粒计算方法解决问题的不同特征,该文归纳了粒计算的 3 种基本模式,回顾了各种模式的相关研究工作;该文讨论了粒计算应用于大数据处理的可行性与优势,并探讨了在大数据的粒计算处理框架中需要解决的各个关键问题.

关键词 粒计算;大数据;云计算;深度学习

中图法分类号 TP181 **DOI 号** 10.11897/SP.J.1016.2015.01497

Review of Big Data Processing Based on Granular Computing

XU Ji[1),2),3)] WANG Guo-Yin[3)] YU Hong[2)]

[1)](*School of Information Science & Technology, Southwest Jiaotong University, Chengdu* 610031)

[2)](*Chongqing Key Laboratory of Computational Intelligence, Chongqing University of Posts and Telecommunications, Chongqing* 400065)

[3)](*Institute of Electronic Information Technology, Chongqing Institute of Green and Intelligent Technology, Chinese Academy of Sciences, Chongqing* 401122)

Abstract In big data age, people from IT industry, academia and government are taking effort to find out the value behind the data resources. Using cloud computing platform to store and process big data distributed has been a common sense and been put into practice extensively. However, this is far from solving the problems completely caused by the 3V-features (volume, variety and velocity) of big data. The problems require joint efforts of researchers from the fields of storage technology, Internet2, processors, and computing paradigm, etc. Granular computing (GrC) is a label of the family of any theories, methodologies, techniques, and tools that make use of information granules in problem solving, especially when the problem is with uncertainty, of hierarchical structure and can be approximately solved. This paper surveys the researches on big data processing and analyzes the limitation in current big data processing researches, summarizes three basic modes of GrC according to the ways in which GrC is used to solve problems and reviews the related works of each mode. The feasibility and advantages of applying GrC to process big data are discussed subsequently, and the related key problems rising in GrC-based big data processing framework are discussed at last.

Keywords granular computing; big data; cloud computing; deep learning

---

收稿日期:2014-08-22;最终修改稿收到日期:2014-11-28.本课题得到国家自然科学基金(61272060,61379114)、重庆市自然科学基金重点项目(CSTC2013jjB40003)资助.**徐 计**,男,1979 年生,博士研究生,高级工程师,中国计算机学会(CCF)学生会员,主要研究方向为数据挖掘、粒计算、软件工程.E-mail:xuji1979@126.com.**王国胤**(通信作者),男,1970 年生,博士,教授,博士生导师,中国计算机学会(CCF)高级会员,主要研究领域为粗糙集、粒计算、认知计算、智能信息处理、数据挖掘、智能信息安全.E-mail:wanggy@ieee.org.**于 洪**,女,1972 年生,博士,教授,研究领域为粗糙集、智能信息处理、Web 智能、数据挖掘等.

1 引 言

随着人类对客观世界各领域数字化程度的不断提高,每天都有大量的数据产生,并且其产生的速度也越来越快.这些数据来源广泛,其中最主要的有:科学研究(天文学、生物学、高能物理等)、社交网络、电子商务、物联网、移动通信等[1].根据IDC公司的研究,与2009年相比,全球数字信息总量将在2020年增长44倍,达到大约35.2ZB[2].

对于如此急剧增长的数据,各个国家、各个领域的人们都意识到了其中蕴含着的巨大价值;同时,信息技术领域的研究者们也感受到了由此带来的巨大挑战.当传统的计算平台、通信网络、DBMS都变得难以满足需求的时候,人们需要在存储技术、下一代网络、处理器、编程语言和计算模型等领域寻求新的解决方案.

粒计算作为一种方法论,旨在有效地建立基于外部世界、并以用户为中心的概念,进而简化我们对物理世界和虚拟世界的认识.并以此为基础,在求解问题的过程中,用粒度合适的"粒"作为处理对象,从而在保证求得满意解的前提下,提高解决问题的效率.自1979年Zadeh发表第一篇关于信息粒度的论文以来,国内外研究人员对粒计算理论和模型进行了深入的研究,同时也将这些理论和模型与其他计算智能、机器学习的技术相结合,取得了大量研究成果.

合适的粒度常常是由问题本身及问题背景决定的,这一点对设计基于粒计算的数据处理框架有重要意义.举一个关于时间的例子,例如张先生问他的朋友:"你什么时候回国的",回答这个问题所选择的时间粒度其实是由他朋友回国的时间到现在有多久决定的.如果没超过一天,那么他会说:"昨天中午";如果有十天左右了,他可以说:"上周";再如果是朋友回国好几年了,张先生才得知消息,那么"2009年"就可以是一个满意的答案了.注意到上面几个答案具有不同的粒度,分别是午、周和年.如果不采用合适的粒度,统一都用计算机上常见的时间戳格式来回答,如:"2013年4月29日下午3时25分",就不太合理,让人觉得别扭.

人工智能和计算智能等学科的诞生,是因为人们试图从人类思维和生物界的一些规律中得到启发,创建相应的计算模型,应用到信息科学中去.人工神经网络、基因计算、群体智能等都是成功的范例.粒计算则在更高层次上模拟了人类的思维规律,也可以将其运用到当前世界面临的大数据挑战中.最近,如何将粒计算应用于大数据处理得到了国际国内粒计算研究者们的重视.

2 大数据

2.1 大数据的定义

为了应对数据大规模增长带来的机遇和挑战,美国《Nature》杂志在2008年9月4日率先提出了"大数据"的概念[3].Gartner公司将大数据定义为:大数据是巨量、高速和多样性的信息资源,它需要合算地、创新地进行信息处理以增强洞察力和决策力.维基百科对大数据的定义是:大数据是这样大而复杂的数据集的汇集,以致使用当前的数据库管理工具和数据处理应用程序很难有效地处理它.另外,IBM也从数据量大、增长快速和来源多样的角度对大数据进行了描述性的定义①.

从上述的定义中,我们不难发现大数据的特征有以下3点:

(1)数据量大.虽然有人认为数据量大并不是大数据的本质,但提到大数据时,数据的大小通常都在PetaBytes到ExaBytes,至少在GigaBytes以上②.

(2)高速增长.每天都有2.5EB(2.5×10^{18} Bytes)的数据产生出来,目前世界上90%的数据都是在过去两年中产生的.除了数据量急剧增大以外,数据的处理也需要"及时"甚至"实时".

(3)结构多样.大数据中的数据,结构多种多样,如文本、声音、视频、传感器信号以及点击流数据等.多种数据交织在一起,由此给数据处理带来了困难.

以上3点就是公认的大数据3V特性(Volume, Velocity, Variety).如果说大数据的特性还包含第4个V,那么对于它的含义就有不同的解释了.微软认为是"价值"(Value),IBM认为是"真实性"(Veracity),还有学者认为是"灵活性"(Vitality).虽然这些观点不同,但都反映了对大数据处理的某种要求.最近,Wu等人[4]提出了大数据的HACE法则,认为大数据的特征是异构(heterogeneous)、自治(autonomous)、复杂(complex)和演化(evolving).

2.2 大数据带来的挑战与机遇

大数据的3V特性给信息技术的各个领域带来

① What is big data?http://www-01.ibm.com/software/data/bigdata/

② http://searchcloudcomputing.techtargetcom/definition/big-data-Big-Data

了重大挑战.不少文献对此进行了详细介绍[5-8].归纳起来,主要包括以下几个方面:

(1)存储与检索.数据量的急剧增长,给数据存储带来了巨大的挑战.虽然随着存储技术的发展,单个磁盘的容量越来越大,单位容量的价格也越来越低,然而无限制地存储所有产生或采集到的数据显然给自然资源、社会资源(生产存储设备)和电力能源形成了极大的压力.同时,数据量的增长,必然增加了检索有价值信息的难度.

(2)通信.Internet 和移动 Internet 在很大程度上催生了大数据;同时,大数据也使得在网络上传输的数据量急剧增加,对网络基础设施形成了巨大的压力.

(3)处理.大数据的价值只有经过处理才能体现.处理流程包括:数据获取→抽取/清洗→集成/表示→分析/建模→解释 5 个阶段,几乎每个阶段都要考虑下面的 5 条需求:异质性与不完备性、数据规模、及时性、隐私保护和人工协同①.

(4)共享与安全.数据量的增长和存储体系的复杂化,给信息共享和信息安全带来了更多问题.如 2012 年 3 月美国犹他州卫生署的数据泄露事件,造成 280000 名病人的社会保障号被下载②.

上述这几类挑战并不单独存在,它们常常交织在一起.例如在大数据处理中就要同时兼顾到存储、通信和安全等问题.由于通信数据量和安全的关系,将多源异构的原始大数据传输到一个中心站点再进行挖掘是不合适的[4].这些不同类型的挑战,吸引了来自不同领域的研究者们,都试图从各自的角度为大数据问题提供解决方案.这些领域包括:计算模型、存储、通信网络、计算机体系结构、数据挖掘、编程语言和系统软件、信息安全等.

大数据产生诸多挑战的同时,也带来了前所未有的战略机遇.大数据在继人力、资本之后,成为一种新的非物质生产要素,作为不可或缺的战略资源以支撑科学研究和各类应用服务[6].这些数据,如果得到合适的处理,将会为社会各个领域带来巨大的价值.文献[7]列举了 5 个具有代表性的领域,分别是:医疗保健、公共区域管理、零售业、制造业和个人位置数据.

为了充分挖掘大数据资源的潜在价值,美国一些大学开始开设数据科学或数据分析课程,为达到要求的学生授数据分析专业硕士学位.这些学生的专业特长可以在电子商务或政府机构工作中得到发挥③.美国政府已将大数据提升到国家战略的高度,继 2012 年由政府斥资 2 亿美元启动"大数据研究与发展计划"之后,2013 年 5 月 3 日,白宫又举办了一次大数据研讨会④.世界经济论坛 2012 年将数据列为与货币和黄金同等重要的一种新的经济资产.2013 年,欧盟大数据论坛分别从音视频文档、考虑建筑节能的知识库构建、数据即服务等方面进行了讨论,并准备制定大数据科学的课程计划⑤.同年,日本政府发布了"创建最尖端 IT 国家宣言",全面阐述了 2013～2020 年期间以发展开放公共数据和大数据为核心的日本新 IT 国家战略⑥;同样,在 2013 年,中国国家自然科学基金委员会、科技部分别启动了大数据处理的研究计划,国家统计局组织阿里巴巴、百度、中国联通等 11 家涉足大数据的国内领军 IT 企业制定了关于大数据的发展战略⑦.

2.3　大数据处理研究现状

当前的大数据处理研究可以从以下 4 个方面来看:大数据处理范式、大数据处理算法、大数据处理平台以及大数据处理应用.它们之间的逻辑关系如图 1 所示.

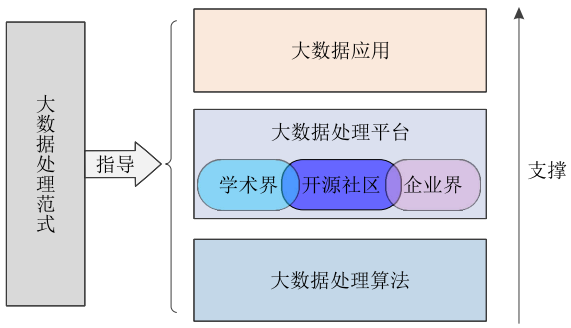


图 1　大数据处理研究的逻辑关系

下面分别介绍这 4 个方面的研究现状.

2.3.1　大数据处理范式

对大数据带来的挑战,信息技术各领域的研究者都在从各自的角度寻求解决方案.计算机系统结构研究者李国杰认为,面对大数据的应用需求,计算机系统结构需要革命性的重构,要由过去的数据围绕计算转变为处理围绕数据转,减少数据的搬运.信

① http://cra.org/ccc/docs/init/bigdatawhitepaper.pdf
② http://www.renci.org/wp-content/uploads/2014/02/0213-WhitePaper-SMW.pdf
③ http://www.nytimes.com/2013/04/14/education/edlife/universities-offer-courses-in-a-hot-new-field-data-science.html?pagewanted=all&_r=0
④ http://www.cccblog.org/2013/05/22/big-data-through-the-years/
⑤ http://2013.data-forum.eu/
⑥ http://www.china-cloud.com/dashujuzhongguo/disanqi/2014/0114/22689.html
⑦ http://news.xinhuanetcom/fortune/2013-12/03/c_118387162.htm

息系统也需要从数据围着处理器转改为处理能力围着数据转.梅宏认为,编程模型和编程语言也需要重新设计①.怀进鹏提出了大数据计算的 3I 理论,即近似性(Inexact)、增量性(Incremental)和归纳性(Inductive),反映了大数据的 3V 特性对数据处理计算模式的变革.

Wu 等人[4]提出了标示大数据演化特征的 HACE 法则,并从数据挖掘视角提出了 3 层大数据处理模型.Fisher 等人[9]讨论了大数据分析中用户交互不便的问题,并且提出了一种新的大数据流程.最近,Chen 等人[10]介绍了大数据的应用、机遇与挑战,以及前沿的大数据技术和处理大数据的基础方法学.

2.3.2 大数据处理算法

大数据处理算法关注数据传输、数据访问以及常规的数据挖掘、机器学习方法等在大数据环境下的改进和迁移.

Havens 等人[11]比较了对大规模数据运用扩展模糊 C 均值聚类的 3 种实现方式的效率,确定了 3 个算法可以较好地近似求解大规模数据的模糊 C 均值聚类.Lu 和 Li[12]研究了在小规模采样的情况下,使用简单随机游走方法进行大数据规模的估计.Sun 等人[13]研究了大数据时代一种新的数据交付模型;Han 等人[14]提出了一种在大数据上进行高效 Skyline 操作的算法;Zhang 等人[15]使用粗糙集方法挖掘大规模数据,在不同的 MapReduce 运行平台(Hadoop/Phoenix/Twister)上,比较了 4 个数据集的不同计算核数运行时间和不同节点数加速比.

