《Data Mining with Big Data》读书报告

计算机科学与技术 18-2 班 2018211958 孙淼

说在前面

本论文取自 IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERIN(TKDE)，在所有 IEEE 出版的文章中（包括所有年份、所有期刊和会议论文)，这篇大数据文章在 2013 年 12 月到 2015 年 6 月连续 19 个月被下载的次数均排第一。所以本读书报告取本论文作为阅读对象。本文中所有引用原文译文皆用斜体且缩进，方便区分。

报告正文

✧ 我们都知道一篇论文的“眼”就是 Abstract，这篇文章的摘要如下：*“大数据涉及具有多个自主数据源的大容量、复杂、不断增长的数据集。随着网络、数据存储和数据收集能力的快速发展，大数据正在包括物理、生物和生物医学在内的所有科学和工程领域迅速扩展。本文提出了一个* ***HACE 定理****，描述了大数据革命的特点，并从****数据挖掘****的角度提出了大数据****处理模型****。这种数据驱动的模型涉及需求驱动的****信息源聚合****、挖掘和分析、用户兴趣建模以及安全和隐私考虑。我们分析了数据驱动模型和大数据革命中的挑战问题。”*

读了这个摘要，我们可以加粗关键词，可以想到作者应该是通过大数据迅速拓展为引，提出 HACE 定理，并根据这个定理，提出了相应的大数据处理模型。这个模型涉及基于需求驱动的信息源聚合、挖掘和分析、用户兴趣建模以及安全和隐私考虑。

✧ 下面看看 Introduction，这可以帮助我们学习作者是在何种条件下发现社会需求，并在此基础上萌生出创新点的，Introduction 的第一段是在这样的：*“莫言获得 2012 年诺贝尔文学奖。这可能是最具争议的诺贝尔奖。在谷歌上搜索“莫言诺贝尔奖”，截至 2013 年 1 月 3 日，**共有 105 万个链接**。莫言最近说：“对于所有的赞扬和批评，我都很感激。”在莫言 31 年的写作生涯中，他收到了哪些类型的赞扬和批评?随着网络和各种新闻媒体的评论不断涌现，我们能否实时总结不同媒体的各种观点，包括不断更新的、相互参照的评论?**这种类型的摘要程序是大数据处理的一个很好的例子，因为信息来自多个异构的、自治的、复杂的、不断发展的关系，并不断增长。”*

可以看出作者选出了论文发表的 2014 年 1 月（但是实际开始准备肯定要早得多，毕竟这样一篇杰出的论文需要时间）时还比较火热的话题，莫言获得诺贝尔奖，这样一个例子很巧妙，我个人认为有三点：

1）莫言是中国人，诺贝尔奖是外国人设立的奖项，这样一个“中西结合”的例子，作为一个中国人团队发表在外国期刊上的论文，能够使得无论是国人还是洋人都不陌生，是非常合适而巧妙的。

2）莫言获诺贝尔奖，是一个很有争议的例子，中外媒体等都对其有着很大量的评论，作者也指出了“*在谷歌上搜索“莫言诺贝尔奖”，截至 2013 年 1 月 3 日，共有 105 万个链接。*”，这契合了大数据的“大”这一条件，提供的足够量的数据集，方便后期的工作进行。

3）最后一点就是作者指出的“*因为信息来自多个异构的、自治的、*

复杂的、不断发展的关系，并不断增长。”，关于这一点，我还不是很懂，可以在后面阅读的过程中慢慢理解。

- ✧ 接下来看 Introduction 的第二段：*“随着上面的例子，大数据时代已经到来[37]，[34]，[29]。每天都有 2.5 万亿字节的数据被创建，当今世界上 90%的数据都是在过去两年产生的[26]。自 19 世纪早期信息技术发明以来，我们生成数据的能力从未如此强大和巨大。另一个例子是 2012 年 10 月 4 日，美国总统奥巴马和州长罗姆尼的首场总统辩论在[46][46]直播的 2 小时内就引发了超过 1000 万条推特。在所有这些推特中，引发讨论最多的具体时刻实际上揭示了公众利益，比如关于医疗保险和代金券的讨论。这种在线讨论提供了一种感知公众利益并实时产生反馈的新手段，与广播、电视等普通媒体相比更具吸引力。另一个例子是 Flickr，一个公共图片分享网站，从 2012 年 2 月到 3 月，平均每天收到 180 万张照片。假设每张照片的大小是 2 兆(MB)，这需要每天 3.6 兆兆(TB)的存储。事实上，正如一句老话：“一张图片胜过一千字，”数十亿图片闪烁是宝箱为我们探索人类社会，社交活动，公共事务，灾害，等等，只有我们有能力利用大量的数据。”*

这一段阐述了一个事实*“大数据时代已经到来”*，具体表现为数据快速生成。这些数据有什么意义呢？作者又举了两个例子，分别是“美国总统巴拉克·奥巴马和州长米特·罗姆尼之间的首场总统辩论推特中**大量讨论（超过 1000 万条）**很有意义”、“Flickr，一个公共图片分享网站，其中的**数十亿**图片是宝箱”。这两个例子中无论是“1000