除了使用"大数据"概念的数据处理算法研究之外,还有针对海量数据和动态数据流开展的研究,这些研究也对大数据处理同样具有指导作用.

对海量数据的处理,使用特征筛选[16]、降维[17]、数据离散化[18]、子空间聚类[19]等方法均有助于解决数据的巨量性问题.

在动态数据和流数据的处理方面,近年来,国内外学者对动态数据"在线增量学习"这一领域开展了大量的研究工作.Zhang 和 Wang[20]在 Suykens 回归最小二乘支持向量机的数学模型基础上,根据分块矩阵计算公式和核函数矩阵本身的特点,提出一种支持向量机的增量式学习算法和在线学习算法,该算法能充分利用过去的训练结果,减少存储空间和计算时间.Lughofer 等人[21-23]针对演化模糊分类器中在线增量式学习进行了一系列的研究;Zhang 等人[24]研究通过在线学习检测恶意网页,提出了在进化模糊分类器中在线增量式为特征赋权重的方法,将特征选择改为了特征赋权重,从而在柔性和连续的属性变化中,获得平滑的学习过程.Hu 等人[25]研究了基于粗糙集的增量式知识获取方法,例如基于基础集的粗糙集增量式约简算法和基于粗糙集和规则树的动态知识获取算法 RRIA[26].Li 等人[27]研究了可变属性集下优势粗糙集近似的增量更新方法,首先引入一种优势矩阵计算 P-优势集和 P-劣势集,然后讨论属性增减时 P-优势集和 P-劣势集的更新原则,最后给出了优势粗糙集近似的增量更新方法和算法.

数据流挖掘方法包括两类:一类是针对数据来考虑,将数据流改造为普通数据集,包括采样、负载削减、草图和概要数据结构等;另一类针对任务,改造现有模型或者创建新技术以应对数据流处理带来的挑战,包括近似算法、滑动窗口、算法输出粒度[28].新竹交通大学 Li 等人[29]开展了在数据流中挖掘频繁项集的研究,提出了通过一次扫描挖掘到整个历史数据流中所有频繁项集的 DSM-FI 算法以及在包含固定事务数目、事务敏感的滑动窗口内挖掘频繁项集的 MFI-TransSW 算法[30].最近,Li 和 Wu 等人[31]提出了追踪数据流中经常性概念漂移的半监督分类算法 REDLLA,其在生成决策树时,使用基于 K-means 的聚类算法为叶节点上的未标记数据进行标记,获得了更好的准确性和时间性能.Gomes 等人[32]研究了如何在动态特征空间中挖掘经常性概念(recurring concepts),同时降低存储已有模型的内存耗费.Reddy 等人[33]提出,数据流挖掘的要求包括"数据约简和概要构建"、"对时变结果进行建模"和"实时响应".

2.3.3 大数据处理平台

以大数据处理算法作为基础,一批来自高校、开源社区和企业研究院的研究人员们致力于大数据处理平台的研发,不断推出性能更好、适用范围更广的处理平台.

中国人民大学高性能数据库研究小组基于 Postgresql 进行了技术原型(LinearDB)的实现,取得了较大的性能提升.中国科学院计算技术研究所针对 MapReduce 开展了索引优化的研究并利用分布式内存 Cache 提高了 MapReduce 的性能[1].此外,R 作为一种广泛使用的数据操作、分析和可视化工具,集成了大量的统计和机器学习软件包.2012 年

① http://news.ciencenetcn/sbhtmlnews/2013/10/279292.shtm

Purdue 大学统计系研发了基于 Hadoop 的 R 环境 RHIPE，可以用来做大规模复杂数据分析的分解和合成①.

大数据时代为开源软件阵营带来了新活力，Apache基金资助了众多大数据相关项目.

大数据处理开源软件主要包括文件系统(Google GFS 的开源实现——HDFS)、MapReduce 运行库(Hadoop MapReduce/Yarn/Phoenix/Twister 等)、NoSQL 数据库系统(HBase/Cassandra/MongoDB/CouchDB 等)、大规模并行数据查询引擎 Cloudera Impala(Google Dremel 系统的开源实现)、静态数据分析工具(Pig/Hive/Mahout/Drill/Shark)、流数据分析工具(Apache S4/Storm)、内存加速集群计算系统 Spark 等.

这些开源软件平台是大数据分析的技术基础，尤其是其中的 Hadoop 已经成为了事实上的大数据处理标准平台. 比如，IBM、微软和 Intel 的大数据解决方案中，均使用了 Hadoop 作为基础；很多科研机构也在 Hadoop 上开展大数据分析实验.

国内外的知名 IT 企业从软/硬件和网络集成的角度开展了针对大数据的研究，IBM、Oracle、EMC、曙光、浪潮分别推出了 PureData、ExaData、EMC Greenplum、XData 和“云海”大数据一体机. 但是这些仅仅是 IT 厂商面对大数据机遇所做的一些尝试②.

除了大数据一体机，IBM、微软、Intel 以及华为等 IT 公司还推出了各自的大数据软件平台. IBM 推出了大数据处理平台 InfoSphere BigInsights③ 和 InfoSphere Streams[34]，将 Hadoop 引入了企业级的应用，在智慧医疗、智慧银行和电信行业实施了多个业务解决方案. 微软的 Power Pivot、Power View 等 BI 工具可以帮助分析 Hadoop 中的非结构化数据，SQL Server 2012 Parallel Data Warehouse 也嵌入了 PolyBase 功能以无缝集成关系和非关系数据④. Intel 公司除了在处理器、固态硬盘、以太网控制器等硬件基础设施上研究对大数据处理更加有利的技术之外，还针对 Intel 硬件优化了开源的 Apache Hadoop 平台，形成了 Intel 的发行版⑤. 据 Intel 的内部测试表明，Intel 发行版获得了 5 倍的性能提升. 国内方面，华为 2013 年 9 月发布了企业级大数据分析平台 FusionInsight Hadoop，用来处理大量非结构化和半结构化类型的数据，另一个产品 FusionInsight Streaming 是 FusionInsight 中的实时数据处理引擎，以事件驱动(Event-driven)模式处理实时数据，解决高速事件流的实时计算问题⑥.

Louridas 和 Ebert[35] 介绍了在嵌入式大数据分析和统计中，部分软件包、工具和语言的应用.

2.3.4　大数据处理应用

大数据处理应用研究在不同的应用领域，如何利用大数据分析发掘出数据背后的价值，使其对科学研究和生产生活具有指导性.

生命科学领域，根据 2013 年 6 月 Marx[36] 在《Nature》杂志上的介绍，使用计算机和大量软件工具进行大数据分析，计算生物学家则不再需要接触活体就可以进行基因组测序和药物发现等领域的研究. Turk-Browne[37] 2013 年 11 月在《Science》上发表了人脑功能交互作为大数据的文章，指出了关于人脑功能的本质和复杂性的 4 个方面：神经表示广泛分布在人脑的各区域内和区域之间；神经过程取决于脑区域之间的动态交互；这些交互随着认知状态系统地改变；可能的交互空间具有很高的维度.

政府公共管理方面，2012 年美国加州大学的人类学家和数学家共同建立了犯罪活动预测平台. 通过分析过去 1300 多万起案件，寻找发案与日期、天气、交通状况及其他相关事件的关系，利用大数据分析优化警方调度管理，在不增加警力的情况下，6 个月间加州 Santa Cruz 的盗窃案件下降了 19%⑦；德国特伦托的大数据平台可以分析出中午时段慕尼黑地区的打出租车容易程度分布情况，以便指导出租车行驶路线，提高顾客满意度[38].

图像处理方面，斯坦福大学的 Le 和 Ranzato 等人与 Google[39] 联合进行研究，使用一个深度神经网络在 1000 台机器(16000 核)组成的集群上运行 3 天，从 37000 幅人脸图片中学习出一个“人脸检测器”；这个神经网络放在猫脸和人的身体图片数据库上进行学习，也得到了很好的结果. Kim 等人[40] 研究通过图像集匹配来进行人脸识别的相互正交子空间的在线学习，提出了一种在线更新判别子空间，使其成

① Divide and Recombine (D&R) with RHIPE-Deep Analysis of Complex Big Data using the R Environment. http://www.datadr.org/

② http://www.zd-lab.com/Html/News/HY/2477.html

③ http://www-01.ibm.com/software/cn/data/infosphere/hadoop/

④ http://www.windowsazure.com/zh-cn/solutions/big-data/#benefits

⑤ http://www.intel.cn/content/www/cn/zh/big-data/apache-hadoop-integration-in-big-data.html

⑥ http://enterprise.huawei.com/ilink/cnenterprise/news-event/news/news-list/HW_280326

⑦ http://www.fbi.gov/stats-services/publications/law-enforcement-bulletin/2013/April/predictive-policing-using-technology-to-reduce-crime

为一种可持续改进识别准确度的方法. Wu 等人[41]研究了使用 Markov 毯进行在线特征流的特征筛选,并将结果应用到火星撞击弹坑的自动检测中.

材料科学方面,创造一种全新的突破性材料极其费时. 为了突破这个科技瓶颈,美国政府成立了材料基因组计划(Materials Genome Initiative,MGI)项目. MGI 的一些研究工作已经开始借助大数据处理技术来进行模拟实验,然后通过分析数据向有潜力的方向进行深入研究①.

金融与商务领域,Chen 等人[42]介绍了商务智能与分析(BI&A)的演进、当前的研究以及大数据分析在 BI&A 中的地位.

大数据在其他领域(如流程工业、医疗保健、电子商务等)也存在很多应用研究,在此就不一一列举了.

2.4 目前大数据处理研究存在的问题

尽管学术界、开源社区和 IT 企业已经为大数据分析作了大量的研究和开发工作,但是距离完善解决大数据带来的挑战,还有很长的路要走. 目前大数据处理研究还存在如下问题:

(1)缺乏人类级机器智能的大数据处理模型

Zadeh[43]认为,由词计算进一步演变到精确化的自然语言计算,是通往人类级机器智能的基石. 现存的大数据处理计算模式,没有考虑到模拟人脑思考问题时对粒度的准确把握. 在研究以人为中心的智能分析处理系统中,中国科学院生物物理研究所陈霖等人通过实验发现,人类认知具有"大范围优先"的规律,视觉系统对全局拓扑特性尤为敏感[44-45]. 因此,直接处理最细粒度的原始数据不符合人类的认知规律.

(2)传统数据处理模式不能适应大数据的处理需求

大数据的 3V 特性使得传统的数据分析、数据挖掘、数据处理的方式方法不再完全适用. 因此,面对大数据,我们需要有新的数据计算范式和计算模式,需要提出数据计算的效率评估方法以及研究数据计算复杂性等基本理论问题.

(3)大数据处理中缺乏降低数据规模的有效方法

既然大数据带来的很多问题是由数据规模太大而导致的,那么在保证数据价值的前提下将数据规模变小不失为一种解决方案. 如果总是从原始的最细粒度出发处理数据,必然造成对原始数据的过度依赖. 大量的检索和分析都在此基础上进行,会给存储、通信和计算带来巨大的压力.

(4)缺乏在约束条件下为大数据领域问题及时提供有效解的方法

基于最细粒度进行问题求解不能从容应对用户的变粒度需求. 很多情况下,用户会根据不同场景提出属于同一问题簇(问题性质相同,但对解的粒度要求不同)的问题. 如果没有预先建立粒结构,那么每次都只能从最细粒度开始计算,而不能在计算环境受限的情况下,退到较粗粒层上及时进行非精确求解.

(5)大数据中的不确定性问题

由于不确定性普遍存在于大数据中,而数据集成和粒化不能够完全消除它. 因此,Pei[46]认为对不确定和概率数据挖掘已成为当前大数据处理中的一个重要问题.

(6)编程模型问题

当前大数据处理技术纷繁复杂,仍然处于产业变革的初期. MapReduce 过于复杂灵活,写出高效 Job 比较困难,而且其运行效率较低[47]. 针对 MapReduce 的缺点,Appache 基金会资助了 Spark 项目进行改善. 但各种新的工具/平台仍不能满足大数据分析的多样性需求[10]. 同时,各种技术的不断涌现,增加了开发人员知识更新和企业数据分析应用升级换代的成本.

(7)大数据分析解决方案的代价问题

大数据为商业发展提供了新的机遇,所以需要大量的实验和分析来确定哪些模式重要,并且将数据洞察转化为商业价值. 为了保证大数据项目有较好的投资回报,降低代价非常关键②.

上述这些问题中,(1)至(4)可以使用粒计算的思想帮助解决. 研究人员已经意识到大数据在数据模型上的问题. 比如,当前想要在 NoSQL 的上面构建一个 SQL 层,但是缺乏抽象数据模型,这种努力就会受到底层技术的限制[48]. 传统的数据挖掘技术总是在最细粒度的原始数据上进行分析,而粒计算观点则认为选择与问题相适应的粒度空间有助于提高求解的质量和求解的时空效率. 文献[49]中对主观评价和货物进行分类以提高协同过滤的准确度即是一个很好的例证. 2010 年图灵奖得主、哈佛大学 Valiant 在其个人主页上的"当前研究兴趣"一栏中

① http://www.nist.gov/mgi/

② http://www.gsma.com/membership/wp-content/uploads/2013/07/The-Top-Challenges-of-Big-Data-Analytics.pdf

指出："人工智能的一个基本问题就是，标示认知中必须的计算构建模块"①，此处的"计算构建模块"可以理解为"信息粒".

3　粒计算

从哲学的观点来看，人类在对任何对象进行认知、度量、形成概念和推理时，粒度都贯穿其中. 因此在智能系统的设计和实现中，粒度也有着重要的作用[50]. 早在 1997 年，Zadeh 就指出粒计算是模糊信息粒化、粗糙集理论和区间计算的超集，是粒数学的子集. Yao[50]更概括地认为，粒计算是在解决问题的过程中使用"粒度"的所有理论、方法、技术和工具的"标签". 粒计算并不是一个具体的模型或方法，而是一种方法论[51]. 在它的"大伞"之下，包含了很多具体的模型，如词计算、粗糙集、商空间、云模型、区间集、邻域系统等等.

3.1　粒与粒化的方法

粒是粒计算的基本要素，它是依照不可区分性、相似性和功能标签聚集到一起的论域中的子集、类、簇和元素. 集合的一个子集、论域中的一个等价类、文章中的一节和系统的一个模块等，都是粒的例子[52]. 此前的粒一般定义在论域上，对应于结构数据中的行. 由于大数据的原因，列数据库近年来得到了学术界和 IT 业界的广泛关注[53]. 在列(属性)上也可以有粒化，具体的实现方法有属性约简(特征选择)、主成分分析等.

进行粒计算的第一步是确定采用哪种具体模型，然后根据相应的粒表示进行粒化，粒化有两个方向：构建和分解(construction and decomposition). 构建指的是如何将更细的或下层的粒合并成为较粗或上层的粒，分解则相反，是将较粗的或上层的粒分解成更细的下层的粒[52]. 粒化的目的是从原始数据中得到合适于问题的粒.

粒的解释是粒构造的语义方面，需要回答"为什么两个对象能放在一个粒里". 钱宇华[54]系统地研究了复杂数据的粒化机理. 文献[55]中的粒被赋予更广泛的含义，它包含了简单类型(如不可区分或相似的类)和复杂类型(决策规则、决策规则集、分类器、聚类、时间窗口及它们的簇、时间窗口序列、过程、agents 或 agents 组).

数据粒化算法由以下两步迭代完成[56]：

(1) 找出最相容的两个粒，合成一个.

(2) 重复上一步直到达到一个满意的抽象标准.

粒化过程中有一点很关键，那就是粒的相容程度的定义，即满足什么条件的对象，可以划分到同一个粒内. 这个相容程度可以是几何上的，也可以是密度驱动的，或者是相似性(形状和方向)驱动的.

粒化的具体方法主要有模糊信息粒化[57]、粗糙集近似[58]、商空间法[59]、基于聚类的粒化[54]和云模型法[60-62]等. 下面分别简要介绍：

(1) 模糊信息粒化[57]

虽然清晰粒在很多方法和技术中扮演着重要的角色，但是粒度的模糊性更是人类处理信息的特征. 例如，考虑"人的头部"这个粒，脸颊、前额、鼻子、耳朵的边界并无清晰的划分，从这个意义上说，这个粒就是模糊的. 并且，模糊粒的属性，如鼻子的长度用"长"、"短"和"很长"来表示，也是模糊的.

模糊信息粒化理论(Theory of Fuzzy Information Granulation，TFIG)的出发点是广义约束. 粒的主要类型包含：可能性的、真实性的、概率性的. 粒的特征是由定义它的广义约束来刻画的. TFIG 的主要推广模式有：模糊化、粒化、模糊粒化. TFIG 受到了人类粒化和处理信息的启发，并且是以数学为基础的. 与 FIG 相关的推广模式主要有：

① 模糊化(Fuzzification，f-generalization). 把一个清晰集用模糊集代替.

② 粒化(Granulation，g-generalization). 一个集合被划分成粒.

③ 随机化(Randomization，r-generalization). 变量被随机变量代替.

④ 通常化(Usualization，u-generalization). 命题 X is A 被 Usually (X is A)代替.

其中，模糊化和粒化的组合——模糊粒化尤为重要. 在模糊逻辑中，它从变量、函数和关系出发，得到语言变量、模糊规则和模糊图.

(2) 粗糙集近似[58]

在粗糙集中主要研究两个相关的问题：信息的粒化和近似. 粒结构由论域中元素之间的相似性决定. "相似性"包含了简单的等价关系、容差关系、自反的二元关系、关系族、层次和邻域系统等.

① 简单的粒化和近似：在等价关系诱导的粗糙集近似中，由不可区分元素构成的每个等价类就可以被看作是一个粒. 对于论域的任意子集，它可能不是恰好为某些等价类的并，这就需要引入上、下近似

① http://people.seas.harvard.edu/~valiant/researchinterests.htm

的定义. 设 U 为论域，X 为 U 的任意子集，$[x]_E$ 表示 x 的等价类，X 的下近似 $\underline{apr}(X)$ 和上近似 $\overline{apr}(X)$ 分别定义为 $\underline{apr}(X)=\bigcup\{[x]_E \mid x\in U,[x]_E\subseteq X\}$，$\overline{apr}(X)=\bigcup\{[x]_E \mid x\in U,[x]_E\cap X\neq\varnothing\}$，其中，$[x]_E=\{y \mid y\in U,xEy\}$.

② 层次粒化与近似：如前所述，两个对象要么有关系，要么没有. 为了避免这种局限性，可以利用对象间其他类型的相似性，考察更一般的粒结构和分层粗糙集近似. 通过把简单粒结构放到一起，可以构成多层复杂粒结构，复杂结构的每一层是一个简单的结构. 嵌套的二元关系序列可以定义粗糙集近似. 如果对所有的 $x\in U$，都有 $[x]_{E_1}\subseteq[x]_{E_2}$ 且 $E_1\subseteq E_2\Rightarrow\underline{apr}_{E_2}(X)\subseteq\underline{apr}_{E_1}(X)\subseteq X\subseteq\overline{apr}_{E_1}(X)\subseteq\overline{apr}_{E_2}(X)$，则称等价关系 E_1 比 E_2 细，粒度越细，准确度越高：

$$E_1\subseteq E_2\Rightarrow\alpha_{E_2}(X)\leq\alpha_{E_1}(X)$$

反之不一定成立. 更一般地，可以考虑 m 个嵌套的等价关系序列，结论相似.

(3) 商空间法[59]

商空间是将不同的粒度世界与数学上的商集概念统一起来表示对象的模型. 用一个三元组 (X,f,T) 描述一个问题. X 表示问题的论域；$f(\cdot)$ 表示论域的属性，可用函数 $f:X\to Y$ 表示（Y 为属性的取值）；T 是论域的结构，指论域 X 中各元素之间的相互关系.

问题的不同粒度表示对应于不同的不可区分关系 R. 也就是说，不同的粒度仅仅是对论域进行不同的划分. 因此，划分是构成不同粒度世界的基本方法，可依据结果 Y 对 $X=f^{-1}(Y)$ 进行分类，也可直接对 X 进行分类. 具体来说，有下述划分法.

① 属性划分法. 即将属性相同或相似的元素归为一类.

② 投影划分法. 若元素 x 的属性函数是多维的，如有 n 个属性函数分量 $f_1,f_2,\cdots,f_n$，若暂不考虑其中的 i 个属性 $f_1,f_2,\cdots,f_i$，将 $f_{i+1},f_{i+2},\cdots,f_n$ 几个属性相同的元素归为一类.

③ 结构划分法. 把结构上或功能上关系密切的元素分为一类.

④ 约束划分法. 设有 n 个约束条件 $C_1,C_2,\cdots,C_n$，那么可按 C_i 进行划分.

当利用分类技术在粗粒度世界讨论问题时，若问题无解，那么在细粒度的原问题上也无解（保假原理）. 一个命题在两个较粗粒度的商空间中是真的，则（在一定条件下）在其合成的商空间中对应的问题也是真的（保真原理）[63]. 这样就可缩小求解范围，加快求解进度，因为粗粒度世界通常比原世界简单.

对于分类有相交的情况：当 X 中的同一元素 x 同时属于 X 中不同的类别，即分类的边界不分明时，可引入模糊逻辑的概念，对分类相交问题进行讨论.

文献[64]指出，商空间方法与粗糙集方法共同之处在于，都是利用等价类（或其他二元关系诱导的划分、覆盖等）来描述“粒度”，用“粒度”来描述概念的粗细. 但是，他们讨论的着重点不同. 商空间强调研究不同粒度世界之间的互相转换、互相依存关系，是描述空间关系学说的理论；而其他的粒度计算（如粗糙集理论等）主要是研究粒度的表示、刻画和粒度与概念之间的依存关系. 商空间是在论域元素之间存在拓扑关系的情况下进行研究的，即论域是一个拓扑空间. 而粗糙集理论的论域是元素集，元素之间没有拓扑关系.

(4) 基于聚类的粒化[54]

聚类的基本思想是先确定对象间相似程度的度量，然后寻求一种方法，对论域进行分割，使得同一类的对象之间相似程度尽量大、不同类的对象之间相似程度尽量小. 根据聚类结果的结构，聚类可以分为划分聚类和层次聚类. 划分聚类得到论域上的一个划分：

设论域 $U=\{x_1,x_2,\cdots,x_n\}$，划分聚类得到 k 类 $C=\{\omega_1,\omega_2,\cdots,\omega_k\}$，$1<k<n$，使得 $\omega_i\neq\varnothing$，$i=1,2,\cdots,k$；$\bigcup_{i=1}^{k}\omega_i=U$；$\omega_i\cap\omega_i=\varnothing$，$i,j=1,2,\cdots,k(i\neq j)$.

每个对象归属性确定的聚类是清晰聚类，与之相对的是模糊聚类，即每个对象以不同的隶属度同时归属于不同的类.

层次聚类得到一个嵌套的树形结构，即 $H=\{H_1,H_2,\cdots,H_Q\}(Q\leq n)$，使得

$$\omega_i\in H_m,\ \omega_j\in H_l,$$

且 $m>l$，对于 $i,j\neq i;m,l=1,2,\cdots,Q$，有 $\omega_i\subseteq\omega_j$ 或 $\omega_i\cap\omega_j=\varnothing$.

两种聚类可以得到相应结构的信息粒.

Pedrycz[65] 详细介绍了多种基于知识的聚类方法，包括模糊聚类、半监督聚类、协同聚类、方向聚类等等. Lin[66] 在 2010 年提出源自于谱聚类的幂迭代聚类（Power Iteration Clustering，PIC）方法，它在样本上定义亲和性矩阵，然后对其规范化，创建隐含数据的一维嵌入（embedding）. 使用这个嵌入就可以进行 K-means 聚类，它不需要找出所有的特征向量，只需要顶层特征向量的线性组合. PIC 算法具有很好的聚类精度和收敛速率. 2013 年 Yan 等人[67] 提出了一种大数据上的并行幂迭代聚类方法 p_PIC.

它改进了 PIC 算法需要将数据和相似矩阵装入内存这一不足，使用并行策略增加了数据的可伸缩性，在设计并行策略和实现细节中使计算和通信花销最小化，而且还重点考虑如何让算法在低端计算机上运行. 文献[54]还讨论了带测量误差的聚类.

(5) 基于云模型的粒化[60-62]

云模型是由我国李德毅创建的不确定性知识表示和推理模型，它可以揭示概念的随机性、模糊性以及随机性和模糊性之间的关联性，用期望、熵和超熵作为数字特征表示定性概念，并通过云变换实现定性概念(概念内涵)和定量数据(概念外延)之间的相互转换. 云模型的定义如下：

设 U 是一个用精确数值表示的定量论域，C 是 U 上的定性概念，若定量值 $x\in U$，且 x 是定性概念 C 的一次随机实现，x 对 C 的确定度 $\mu(x)\in[0,1]$ 是有稳定倾向的随机数，即 $\mu:U\to[0,1]$，$\forall x\in U$，$x\to\mu(x)$，则 x 在论域 U 上的分布称为云模型，简称为云，每一个 x 称为一个云滴[68].

云模型用期望 Ex、熵 En 和超熵 He 这 3 个数字特征来表征一个概念，它们反映了定性概念 C 整体上的定量特征.

使用云模型进行粒化的思路是使用逆向云发生器，从数据集中学习得到反映定性概念的数字特征集$\{(Ex_i,En_i,He_i)\}$，$i=1,2,\cdots,n$[62]. 然后，将数据划分成 n 个粒.

利用云模型进行图像分割[60]，本质上就是基于逆向云发生器的粒化过程，分割后得到的每个区域对应于一个粒. 将图像分割中对每个像素的处理推广到对其他类型数据集中对象的处理，即经过如下的关键步骤就可以实现粒化：① 云变换；② 泛概念树的自动生成和爬升；③ 使用极大判定法[61].

除了上述 5 种，还有其他的粒化方法，如基于概念格[69]和基于领域系统[70]等，在此不再一一介绍. 值得注意的是，有时各种不同的信息粒之间需要进行通信，文献[71]中讨论了不同信息粒的交互机制.

3.2　应用“粒”求解问题

自粒计算概念提出以来，过去十多年间吸引了大量研究人员和从业者的兴趣. 这些研究工作按其目的大致分为两类：一类是使用粒计算思想作为指导，对计算智能相关模型进行扩展的基础理论研究[72-77]；另一类是在处理数据模型中使用信息粒的应用研究[78-81].

根据运用“粒”求解问题时对粒度层次的使用特征，这些研究工作又可以归纳为 3 个基本模式：粒度空间优化、粒度层次切换和多粒度联合计算.