万条”还是“数十亿”，实际上结论已经呼之欲出，在这个信息爆炸的时代，借助计算机去处理数据，才能高效率地得到我们真正想要的东西，那么怎么去处理和利用这些大量数据来为我们探索人类社会，社交活动，公共事务，灾害提供帮助呢？相信作者在后面将给出解决方案。

- 接下来看 Introduction 的第三段：*“上面的例子展示了大数据应用程序的崛起，其中的数据收集已经极大地增长，超出了常用软件工具在“可容忍的时间流逝”内捕获、管理和处理的能力。大数据应用程序最基本的挑战是探索大量数据，提取有用的信息或知识，为未来的行动[40]。在许多情况下，知识提取过程必须非常高效和接近实时，因为存储所有观测数据几乎是不可行的。例如，射电天文学中的平方公里阵列[17]在中心 5 公里范围内由 1000 到 1500 个 15 米的天线组成。它的视觉灵敏度是现有射电望远镜的 100 倍，回答了有关宇宙的基本问题。然而，由于每秒 40 g (GB)的数据量，从 SKA 生成的数据非常大。虽然研究人员已经证实了一些有趣的模式，比如从 SKA 数据中可以发现瞬时无线电异常[41]，但现有的方法只能在离线方式下工作，无法实时处理这种大数据场景。因此，前所未有的大数据量需要一个有效的数据分析和预测平台来实现对大数据的快速响应和实时分类。”*

这一段实际上和我对上一段总结的中心思想是一致的，但是作者说的更好，这其实是一种功力，在论文中想说清、说好、让别人理解和赞成你想说的东西，实际上不是一件简单的事情。作者在这一段又举了

一个 SKA 射电望远镜的例子，至此，例子应该结束了，因为例子已经从小到大，从我们身边的医疗保险、诺贝尔奖到人类的图片数据、探索宇宙的射电设备。很好的说明了大数据无论与鸡毛蒜皮的小事，还是国家，甚至全人类都是息息相关的。

- ✧ 接下来看 Introduction 的第四段：*“本文的其余部分结构如下:在第 2 节中，我们提出一个 HACE 定理来建模大数据特征。第 3 节总结了大数据挖掘。第四部分概述了作者在该领域的一些重点研究计划和国家研究项目。相关工作在第 5 节进行了讨论，我们在第 6 节对论文进行了总结。”*

这是一篇论文的正常 Abstract 结尾，对文章后面要说的内容做一个划分和概述，看到这里，我对后面的内容是很感兴趣的，作为一个对大数据了解不多的学生，都能因为一个 Abstract 而对整片文章产生兴趣，足见作者功力。

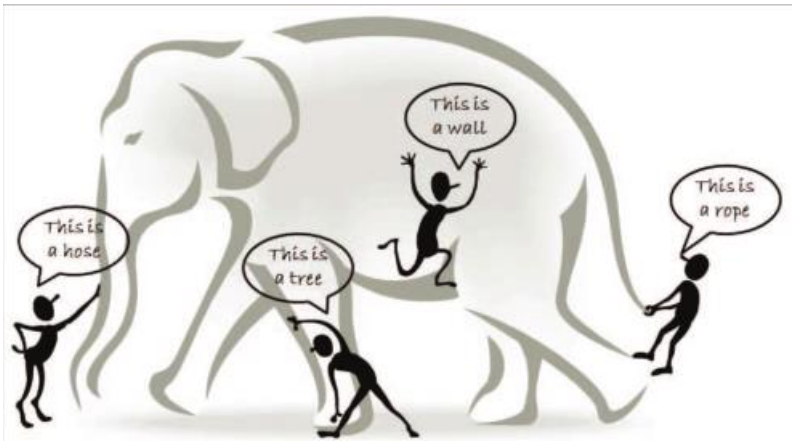
- ✧ *第二节 BIG DATA CHARACTERISTICS: HACE THEOREM 大数据特征：“hace 定理：HACE 定理。大数据始于具有分布式和分散控制的大容量、异构、自治的数据源，并试图探索数据之间复杂和不断发展的关系。这些特征使得从大数据中发现有用的知识成为一个极端的挑战。在 naive 的意义上，我们可以想象一些盲人正在试图衡量一个巨大的大象(见图 1)，这将是大数据在此背景下。每个盲人的目标是根据他在这个过程中收集到的信息，画出大象的图片(或结论)。因为每个人的视野都局限于自己所在的区域，所以毫不奇怪，盲人会各自得出这样的结论:大象“摸起来”就像一根绳*