3.2.1　粒度空间优化

粒度空间优化是指针对问题性质和计算的约束条件(时间、计算资源和通信带宽等)，在问题数据的多粒度表示空间中选择合适的粒层. 问题本身可能显式地规定了解的粒度，也可能没有，但可从问题背景推测出满意解的粒度. 这个由问题本身和问题背景决定的满意解的粒度，简称为问题粒度. 最后计算得出的解本身也有粒度，从问题粒度到解的粒度存在着映射关系 M_{PS}，解的粒度应该等于问题粒度或者比问题粒度更细. 而要得到特定粒度上的解，在求解过程中，计算对象就应该选择在合适的粒度层次上. 将计算过程中所处理的信息粒的粒度简称为计算粒度. 从解的粒度到计算粒度，也存在着映射关系 M_{SC}，这个关系是由信息的粒化模型和基于信息粒的问题求解模型二者共同决定的. 利用 M_{PS} 和 M_{SC}，就可以得出从问题粒度到计算粒度的映射 M_{PC}. 粒度空间优化的实质就是求得映射 M_{PC}，如图 2 所示.

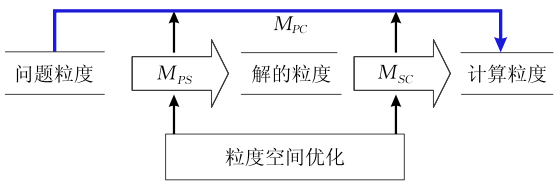


图 2　问题粒度到计算粒度的寻优与映射

在粒度空间寻优问题上，近年涌现了很多相关研究成果. 比如，Tang 和 Zhu[82]研究了粒度空间的层次聚类和模糊近似关系分析，其中层次聚类问题包括层次聚类的粒度表示、一致聚类、最优化聚类、协同聚类和层次聚类分析等. Nandedkar 和 Biswas[83]提出了粒度反射模糊最小-最大神经网络(GrRFMN)，能够对粒度数据进行学习和分类. 粒计算领域的概念和算法在不同的形式框架(模糊集、粗糙集等)下各不相同，因此该领域需要一种可靠的基础理论. Pedrycz[84]指出粒模型是数值模型的泛化(generalization)，说明了信息粒的分布如何由数值函数的参数确定，可以被描述为一个优化问题；并指出粒模型设计是一种新的模型构造，它可以将现存的数值模型提升到下一个抽象水平. 类似地，Pedrycz 还与 Homenda 一道，提出了“合理粒度”(justifiable granularity)的概念，将信息粒的构建与优化问题联系起来[85]. 还有一些研究，虽然不是来自粒计算领域的研究者，也没有明确使用粒计算的术语描述其工作，但他们解决问题的思路本质上采用了粒度的思想. 例如 2013 年 Nakatsuji 和 Fujiwara[49]对主观

评价和货物进行分类以提高协同过滤准确度，就是把最细粒度数据抽象到一个合适的较粗粒度上，再以此为基础进行问题求解.

3.2.2 粒度层次切换

粒度层次切换是指问题的解可以在不同粒层上求得，重点研究相邻粒层上解的快速重构方法，如图 3 所示. 图中，$G(V,E)$ 和 $G'(V',E')$ 是两个粒层，f 是从 $G(V,E)$ 到 $G'(V',E')$ 的映射关系，S 和 S' 分别是问题 Problem 在粒层 G 和 G' 上的解，求解过程用函数 $h()$ 表示. 在问题和粒层映射已知的情形下，问题在较粗粒层 G' 上的解其实可以通过两种途径求得，一种是 $S'=h(\text{Problem}(G'))$，即直接在 G' 上“从头求解”；另一种是先找出粒层关系 f 和问题 Problem 所决定的不同粒层上解之间的映射关系 f'，然后求出 $S'=f'(S)$. 由于 G 和 G' 之间具有相对简单的逻辑关系，我们希望通过 $S'=f'(S)$ 重构出解比从头求解更为高效. 如何求得 f' 是这个问题的关键.

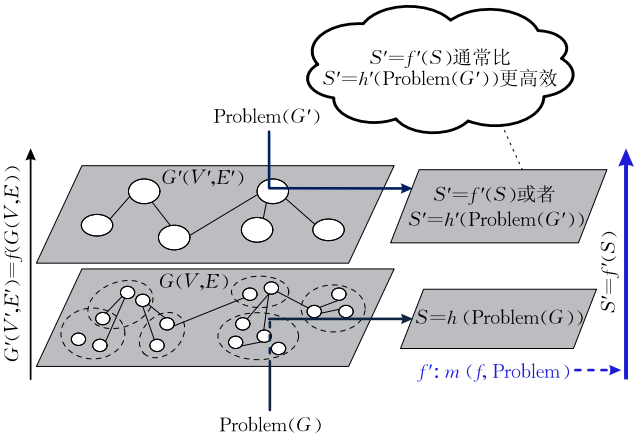


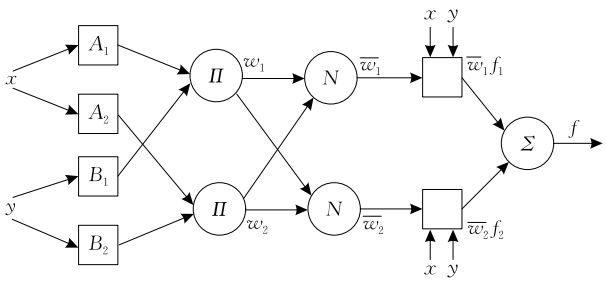
图 3 基于粒层切换的解的快速重构机制

很多研究者已经关注到粒度层次切换问题，并做了一些有意义的工作. Zhang 和 Zhang 等人[59,86]提出的商空间理论较早研究了不同粒度空间上解之间的关系. Zhu 和 Hu[87]研究了使用集成学习方法，在不同的粒度空间中训练基分类器，然后使用边缘分布优化来衡量各个基分类器的重要性，最终实现自适应地选择邻域粒度和组合. Hobss 和 McCalla 等人[88-89]研究了粒度的转换，指出粒度具有两种含义：聚合和抽象，类似于面向对象程序设计中的 ISA 和 Part-Of 关系. 王国胤等人[90]综述了不同粒度下粗糙集的不确定性，并讨论了不同知识粒度下粗糙集的粗糙度和模糊度的变化关系；开展了确定性和不确定性在不同粒层的转化规律研究[91-92].

3.2.3 多粒度联合计算

多粒度联合计算是指，将复杂问题的求解分配到数据表示的多个粒度层次上成为子任务，各个粒度层次上相对简单的功能协同起来，最终完成复杂问题的求解.

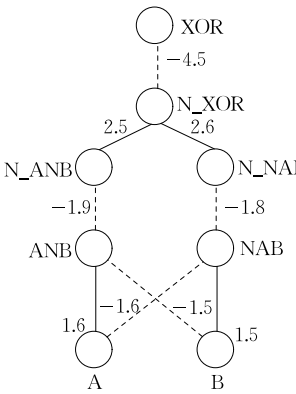
多粒度联合计算的研究始于 20 世纪 90 年代. Jang[93]提出了基于自适应网络的模糊推理系统 ANFIS，该系统使用 5 层神经元实现模糊推理，如图 4(a)对于只有两条 Takagi 与 Sugeno 类型模糊规则的推理系统来说，每层中神经元的功能依次是计算隶属度、t-norm、规范化的激活强度、内部活动函数与激活强度的乘积、求和. Wang 和 Shi[94]提出了三值/多值神经网络 TMLNN，构建了能够正确处理三值/多值逻辑问题的神经元模型，并对其逻辑特性进行了详细分析. 由这种神经元构造的多层神经网络能够用于表达处理任意的三值/多值逻辑推理问题，图 4(b)是一个实现“异或”的例子. 2013 年被 MIT 技术评论选为 10 大突破性科技之首的深度学习①，本质上就是多粒度联合计算. 因为不管是使用受限玻尔兹曼机(Restricted Boltzmann Machine，RBM)[95]、自动编码器(auto-encoders)[96]，还是使用其他的方法(如半监督嵌入[97])，深度学习都遵



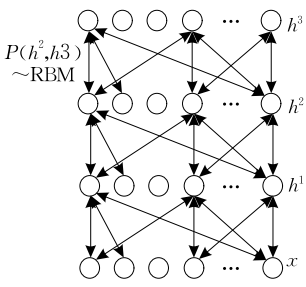
$w_i=\mu A_i(x)\times\mu B_i(x),\ i=1,2.$ $\quad f=\dfrac{w_1f_1+w_2f_2}{w_1+w_2}=\overline{w}_1f_1+\overline{w}_2f_2$

$f_1=p_1x+q_1y+r_1,\ \ f_2=p_2x+q_2y+r_2$

(a) ANFIS使用5层神经网络实现3型模糊推理[93]



(b) TMLNN 使用5层神经网络实现“异或”运算[94]



(c) 观测向量为$\boldsymbol{x}$，隐含层为h^1、h^2、h^3的深度信念网(DBN)模型[98]

图 4 多粒度联合计算的 3 个代表实例

① http://www.technologyreview.com/lists/breakthrough-technologies/2013/

循如下的原则：使用无监督学习训练中间表示层[98]，然后再由若干较低层次组件抽象出高层次的概念.

深度学习作为多粒度联合计算的一种范例，取得了巨大的成功. 最近 Facebook 研发的 DeepFace 面部识别技术精准度为 97.25%，接近普通人类 97.5% 的识别精准度①.

此外，与 agents 团队组建相关的研究工作，都可以借鉴到多粒度联合计算中来，因为可以将每一个 agent 看作是一个功能粒. Liemhetcharat 和 Veloso[99] 还提出了使用带权协同图组件将异构临时 agents 组成团队的方法，解决复杂团队组建中存在的问题：单个 agent 在团队中的功能预先不知道，以及整个团队的能力依赖于 agents 协同的方式且超越单个 agent 的功能.

4　使用粒计算处理大数据

4.1　使用粒计算处理大数据的可行性分析

信息粒化旨在建立基于外部世界的、有效的、以用户为中心的概念，同时简化我们对物理世界和虚拟世界的认识，可以高效地提供“实用”的非精确解[50]. 粒计算已经成为一种发展迅速的信息处理范式[52]. 澳门大学陈俊龙[10]将粒计算列为驾驭大数据的第一方法.

美国计算机研究协会计算社区联盟(CCC)在《Challenges and Opportunities with Big Data》白皮书②中给出了典型的大数据信息处理过程，包括：数据获取与存储、信息提取与清洗、数据集成/聚集/表示/查询处理、数据建模与分析和解释等阶段. 在上述大数据分析的各阶段中，都必须面对如下挑战：

(1) 异质性和不完备性. 数据具有不确定性，通常与领域和具体应用相关.

(2) 数据规模. 数据的规模和增长速度一直都是一个挑战.

(3) 及时性. 很多任务需要在规定时间内完成，否则分析的结果就会失去意义.

(4) 隐私保护. 如何让数据在得到充分利用的同时，又保证公民的隐私不受侵犯是大数据面临的挑战之一.

(5) 人工协同. 虽然计算分析已经取得长足的进步，但有时候人类可以轻松检测到某个模式，而计算机算法却难以完成，因此大数据分析必须支持多位专家协同决策.

表 1 分析了利用多粒度智能计算是否可以应对大数据处理面临的挑战.

表 1　粒计算应对大数据处理面临的挑战

挑战	粒计算是否能解决	理由
不确定性	适合	通过粒度变换实现确定信息和不确定信息的转换.
异质性	配合解决	通过不同类型粒数据模型对不同的非结构数据(文本、音频、视频等)进行建模.
数据规模	适合	信息粒结构是对最细粒度原始数据的抽象概要表示，大幅度降低了原有的数据规模.
及时性	可以	在具体应用需求提出之前，先建立起面向领域的粒结构. 由于数据规模的减小，必将提高求解的时间效率.
隐私保护	适合	信息粒结构可以隐藏细节信息. 隐私信息一般是以最细粒度原始数据的形式存在；经过粒化之后，在传输和处理过程中，规避了隐私泄露的风险.
人工协同	可以	将领域专家的知识表示为人工信息粒，添加到由原始数据产生的信息粒中，以及将领域专家知识编码为高粒度形式化知识，这些都可以作为大数据处理人工协同的方式.

通过表 1 可以看出，粒计算对于大数据处理中面临的主要挑战均有着十分积极的作用. 尤其是在降低数据规模方面，很多粒计算的研究都认为使用信息粒可以降低数据量，或者说可以实现数据压缩(一般是有损压缩)[100-101]. Wang 等人[102]研究了在覆盖粗糙集中，通过引入“一致函数”的概念，进一步得到基于覆盖的信息系统的同态，最终实现对数据集的压缩. Panoutsos 和 Mahfouf[103]在对工业数据的处理中，使用粒计算进行数据压缩和过滤. 通过定义两个粒之间的相容性，使用文献[56]中提到的粒化算法，即可实现对原始数据的压缩，同时还可以排除噪声和不精确数据的影响.

在使用粒计算方法处理大规模数据集方面，已经有了一些研究成果. 如 Yang 等人[104]针对时空相关性对云平台上的图大数据进行压缩后，再以数据驱动的方式进行计算资源调度，在减小数据信息损失的前提下，获得了更高的时间效率和资源利用率，其中的压缩技术主要采用了聚类方法. Ruan 等人[105]使用模糊信息粒化方法先对时间序列进行粒化，然后使用 SVMs 对粒化了的时间序列进行回归分析和预测，提高了大规模时间序列分析的速度. 根据文中实验描述，其数据规模其实能在单个 PC 上进行处理.

4.2　粒计算代表模型在大数据处理中的应用

运用粒计算模型处理大数据，需要考虑的首要

① http://www.cs.tau.ac.il/~wolf/papers/deepface_11_01_2013.pdf

② http://cra.org/ccc/docs/init/bigdatawhitepaper.pdf

因素是计算复杂度,因为在大数据分析中对时间复杂度的要求会更加苛刻;粒计算是模拟人类思维的一种计算范式,所以要分析各个模型如何融入分析者的主观性;大数据分析强调数据之间的关系,因此要比较各个模型的粒结构(粒层之间的关系和粒层内关系)表达能力;由于大数据的几个典型特征在具体场合中不一定同时具备,因此分析各个模型对大数据不同特征的适应性可以为模型选择提供建议.

此处我们尝试归纳 5 个粒化模型在大数据处理中的优缺点及适用场合,见表 2.

表 2 粒计算的代表模型在大数据处理中的适用性

粒化方法	典型时间复杂性(n 为样例数,m 为属性数,s 为分类数)	主观性表达方式	粒结构表达能力	对大数据特性(巨量/异构/高速)的适应能力
模糊信息粒化	$O(n\times l)$,l 为粒化的属性数	隶属度函数的设定	较弱	对巨量性适应能力较强,但对复杂关系适应性较弱;宜与其他模型结合应用
粗糙集近似	$O(n^2\times s)$或 $O(n\times s^2)$或 $O(n\times s\times(s+\log n))$[104,106]	相容关系、相似关系或限制容差关系等的定义	较强	已有针对大数据的增量式并行算法[107],适应性较强
商空间法	覆盖法:$O(n^2m/k)$,k 为每个覆盖下的样例数目	划分方法的指定	强	对大数据的巨量性和复杂结构均有较强适应能力
聚类粒化	K-Means:$O(tsn)$,t 为循环次数;CURE:$O(n^2)$;	类别数目的指定	较强	多数的聚类算法容易并行化,对大数据环境适应性强
云模型粒化	H-GCT[108]:$O(n\times s)$ A-GCT[108]:$O(n^2)$	类别数的指定(H-GCT)或概念清晰度 β 的指定(A-GCT)	较弱	算法复杂度较低,容易处理大规模数据

4.3 大数据环境下的粒计算处理框架

针对大数据的特性,我们提出统一的大数据问题粒计算解决框架,如图 5 所示. 大数据的 3V 特性在一轮中可以按如下的顺序处理:多样性→巨量性→高速性(当然,有些数据本身不同时具有这 3 个特性,需根据实际情况进行取舍).

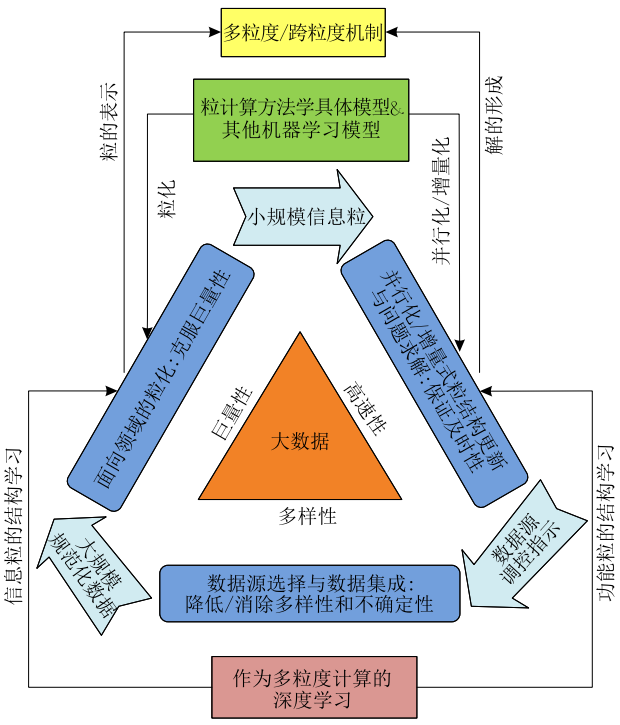


图 5 大数据环境下的粒计算处理框架

(1) 使用数据过滤和数据集成将分布式存储的多样、异质数据进行转换、抽取、粒化,得到较为规范的数据表,消除其中的不确定性.

(2) 针对问题,使用粒计算“大伞”下的具体模型和技术将原数据粒化为粒度大小合适的粒,降低数据规模,并构建相应的粒层及各粒层上的结构.

(3) 在其他机器学习方法的辅助下,对信息粒进行数据挖掘或者机器学习.

(4) 将用到的方法改造为分布式的、在线增量学习的版本以满足大数据处理的及时性要求.

(5) 在处理大数据中,粒度的自由切换,需要考虑多个粒度层次上粒的分解与合并,还有相应解的快速构建;对某些特定问题,需要同时考虑多个粒度层次的信息,使用“跨粒度”机制求解问题.

(6) 从整个处理过程中,可以发现原始数据是否具有合适的粒度,为是否需要调整及如何调整原始数据的产生或采集提供指导.

(7) 借鉴深度学习(Deep Learning)思想,将关键的处理流程调整为多个层次,让具体参数(如粒的大小和粒层次数)在学习中得到优化,从而优化最终学习结果.

图 5 所示的框架与大数据处理流程(数据获取→抽取/清洗→集成/表示→分析/建模→解释)之间有明确的对应关系. 右下角的箭头“数据源调控指示”实际上是根据前一阶段的分析应用对数据粒度(采集或产生的精确度和频度等)做出调整,对应于“数据获取”;接着的“数据源选择与数据集成”对应于“抽取/清洗”;“面向领域的粒化”对应于数据的“集成/表示”;上方的“粒计算方法学模型&其他机器学习模型”和右上方圆角矩形内的“并行化/增量式粒结构更新与问题求解”对应了“分析/建模”;由

于信息粒本身就具备明确的语义，因此粒化和运用挖掘/学习模型进行分析的过程都有明确“解释”.

接下来，介绍在图 5 所示的框架中，用圆角矩形表示的 3 个主要环节. 重点分析在每个环节中需要解决的问题及可以采用的方法.

4.3.1　数据源选择与数据集成

大数据处理的第一个环节是要确认哪些数据对于问题的解决可能会有帮助，哪些是与主题无关的. 麦肯锡认为这是大数据分析的 3 个关键挑战之一①.

大数据的原始形式一般具有“多样性”，包括语法异构和语义异构. 其中语法异构保持了数据的原子性，仅仅是命名不同或者类型不一致，这种情况较容易处理. 语义异构则涉及数据粒度和数据类型等多个方面的不同，需要仔细分析，然后用元数据来对原始数据进行描述. 例如视频数据，有的应用只需要它的一些基本信息(如场景类别、时长等).

对异构数据的处理方面，Pal[109] 讨论了如何在数据预处理阶段处理数据的异构性，提到的方法有降维、数据浓缩(data condensation)和数据封装(data wrapping). Pedrycz[65] 介绍了对于异构数据，如何进行模糊聚类. 作为大数据分析的准备阶段，数据集成是必不可少的. 关于数据集成的研究已经较为成熟，读者可以参阅文献[110-117].

4.3.2　面向领域的粒化

利用粒计算方法研究大数据问题的非精确求解，目的是将问题的输入从最细粒度原始数据转换为信息粒表示，在保留数据中蕴含的信息和价值的前提下，大幅降低数据量.

面向领域的粒化意味着在具体数据分析需求提出之前，根据领域知识将原始数据先构建为多粒度信息知识表达模型(Multi-Granular Information/Knowledge Representation model，MGrIKR). 构建 MGrIKR 的意义在于为一族解粒度粗细不同的问题提供合适的计算输入.

粒化首先需要研究信息粒、粒层和整个粒结构的表示，然后针对表示方法进行构建.

(1) 信息粒的表示

借鉴商空间中对拓扑空间的表示方法，采用三元组对信息粒进行形式化描述，即 $IG=(KVS, GM, VM)$. KVS(Key Value pair Set)表示描述信息粒的特征子向量，也可以称为键值对集合，即 $KVS=\{\langle key_1, value_1\rangle, \cdots, \langle key_n, value_n\rangle\}$. $value_i$ 表示信息粒中名为 key_i 的特征所取的值，$i=1,2,\cdots,n$. GM 表示该信息粒的粒度度量(Granularity Measure)，即信息粒的粗细程度. VM 表示该信息粒的价值度量(Value Measure).

数据粒化从实例(examples/instances)粒化和特征(features/attributes)粒化两个方向进行. 特征粒化主要指特征的筛选和组合，可以借鉴机器学习中的核函数方法. 实例的粒化可以采用数据挖掘的聚类思想，即先确定一个信息粒层所包含的细粒度数据间相似度的度量标准，然后对论域进行分割，使得同一个信息粒内部各数据相似程度最大、不同信息粒的数据间相似程度最小.

关于信息粒粒度度量 GM 的表示，可结合已有的粒度度量方法进一步研究. 比如，Yao[118] 提出的粒度度量方法，即

$$\mathrm{GM}(\pi)=\sum_{i=1}^{m}\frac{|X_i|}{|U|}\log|X_i|,$$

其中，$\pi=\{X_1, X_2, \cdots, X_m\}$ 是对论域 U 的一个划分，X_i 是 U 的子集. 当粒度最细时，即每个粒为单点集，有 $\mathrm{GM}(\pi)=0$；当粒度最粗时，即整个论域为一个粒，$\mathrm{GM}(\pi)=\log|U|$. 信息粒的粒度度量有助于在问题求解过程中寻找合适的粒层，即粒度空间的优化.

关于信息粒的价值度量 VM，主要从粒度度量、不确定性和领域知识 3 个方面确定：① 信息粒的粒度度量与数据分析需求适应度越高，价值越大；粒度过粗或者过细的信息粒，其价值都会降低；② 可以采用信息论中信息熵和统计学中方差分析方法确定信息粒的价值度量；③ 允许通过领域知识和专家经验指定特定信息粒的价值度量.

(2) 粒层的表示

粒层(Layer)由基于某种粒化准则得到的所有信息粒以及信息粒之间的关系构成. 粒层可以形式化表示为一个二元组，即 Layer=(IGS, Intra-LR). 其中，IGS 表示粒层中信息粒 IG 的集合(Information Granule Set，IGS)，IGS 可表示为 $\mathrm{IGS}=\{\mathrm{IG}_1, \mathrm{IG}_2, \cdots, \mathrm{IG}_M\}$；Intra-LR(Intra-Layer Relationships，Intra-LR)表示粒层中信息粒之间可能存在的关系，如果信息粒 IG_p 与 IG_q 存在关系，那么，Intra-LR 可表示为 $\text{Intra-LR}=\{E \mid E=(\mathrm{IG}_p, \mathrm{IG}_q),\ \mathrm{IG}_p, \mathrm{IG}_q \in \mathrm{IGS}\}$.

(3) 粒结构的表示

MGrIKR 中的粒结构是由不同粒化准则得到的多个粒层、不同粒层中信息粒之间的相互关系以

① http://www.mckinsey.com/insights/business_technology/making_data_analytics_work

及同一粒层中信息粒之间的相互关系构成的拓扑结构.因此,粒结构的形式化表示类似于信息粒 IG 和粒层 Layer,也可用元组形式表示粒结构(Granular Structure,GS),即

$$GS=(LS,\text{Inter-LR}).$$

其中,$LS=\{\text{Layer}_1,\cdots,\text{Layer}_{m-1},\text{Layer}_m\}$ 表示 m 个粒层集合(Layer Set,LS),其中粒层 Layer_j 是粒结构中的一个粒层.Inter-LR(Inter-Layer Relationships,Inter-LR)表示某两粒层 Layer_j 与 Layer_k 的信息粒之间的转换关系集,Inter-LR 可表示为

$$\text{Inter-LR}=\{r^{\prec}\mid r^{\prec}(\text{Layer}_j,\text{Layer}_k)\},$$

或者

$$\text{Inter-LR}=\{r^{\prec}\mid r^{\prec}(\text{IG}^j,\text{IG}^k),\text{IG}^j\in\text{IGS}_j,\text{IG}^k\in\text{IGS}_k\}.$$

这里,$r^{\prec}$ 表示粒层 Layer_j 与 Layer_k 中信息粒之间满足的偏序关系,$j,k=1,\cdots,m$.其中,$r^{\prec}$ 可以是相邻两粒层中信息粒之间的关系,也可是跨层的信息粒之间的关系.

大数据的粒化就是参照信息粒、粒层和粒结构的形式化表示,计算它们各自元组中的每个元素.

4.3.3 并行化/增量式粒结构更新与问题求解

大数据的"高速性"特征要求对其分析的速度要快,采取的响应动作要及时.目前可用的技术方案主要有并行化计算和增量式更新.其中并行计算包括使用分布式并行计算平台、并行使用多核 CPU 的多个计算单元以及使用 GPU 进行协同计算等.增量式更新则借鉴前文 2.3.2 小节提到的各研究成果,当大规模数据集中少部分数据发生变动时,使用增量更新的思想维护整个 MGrIKR 以及修正基于 MGrIKR 问题求解到的结果.

在保障大数据分析的及时性上,从信息粒更新的及时性和问题求解的及时性两个方面进行分析.

(1)信息粒更新的及时性——多源异构粒结构的动态更新

不失一般性,此处我们考虑复杂情况下(多源异构动态数据流)粒结构的动态更新,其余的简单情形类似可得.首先对各个数据源分别建立初始粒结构,然后将各个初始粒结构依照一定的关系进行整合,最终形成一个全局粒结构.

第一步,整合粒结构.首先形式化描述整合两个粒结构 $GS_i=(LS_i,\text{Inter-LR}_i)$ 和 $GS_j=(LS_j,\text{Inter-LR}_j)$.可以定义一个逻辑运算,即定义一个二元映射 $f:GS\times GS\rightarrow GS$,其中 GS 是整个问题论域,即粒结构的集合,这个二元映射应该满足运算规则:

$$f(GS_i,GS_j)=(f_1(LS_i,LS_j),f_2(\text{Inter-LR}_i,\text{Inter-LR}_j)).$$

其中,二元映射 f_1 将两个粒结构的层次进行整合,形成一个新的全局粒层;二元映射 f_2 将两个粒结构进行重新整合,在粒层与粒层之间信息粒的关系集的整合过程中,需要对不同粒层之间和同一粒层之间信息粒转换的关系集进行整合,包括关系的合并、删除与更新.

第二步,各个分量粒结构的动态更新.粒结构的动态更新可以形式化为:$UpdateGS(GS_i)=(UpdateL(LS_i),UpdateR(\text{Inter-LR}_i))$.其中,$UpdateL$ 是粒层的动态更新方法,$UpdateR$ 是层与层中信息粒关系集的动态更新方法.

第三步,全局粒结构的增量更新.通过每个数据源的动态更新结果设计全局粒结构的更新方法,形式化表示为

$$Update(\text{global }GS)=Update(UpdateGS(GS_1),UpdateGS(GS_2),\cdots,UpdateGS(GS_n)).$$

多源异构粒结构的实时动态更新机制如图 6 所示.

(2)问题求解的及时性——基于 MGrIKR 求解的适用问题类型分析

由于粒计算本身具有"非精确"的性质,因此它并不能满足所有类型的大数据处理需求.针对合适的问题类型,基于粒结构的计算可以加速求解过程,保证及时性.确定哪些类型的大数据问题适宜采用粒计算方法非常重要.此处,我们暂提出两类问题作为例子.更多类型的问题可以在进一步的研究工作中发现.

例 1. 粒度空间寻优问题.采用优化理论描述粒层选择问题,确定有效解的计算粒度,从而在最短时间内获得有效解.

定义 1. 解的有效性可以通过一个二元组来定义 $SolutionEffectiveness=(\text{GM}(R),T_u)$.$R$ 是计算的结果,$\text{GM}(R)$是该结果的粒度度量,T_u是时限需求.如果一个解的 $\text{GM}\leq\text{GM}(R)$,并且获得这个解的时间小于 T_u,则这个解具有有效性,称为有效解.

为了从面向领域的粒结构中选择一个"合适"的粒层进行计算,以降低计算的实际时空耗费,需进行粒度空间寻优.粒度空间寻优就是在粒结构的 m 层中找到这样的粒层 Layer_i:

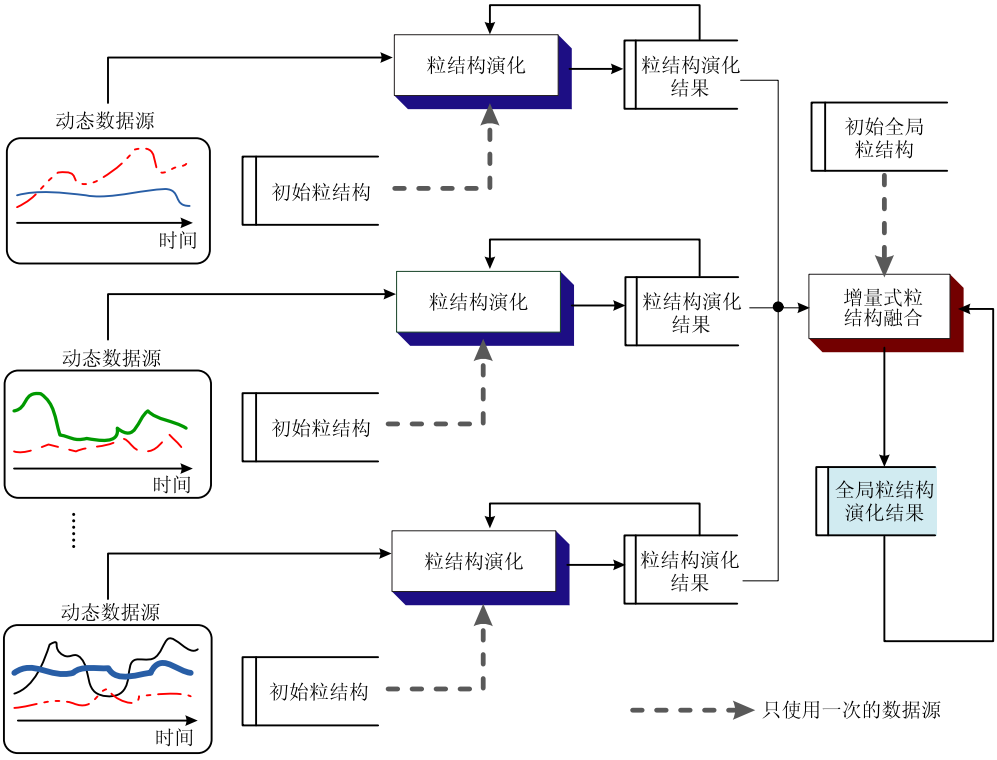


图 6　粒结构动态实时动态更新

$$\text{Max GM}(\text{Layer}_i)$$

$$\text{s. t.}$$

$$\text{GM}(R_i := \text{Solve}(\text{Layer}_i)) \leq \text{GM}(R_u),$$

$$T_i \leq T_u, 1 \leq i \leq m.$$

其中，R_i、T_i分别是在第 i 层上求解的结果和所花费时间. 考虑如图 7 所示的问题.

Layer_3上的解粒度满足需求，但时间不能满足时限约束；Layer_1上的解时间可以满足时限约束，但粒度又太粗. 这两个粒层上的解都不是有效解. 在Layer_2上的解同时满足了解的粒度要求和时限约束，是有效解.

例 2.　人机协同可渐进计算问题. 在由人和计算系统构成的决策系统中，如果由决策指导的行动具有可拆分性、决策的计算可以逐步细化，且从当前状态出发，更细化的解可用于指导下一步的行动，针对这一类型的问题，可构建“抢先行动”和“边计算边行动”的人机协同决策响应模型.

在相邻粒层，下层解是上层解的细化，记作 $R_{i-1} \prec R_i$，每个解对应用户下一步采取的行动步骤(ActionStep，AS)，记作 $R_i \rightarrow AS_i$，并且整个决策对应的行动 A 具有可拆分性，记作

$$A = \sum_i AS_i,\ i = 1,2,\cdots,n.$$

根据行动的步骤数，确定 n 的取值，也就是确定了求解的阶段和并行粒度，然后从预先建立的面向领域的粒结构中筛选出合适的 n 个粒层，人机协同可渐进求解问题模型如图 8 所示.

如果不采用人机协同渐进计算方式，那么 Action-Step_1只能从 t_3时间点开始执行，整个决策和行动的最后完成时间会大幅度延后.

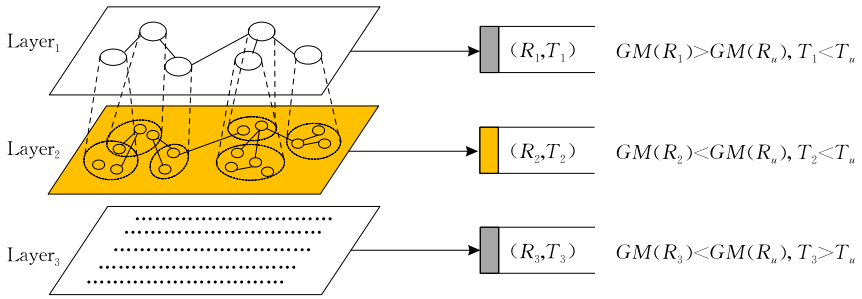


图 7　选择合适的粒层，满足粒度度量需求和时限约束

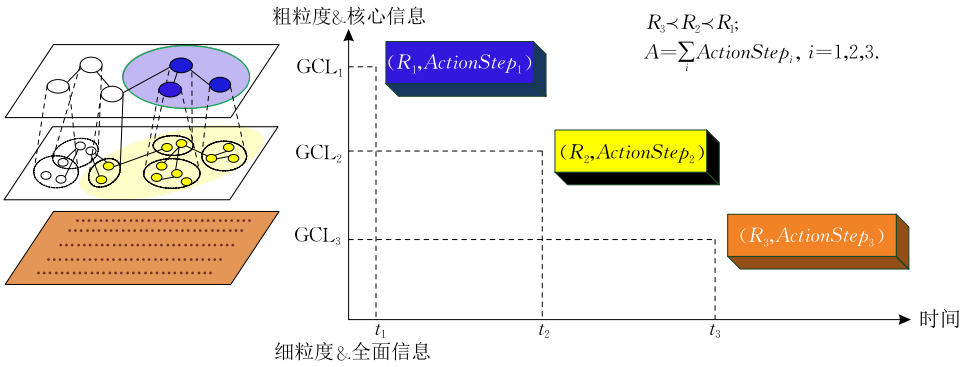


图 8 人机协同决策响应模型

4.4 粒计算应用于大数据处理的其他问题

在图 5 所示的大数据粒计算处理框架之外，还有一些问题需要考虑. 这些问题在基于粒计算的大数据分析实际工作中，会直接影响到数据分析和决策行为的效果.

(1) 研究问题领域粒度层次的逻辑设计

对大数据所表示的领域信息进行粒度分析，确定可能的粒度层次数目、各层次上信息粒的语义以及根据领域知识能够断言的信息粒之间的相关关系，这些是进一步构建和分析 MGrIKR 的基础. 这一粒度分析阶段得到的结果及其质量直接影响后续的大数据处理的准确性和效率. 此项研究工作含有较多的经验成分，可以借鉴数据库系统理论中模式设计和软件工程中需求分析的理论和方法，制定问题领域粒度层次的逻辑设计规范，以确保粒度层次数目和信息粒语义等关键信息设置的科学性.

(2) 多粒度大数据分析的实现技术

多粒度大数据分析理论模型的科学性和有效性最终要通过实验来进行检验. 由于目前大数据开源平台的蓬勃发展，适用于不同应用场合的系统层出不穷，为实验提供了很多选择. 如何针对具体数据挑选最适合于多粒度建模与计算的实验平台，并通过各种参数和配置的调优，实现对特定粒计算模型的最佳支持，是成功开展实验的关键.

(3) 计算结果验证、评价和应用

粒计算是基于非精确数据表达的计算，在实验阶段需要验证模型的正确性. 常见的一种方法是使用精确计算实现同样的分析需求，从而可以验证产生自粒计算的非精确解的正确性. 除此之外，我们还可以考虑使用其他途径，如“形式化描述＋数学证明”，或者借助于自动推理.

采用 MGrIKR 进行大数据分析是因为它具备如前文所述的一些优点(较高的时空效率、及时提供有效解以及便于人工协同等)，如何设计对比实验，如何设置定性和(或)定量指标对这些性能进行评价，也是值得考虑的问题.

大数据分析的结果需要应用到具体的决策支持中才能体现其最终价值. 这一环节的关键在于确认分析结果的语义解释及该结果的使用方式，如指导人的行动或者是作为其他软件系统的输入.

5 研究展望

归纳关于将粒计算应用到大数据处理的可能性和模型框架等问题，未来的研究工作可以从以下方面展开：

(1) 研究大数据的粒化. 重点针对大数据的“高速性”和“巨量性”，继续对粒计算基本模型和算法开展理论研究，获取更加快速的粒化方法. 如前文表 2 的分析，模糊集、粗糙集、商空间和云模型等各个具体粒化模型的代表性实现方法具有不同的时间复杂度. 在降低粒计算模型的时间复杂度方面，已经有了一些研究成果[106,119]. 加快知识获取速度的一种常用方法是增量式更新，近年来在粗糙集的增量式更新方面已经有了一些很好的成果[120-122].

今后还需要继续研究满足大数据处理“高速性”和“巨量性”需求的改进模型及其实现算法，例如研究更加接近线性时间复杂度的粗糙集约简算法；继续研究其他各个粒计算模型的增量式更新知识获取方法，尤其是动态多数据源的粒结构增量更新方法.

(2) 研究粒度空间优化、粒度层次切换和多粒度联合计算这 3 种粒计算模式在大数据环境下的应用. 例如，除了 4.3.3 节提到的两个例子中介绍的问题类型，还可以将例 2 的人机协同决策响应模型转化成另外一种问题求解模型，即精度渐进求解模型：从给出最粗粒度解开始，朝着更细粒度层次的方向递阶进行计算，用户在任何时刻都可以获取到一个当前可得的最细粒度解. 这一计算模型的意义是在保证及时性的前提下，获得有实用价值的非精确解.

(3)研究和验证大数据环境下粒计算处理框架(如图 5 所示)的指导作用.该处理框架考虑了在大数据处理的各个环节中如何使用粒计算思想,同时兼顾粒计算具体模型和数据挖掘/机器学习算法的应用.可结合具体的领域背景和数据分析需求,将其用于指导大数据分析的研究和实践,并且根据指导的过程中发现的新问题使之得到修正和完善.

(4)研究粒计算处理大数据问题的并行实现方法.紧密结合快速发展的 IT 基础设施和软件平台,开发并行计算在大数据的粒计算方法分析中的加速作用.对于数据可并行的计算密集型任务,研究粒计算的 GPU+CPU 高性能计算集群解决方案;对于数据量巨大且数据整体关联性较强、并行性较弱的问题,研究 Hadoop、Spark/Storm 等开源平台上的处理方法.

(5)结合具体应用背景,在科学研究和工程应用中使用基于粒计算的大数据处理方法.例如在音乐大数据的哼唱检索中,将每首歌曲以乐句或乐段为信息粒表达为粒结构,再进行检索;在生态环境监测和预警中,针对水质的连续性物理化学过程,在时间、空间以及水质属性的取值等多个维度上应用多粒度建模,高效计算下一时段指定位置的水质指标;在大规模视频监控系统中,将监控录像按照场景分类信息进行粒化后,组织成具有场景语义的粒结构,从而实现监控视频的压缩存储和高效检索.这些具体的研究工作将会不断丰富基于粒计算的大数据处理这一方向的理论模型和技术手段.

6　总　结

大数据是当前信息技术研究的一个热点课题.成功应对大数据带来的挑战、充分挖掘出其作为一种资源的价值,需要多个领域的研究人员与从业人员共同努力.本文分析了使用粒计算处理大数据的可能性,提出一种基于粒计算的大数据处理框架,并且综述了大数据和粒计算领域的相关研究基础.未来需要开展的工作主要是结合具体的应用领域和分析需求,参照本文基于粒计算的大数据处理框架,研究大数据的 MGrIKR 构建、演化以及基于大数据 MGrIKR 的智能计算方法.

粒计算作为一种计算范式已经在智能信息处理领域发挥了重要的作用,但将其应用于大数据分析还处于起步阶段.本文着重梳理了粒计算对于大数据处理的指导作用,希望对关注粒计算应用和大数据处理技术的研究者能够提供一些有益的借鉴和帮助.

致　谢　评审专家提供了宝贵意见和建议,许昌林博士和马希骜博士参加了本文的讨论,在些表示感谢!

参　考　文　献

[1] Qin Xiong-Pai, Wang Hui-Ju, Du Xiao-Yong, et al. Big data analysis—Competition and symbiosis of RDBMS and MapReduce. Journal of Software, 2012, 23(1): 32-45(in Chinese)
(覃雄派, 王会举, 杜小勇等. 大数据分析——RDBMS 与 MapReduce 的竞争与共生. 软件学报, 2012, 23(1): 32-45)

[2] Wang S L, Ding G Y, Zhong M. Big spatial data mining//Proceedings of the 2013 IEEE International Conference on Big Data. Santa Clara, CA, USA, 2013: 13-21

[3] Bill B, et al. Big data: The next Google. Interview by duncan graham-rowe. Nature, 2008, 455(7209): 8-9

[4] Wu X D, Zhu X Q, Wu G Q, Ding W. Data mining with big data. IEEE Transactions on Knowledge and Data Engineering, 2014, 26(1): 97-107

[5] Labrinidis A, Jagadish H V. Challenges and opportunities with big data//Proceedings of the VLDB 2012. Istanbul, Turkey, 2012: 2032-2033

[6] Ma Shuai, Li Jian-Xin, Hu Chun-Ming. Challenges and thought of big data science and engineering. Communications of China Computer Federation, 2012, 9(8): 22-30 (in Chinese)
(马帅, 李建欣, 胡春明. 大数据科学与工程的挑战与思考. 中国计算机学会通讯, 2012, 9(8): 22-30)

[7] Manyika J, et al. Big data: The next frontier for innovation, competition, and productivity. San Francisco, CA, USA: McKinsey Global Institute, Technical Report: May 2011

[8] Bizer C, et al. The meaningful use of Big data: Four perspectives-four challenges//Proceedings of SIGMOD 2011. Athens, Greece, 2011, 40(4): 56-60

[9] Fisher D, DeLine R, Czerwinski M, et al. Interactions with big data analytics. Interactions, 2012, 19(3): 50-59

[10] Chen C L P, Zhang C Y. Data-intensive applications, challenges, techniques and technologies: A survey on big data. Information Sciences, 2014, 275: 314-347

[11] Havens T C, Bezdek J C, Leckie C, Hall L O, Palaniswami M. Fuzzy c-means algorithms for very large data. IEEE Transactions on Fuzzy Systems, 2012, 20(6): 1130-1146

[12] Lu J G, Li D D. Bias correction in a small sample from big data. IEEE Transactions on Knowledge and Data Engineering, 2013, 25(11): 2658-2663

[13] Sun W Q, et al. Store, schedule and switch-A new data delivery model in the big data era//Proceedings of the 15th International Conference on Transparent Optical Networks (ICTON). Cartagena, Spain, 2013: 1-4

[14] Han X X, Li J Z, Yang D H, et al. Efficient skyline computation on big data. IEEE Transactions on Knowledge and Data Engineering, 2013, 25(11): 2521-2535

[15] Zhang J B, Wong J S, Li T R, Pan Y. A comparison of parallel large-scale knowledge acquisition using rough set

theory on different MapReduce runtime systems. International Journal of Approximate Reasoning, 2014, 55(3): 896-907

[16] Guyon I, Elisseeff A. An introduction to variable and feature selection. The Journal of Machine Learning Research, 2003, 3: 1157-1182

[17] Yan S, Xu D, Zhang B, et al. Graph embedding and extensions: A general framework for dimensionality reduction. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2007, 29(1): 40-51

[18] Nguyen H S, Nguyen S H. Discretization methods in data mining. Rough Sets in Knowledge Discovery, 1998, 1: 451-482

[19] Parsons L, Haque E, Liu H. Subspace clustering for high dimensional data: A review. ACM SIGKDD Explorations Newsletter, 2004, 6(1): 90-105

[20] Zhang Hao-Ran, Wang Xiao-Dong. Incremental and online learning algorithm for regression least squares support vector machine. Chinese Journal of Computers, 2006, 29(3): 400-406(in Chinese)
(张浩然，汪晓东. 回归最小二乘支持向量机的增量和在线式学习算法. 计算机学报，2006，29(3)：400-406)

[21] Lughofer E. On-line incremental feature weighting in evolving fuzzy classifiers. Fuzzy Sets and Systems, 2011, 163(1): 1-23

[22] Angelov P, Lughofer E, Zhou X W. Evolving fuzzy classifiers using different model architectures. Fuzzy Sets and Systems, 2008, 159(23): 3160-3182

[23] Lughofer E, Angelov P, Zhou X W. Evolving single- and multi-model fuzzy classifiers with FLEXFIS-class//Proceedings of the 2007 IEEE International Conference on Fuzzy Systems. London, UK, 2007: 1-6

[24] Zhang W, et al. Malicious Web page detection based on on-line learning algorithm//Proceedings of the 2011 International Conference on Machine Learning and Cybernetics. Guilin, 2011: 1914-1919

[25] Hu F, Wang G Y, Huang H, Wu Y. Incremental attribute reduction based on elementary sets//Proceedings of the RSFDGrC2005, Part 1, LNAI3641. Regina, Canada, 2005: 185-193

[26] Zheng Z, Wang G Y. RRIA: A rough set and rule tree based incremental knowledge acquisition algorithm. Fundamenta Informaticae, 2004, 59(2): 299-313

[27] Li S Y, Li T R, Liu D. Incremental updating approximations in dominance-based rough sets approach under the variation of the attribute set. Knowledge-Based Systems, 2013, 40: 17-26

[28] Mohamed M G, Zaslavsky A, Krishnaswamy S. Mining data streams: A review. ACM Sigmod Record, 2005, 34(2): 18-26

[29] Li H F, Lee S Y, Shan M K. An efficient algorithm for mining frequent itemsets over the entire history of data streams//Proceedings of the 1st International Workshop on Knowledge Discovery in Data Streams. Pisa, Italy, 2004: 1-10

[30] Li H F, Lee S Y. Mining frequent itemsets over data streams using efficient window sliding techniques. Expert Systems with Applications, 2009, 36(2): 1466-1477

[31] Li P P, Wu X D, Hu X G. Mining recurring concept drifts with limited labeled streaming data. ACM Transactions on Intelligent Systems and Technology (TIST), 2012, 3(2): 29

[32] Gomes J B, Gaber M M, Menasalvas E, Sousa P. Mining Recurring Concepts in a Dynamic Feature Space. IEEE Transactions on Neural Networks and Learning Systems, 2014, 25(1): 95-110

[33] Reddy Dr T V, Govardhan A, Sidda V. Benchmark algorithms and models of frequent itemset mining over data streams: Contemporary affirmation of state of art. Global Journal of Computer Science and Technology, 2013, 13(5): 1-9

[34] Zikopoulos P C, Eaton C, DeRoos D, et al. Understanding Big Data. New York: McGraw-Hill, 2012

[35] Panos L, Christof E. Embedded analytics and statistics for big data. IEEE Software, 2013, 30(6): 33-39

[36] Marx V. Biology: The big challenges of big data. Nature, 2013, 498(7453): 255-260

[37] Turk-Browne N B. Functional Interactions as big data in the human brain. Science, 2013, 342(6158): 580-584

[38] Bedini I, Elser B, Velegrakis Y. The trento big data platform for public administration and large companies: Use cases and opportunities//Proceedings of the VLDB 2013. Riva del Garda, Trento, Italy, 2013: 1166-1167

[39] Le Q V, Ranzato M A, Monga R, et al. Building high-level features using large scale unsupervised learning//Proceedings of the 2013 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Vancouver, Canada, 2013: 8595-8598

[40] Kim T K, et al. On-line learning of mutually orthogonal subspaces for face recognition by image sets. IEEE Transactions on Image Processing, 2010, 19(4): 1067-1074

[41] Wu X D, et al. Online feature selection with streaming features. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2013, 35(5): 1178-1192

[42] Chen H, Chiang R H L, Storey V C. Business intelligence and analytics: From big data to big impact. MIS Quarterly, 2012, 36(4): 1165-1188

[43] Zadeh L A. Toward human level machine intelligence—Is it achievable? The need for a paradigm shift. IEEE Computational Intelligence Magazine, 2008, 3(3): 11-22

[44] Chen L. Topological structure in visual perception. Science, 1982, 218(12): 699

[45] Chen Lin. Neural correlation of "global-first" object formation: Anterior temporal lobe. Chinese Bulletin of Life Sciences, 2008, 20(5): 718-721(in Chinese)
(陈霖. "大范围优先"对象形成的神经关联：前颞叶. 生命科学，2008，20(5)：718-721)

[46] Pei J. Some new progress in analyzing and mining uncertain and probabilistic data for big data analytics//Proceedings of the Rough Sets, Fuzzy Sets, Data Mining, and Granular Computing. Berlin Heidelberg: Springer, 2013: 38-45

[47] Lee K H, et al. Parallel data processing with MapReduce: A survey. ACM SIGMOD Record, 2012, 40(4): 11-20

[48] Tan W, et al. Social-network-sourced big data analytics. IEEE Internet Computing, 2013, 17(5): 62-69

[49] Nakatsuji M, Fujiwara Y. Linked taxonomies to capture users' subjective assessments of items to facilitate accurate collaborative filtering. Artificial Intelligence, 2014, 207: 52-68

[50] Yao Y Y. Granular Computing: Basic issues and possible solutions//Proceedings of the 5th Joint Conference on Information Sciences, Vol. I. Atlantic, NJ: Association for Intelligent Machinery, 2000: 186-189

[51] Miao Duo-Qian, Li De-Yi, Yao Yi-Yu, et al. Uncertainty and Granular Computing. Beijing: Science Press, 2011 (in Chinese)
(苗夺谦，李德毅，姚一豫等. 不确定性与粒计算. 北京：科学出版社，2011)

[52] Yao J T, Vasilakos A V, Pedrycz W. Granular Computing: Perspectives and Challenges. IEEE Transactions on Cybernetics, 2013, 43(6): 1977-1989

[53] Li Chao, Zhang Ming-Bo, Xing Chun-Xiao Hu Jin-Song. Survey and review on key technologies of column oriented data base systems. Computer Science, 2010, 37(12): 1-7(in Chinese)
(李超，张明博，邢春晓，胡劲松. 列存储数据库关键技术综述. 计算机科学，2010，37(12)：1-7)

[54] Qian Yu-Hua. Granulation Mechanism and Data Modeling for Complex Data [Ph. D. dissertation]. School of Computer and Information Technology, Shanxi University, Taiyuan, 2011(in Chinese)
(钱宇华. 复杂数据的粒化机理和数据建模[博士学位论文]. 山西大学计算机与信息技术学院，太原，2011)

[55] Skowron A, Wasilewski P. Information systems in modeling interactive computations on granules. Theoretical Computer Science, 2011, 412(42): 5939-5959

[56] Panoutsos G, Mahfouf M. A neural-fuzzy modeling framework based on granular computing: Concepts and applications. Fuzzy Sets and Systems, 2010, 161: 2808-2830

[57] Zadeh L A. Toward a theory of fuzzy information granulation and its centrality in human reasoning and fuzzy logic. Fuzzy Sets and Systems, 1997, 90: 111-127

[58] Yao Y Y. Information granulation and rough set approximation. International Journal of Intelligent Systems, 2001, 16(1): 87-104

[59] Zhang Yan-Ping, Zhang Ling, Wu Tao. The representation of different granular worlds: A quotient space. Chinese Journal of Computers, 2004, 27(3): 328-333(in Chinese)
(张燕平，张铃，吴涛. 不同粒度世界的描述法——商空间法. 计算机学报，2004，27(3)：328-333)

[60] Ma Hong-Yao, Wang Guo-Yin, Zhang Qing-Hua, et al. Multi-granularity color image segmentation based on cloud model. Computer Engineering, 2012, 38(20): 184-187(in Chinese)
(马鸿耀，王国胤，张清华等. 基于云模型的多粒度彩色图像分割. 计算机工程，2012，38(20)：184-187)

[61] Qin Kun, Li De-Yi, Xu Kai. Image segmentation based on cloud model. Journal of Geomatics, 2006, 31(5): 3-5(in Chinese)
(秦昆，李德毅，许凯. 基于云模型的图像分割方法研究. 测绘信息与工程，2006，31(5)：3-5)

[62] Liu Chang-Yu, Feng Mang, Dai Xiao-Jun, et al. A new algorithm of backward cloud. Journal of System Simulation, 2004, 16(11): 2417-2420(in Chinese)
(刘常昱，冯芒，戴晓军等. 基于云 X 信息的逆向云新算法. 系统仿真学报，2004，16(11)：2417-2420)

[63] Zhang Qing-Hua. Research on Hierarchy Granular Computing Theory and Its Application [Ph. D. dissertation]. School of Information Science and Technology, Southwest Jiaotong University, Chengdu, 2009(in Chinese)
(张清华. 分层递阶粒计算理论及其应用研究[博士学位论文]. 西南交通大学信息科学与技术学院，成都，2009)

[64] Zhang Ling, Zhang Bo. Theory of fuzzy quotient space (methods of fuzzy granular computing). Journal of Software, 2003, 14(4): 770-776(in Chinese)
(张铃，张钹. 模糊商空间理论(模糊粒度计算方法). 软件学报，2003，14(4)：770-776)

[65] Pedrycz W. Knowledge-Based Clustering: From Data to Information Granules. Hoboken, New Jersey: John Wiley & Sons, 2005

[66] Lin F, Cohen W W. Power iteration clustering//Proceedings of 2010 International Conference on Machine Learning (ICML). Haifa, Israel, 2010: 655-662

[67] Yan W, Brahmakshatriya U, Xue Y, et al. p-PIC: Parallel power iteration clustering for big data. Journal of Parallel and Distributed Computing, 2013, 73(3): 352-359

[68] Wang Guo-Yin, Li De-Yi, Yao Yi-Yu, et al. Cloud Model and Granular Computing. Beijing: Science Press, 2012 (in Chinese)
(王国胤，李德毅，姚一豫等. 云模型与粒计算. 北京：科学出版社，2012)

[69] Du Wei-Lin, Miao Duo-Qian, Li Dao-Guo, et al. Analysis on relationship between concept lattice and granule partition lattice. Computer Science, 2005, 32(1): 182-187 (in Chinese)
(杜伟林，苗夺谦，李道国等. 概念格与粒度划分的相关性分析. 计算机科学，2005，32(1)：182-187)

[70] Hu Qing-Hua, Yu Da-Ren, Xie Zong-Xia. Numerical attribute reduction based on neighborhood granulation and rough approximation. Journal of Software, 2008, 19(3): 640-649(in Chinese)
(胡清华，于达仁，谢宗霞. 基于邻域粒化和粗糙逼近的数值属性约简. 软件学报，2008，19(3)：640-649)

[71] Pedrycz W. Granular computing—The emerging paradigm. Journal of Uncertain Systems, 2007, 1(1): 38-61

[72] Liu H B, et al. Hyperspherical granular computing classification algorithm based on fuzzy lattices. Mathematical and Computer Modelling, 2013, 57(3-4): 661-670

[73] Yao Y Y, Deng X F. A granular computing paradigm for concept learning//Proceedings of the Emerging Paradigms in Machine Learning. Berlin Heidelberg: Springer, 2013

[74] Zhang X, Miao D. Two basic double-quantitative rough set models of precision and grade and their investigation using granular computing. International Journal of Approximate Reasoning, 2013, 54(8): 1130-1148

[75] Pal S K, Meher S K, Dutta S. Class-dependent rough-fuzzy granular space, dispersion index and classification. Pattern Recognition, 2012, 45(7): 2690-2707

[76] Pedrycz A, Hirota K, Pedrycz W, et al. Granular representation and granular computing with fuzzy sets. Fuzzy Sets and Systems, 2012, 203: 17-32

[77] Peters G. Granular box regression. IEEE Transactions on Fuzzy Systems, 2011, 19(6): 1141-1152

[78] Solis A R, Panoutsos G. Granular computing neural-fuzzy modelling: A neutrosophic approach. Applied Soft Computing, 2013, 13(9): 4010-4021

[79] Niu W J, et al. Multi-granularity context model for dynamic Web service composition. Journal of Network and Computer Applications, 2011, 34: 312-326

[80] Meher S K, Pal S K. Rough-wavelet granular space and classification of multispectral remote sensing image. Applied Soft Computing, 2011, 11(8): 5662-5673

[81] Gacek A. Granular modelling of signals: A framework of granular computing. Information Sciences, 2013, 221: 1-11

[82] Tang X Q, Zhu P. Hierarchical clustering problems and analysis of fuzzy proximity relation on granular space. IEEE Transactions on Fuzzy Systems, 2013, 21(5): 814-824

[83] Nandedkar A V, Biswas P K. A granular reflex fuzzy min-max neural network for classification. IEEE Transactions on Neural Networks, 2009, 20(7): 1117-1134

[84] Pedrycz W. Allocation of information granularity in optimization and decision-making models: Towards building the foundations of granular computing. European Journal of Operational Research, 2014, 232(1): 137-145

[85] Pedrycz W, Homenda W. Building the fundamentals of granular computing: A principle of justifiable granularity. Applied Soft Computing, 2013, 13(10): 4209-4218

[86] Zhang L, Zhang B. The quotient space theory of problem solving. Fundamenta Informaticae, 2004, 59(2): 287-29

[87] Zhu P, Hu Q. Adaptive neighborhood granularity selection and combination based on margin distribution optimization. Information Sciences, 2013, 249: 1-12

[88] Hobbs J R. Granularity//Proceedings of the 9th International Joint Conference on Artificial Intelligence. Los Angeles, USA, 1985: 432-435

[89] McCalla G, Greer J, Barrie B, et al. Granularity hierarchies. Computers & Mathematics with Applications, 1992, 23(2): 363-375

[90] Wang Guo-Yin, Zhang Qing-Hua. Uncertainty of rough sets in different knowledge granularities. Chinese Journal of Computers, 2008, 31(9): 1588-1598(in Chinese)
(王国胤，张清华. 不同知识粒度下粗糙集的不确定性研究. 计算机学报，2008, 31(9): 1588-1598)

[91] Wang Guo-Yin, Zhang Qing-Hua, Ma Xi-Ao, et al. Granular computing models for knowledge uncertainty. Journal of Software, 2011, 22(4): 676-694(in Chinese)
(王国胤，张清华，马希骜等. 知识不确定性问题的粒计算模型. 软件学报，2011, 22(4): 676-694)

[92] Zhang Qing-Hua, Wang Guo-Yin, Xiao Yu. Approximation sets of rough sets. Journal of Software, 2012, 23(7): 1745-1759(in Chinese)
(张清华，王国胤，肖雨. 粗糙集的近似集. 软件学报，2012, 23(7): 1745-1759)

[93] Jang J S R. ANFIS: Adaptive-network-based fuzzy inference system. IEEE Transactions on Systems, Man and Cybernetics, 1993, 23(3): 665-685

[94] Wang G Y, Shi H B. TMLNN: Triple-valued or multiple-valued logic neural network. IEEE Transactions on Neural Networks, 1998, 9(6): 1099-1117

[95] Hinton G E, Salakhutdinov R R. Reducing the dimensionality of data with neural networks. Science, 2006, 313: 504-507

[96] Coates A, Ng A Y, Lee H. An analysis of single-layer networks in unsupervised feature learning//Proceedings of the International Conference on Artificial Intelligence and Statistics. Fort Lauderdale, USA, 2011: 215-223

[97] Weston J, Ratle F, Mobahi H, et al. Deep learning via semi-supervised embedding//Montavon G, Orr G B, Müller K-R eds. Neural Networks: Tricks of the Trade. Berlin Heidelberg: Springer, 2012: 639-655

[98] Bengio Y. Learning deep architectures for AI. Foundations and Trends in Machine Learning, 2009, 2(1): 1-127

[99] Somchaya L, Veloso M. Weighted synergy graphs for effective team formation with heterogeneous ad hoc agents. Artificial Intelligence, 2014, 208: 41-65

[100] Pedrycz W. Granular computing as a framework of system modeling. Journal of Control, Automation and Electrical Systems, 2013, 24(1-2): 81-86

[101] Slezak D, Synak P, Wróblewski J, et al. Infobright analytic database engine using rough sets and granular computing//Proceedings of the 2010 IEEE International Conference on Granular Computing (GrC). Silicon Valley, USA, 2010: 432-437

[102] Wang C Z, Chen D G, Wu C, Hu Q H. Data compression with homomorphism in covering information systems. International Journal of Approximate Reasoning, 2011, 52(4): 519-525

[103] Panoutsos G, Mahfouf M. Modelling imprecise and scattered multidimensional data using granular data compression and multiple granularity modeling//Proceedings of the 2008 IEEE International Conference on Granular Computing. Hangzhou, China, 2008: 512-517

[104] Yang C, Zhang X, Zhong C, et al. A spatiotemporal compression based approach for efficient big data processing on cloud. Journal of Computer and System Sciences, 2014, 80(8): 1563-1583

[105] Ruan J H, Wang X P, Shia Y. Developing fast predictors for large-scale time series using fuzzy granular support vector machines. Applied Soft Computing, 2013, 13(9): 3981-4000

[106] Hu Feng, Wang Guo-Yin. Quick reduction algorithm based on attribute order. Chinese Journal of Computers, 2007, 30(8): 1429-1435(in Chinese)
(胡峰，王国胤. 属性序下的快速约简算法. 计算机学报，2007，30(8)：1429-1435)

[107] Zhang J B, Li T R, Ruan D, et al. A parallel method for computing rough set approximations. Information Sciences, 2012, 194: 209-223

[108] Liu Y C, Li D Y, He W, Wang G Y. Granular computing based on Gaussian cloud transformation. Fundamenta Informaticae, 2013, 127: 385-398

[109] Pal S K. Soft data mining, computational theory of perceptions, and rough-fuzzy approach. Information Sciences, 2004, 163(1-3): 5-12

[110] Halevy A, Rajaraman A, Ordille J. Data integration: The teenage years//Proceedings of the VLDB2006. Seoul, Korea, 2006: 12-15

[111] Kanehisa M, et al. KEGG for integration and interpretation of large-scale molecular data sets. Nucleic Acids Research, 2012, 40(Database Issue): D109-D114

[112] Dong X L, Halevy A, Yu C. Data integration with uncertainty. The VLDB Journal, 2009(18): 469-500

[113] Achard F, Vaysseix G, Barillot E. XML, bioinformatics and data integration. Bioinformatics Review, 2001, 17(2): 115-125

[114] Shawn R J. Arnold: Declarative crowd-machine data integration//Proceedings of the 6th Biennial Conference on Innovative Data Systems Research (CIDR'13). Asilomar, California, USA, 2013: 6-9

[115] Lenzerini M. Data integration: A theoretical perspective//Proceedings of the ACM Symposium on Principles of Database Systems (PODS). New York, NY, USA, 2002: 233-246

[116] Shvaiko P, Euzenat J. A survey of schema-based matching approaches. Journal on Data Semantics, 2005, IV: 146-171

[117] Friedman M, Levy A, Millstein T. Navigational plans for data integration//Proceedings of the National Conference on Artificial (AAAI). 1999: 67-73

[118] Yao Y Y. Notes on rough set approximations and associated measures. Journal of Zhejiang Ocean University (Natural Science), 2010, 29(5): 399-410

[119] Qian Y H, Wang Q, Cheng H, et al. Fuzzy-rough feature selection accelerator. Fuzzy Sets and Systems, 2015, 258: 61-78

[120] Chen H M, Li T R, Ruan D, et al. A rough-set-based incremental approach for updating approximations under dynamic maintenance environments. IEEE Transactions on Knowledge and Data Engineering, 2013, 25(2): 274-284

[121] Chen H M, Li T R, Luo C, Horng S J, Wang G Y. A rough set-based method for updating decision rules on attribute values' coarsening and refining. IEEE Transactions on Knowledge and Data Engineering, 2014, 26(12): 2886-2899

[122] Liang J Y, Wang F, Dang C Y, et al. A group incremental approach to feature selection applying rough set technique. IEEE Transactions on Knowledge and Data Engineering, 2014, 26(2): 294-308

XU Ji, born in 1979, Ph. D. candidate. His current research interests include data mining, granular computing and software engineering.

WANG Guo-Yin, born in 1970, Ph.D., professor, Ph. D. supervisor. His research interests include rough set, granular computing, cognitive computing, intelligent information processing, data mining, intelligent information security.

YU Hong, born in 1972, Ph.D., professor. Her research interests include rough set, intelligent information processing, web intelligence, data mining.

Background

Big data processing is a hot topic for researchers from IT industry and academia, and is concerned by the governments around the world. Many significant progresses in this field have been done in terms of a few theoretical aspects and application realms, but there are still many challenges in the big data processing. Researchers did not intend to introduce granular computing methodology into big data processing until recently. This paper reviews the researches in big data processing and granular computing firstly, and then analyzes the potential of harnessing big data with granular computing technology. The related key problems in this regard are discussed at last.

Our work is partly supported by the National Natural Science Foundation of China under Grant Nos. 61272060 and 61379114, and Natural Science Foundation Key Project of Chongqing of P. R. China under Grant No. CSTC2013jjB40003. The projects aim to construct the theories and methods of bi-directional cognitive transformation between the intension and extension of concepts, and three-way decision clustering in massive complex data. Some works of the research fields have been published in international journals and the international conferences, such as IJCAI, RSKT, etc.