子、一根水管或一堵墙，这取决于他们每个人所局限的区域。让问题更加复杂的是,让我们假设

1)大象不断快速增长及其构成变化,

2）每个盲人都可能有自己的(可能不可靠和不准确的)关于大象的信息来源(例如，一个盲人可能与另一个盲人交换他对大象的感觉，而交换的知识本身就有偏见)。

在这种情况下，探索大数据就相当于聚合来自不同来源(盲人)的异构信息，以帮助绘制出尽可能好的图像，以实时方式揭示大象的真实姿态。事实上，这项任务并不像让每个盲人描述他对大象的感觉，然后让专家画一幅综合视图的画那么简单，考虑到每个人可能说不同的语言(异构和不同的信息源)，他们甚至可能对他们在信息交换过程中考虑的消息有隐私问题。”



作者通过中国传统故事“盲人摸象”类比，指出在大数据时代下，人们想要从海量数据中得到有用的核心知识或者是一个“综合”的知识是很困难的，紧接着就提出了大数据技术的意义就在于解决这个问题，但是事实上又与这个问题有所不同，因为“考虑到每个人可能说不同的语言(异构和不同的信息源)，或者可能在信

息交换过程中考虑隐私问题”

- *第 2.1 节：Huge Data with Heterogeneous and Diverse Dimensionality 具有**异构和不同维度**的海量数据：“大数据的基本特征之一是由异构和不同维度表示的海量数据。这是因为不同的信息收集者更喜欢他们自己的数据记录模式或协议，并且不同应用的性质也导致不同的数据表示。例如，生物医学世界中的每个人都可以通过使用简单的人口统计信息来表示，例如性别、年龄、家族疾病史等。对于每个人的 x 光检查和 CT 扫描，图像或视频被用来表示结果，因为它们为医生进行详细的检查提供了视觉信息。对于 DNA 或基因组相关的测试，微阵列表达图像和序列被用来表示遗传密码信息，因为这是我们目前的技术获取数据的方式。在这种情况下，**异质特征指的是同一个体的不同类型的表征，而多样特征指的是所涉及的表征每个单一观察的特征的多样性。**想象一下，不同的组织(或医疗从业者)可能有他们自己的模式来代表每个患者，如果我们试图通过组合来自所有来源的数据来实现数据聚合，数据异构性和不同的维度问题将成为主要的挑战。”*

作者提出了“heterogeneous features（异构）”和“diverse features（多维）”的概念，核心思想通过例子表达的很清楚，我的理解如下：每个不同领域有着自己的表达方式，这导致对于同一个事物，其相关数据的“量纲”是不同的，这造成了大数据处理的难度。举个最简单的例子，每个国家的货币单位都是不一样的，如果对货币之间的汇率不加考虑就进行使用，无疑会导致错误，100 元人民币和 100 美金，

虽然数字都是100，但是实际的价值却不一样。弄清数据之间的“汇率”很关键！

- *第 2.2 节：Autonomous Sources with Distributed and Decentralized Control 具有分布式和分散控制的自治源：“具有分布式和分散控制的自治数据源是大数据应用程序的主要特征。由于是自治的，每个数据源都能够生成和收集信息，而不涉及(或依赖)任何集中控制。这类似于万维网(WWW)设置，其中每个网络服务器提供一定量的信息，并且每个服务器能够完全运行，而不必依赖于其他服务器。另一方面，如果整个系统必须依赖于任何中央控制单元，那么庞大的数据量也会使应用程序容易受到攻击或出现故障。对于与大数据相关的主要应用程序，如谷歌、Flicker、脸书和沃尔玛，在世界各地部署了大量服务器场，以确保为当地市场提供不间断的服务和快速响应。这种自主性来源不仅是技术设计的解决方案，也是不同国家/地区立法和监管规则的结果。例如，沃尔玛的亚洲市场在季节性促销、畅销商品和顾客行为方面与北美市场有着内在的不同。更具体地说，当地政府法规也影响批发管理流程，并导致为当地市场重建数据表示和数据仓库。”*

本节指出大数据应用程序主要特征：分布式、分散控制，究其原因，有几点：自治带来独立，防止只有一个数据源导致受到攻击或是故障，此外，还能为当地市场提供相应的服务和快速响应，也方便遵从当地的立法和监管规则，比如在美洲的网站贩卖枪支是合法的，换成中国无疑就不可以了。

✧ *第 2.3 节：复杂且不断发展的关系 Complex and Evolving Relationships：“随着大数据量的增加，数据的复杂性和数据下的关系也在增加。在数据集中信息系统的早期阶段，重点是寻找最佳特征值来表示每个观测值。这类似于使用许多数据字段，如年龄、性别、收入、教育背景等来表征每个人。这种类型的样本特征表示固有地将每个个体视为独立的实体，而不考虑它们的社会联系，这是人类社会。我们的朋友圈子可能是基于共同的爱好形成的，也可能是人们通过生物关系联系在一起的。这种社会关系不仅普遍存在于我们的日常活动中，而且在网络世界中也非常流行。例如，主要的社交网站，如脸书或推特，主要以社交功能为特征，如朋友联系和追随者(在推特中)。个体之间的相关性本质上使整个数据表示和任何数据推理过程变得复杂。在样本-特征表示中，如果个体共享相似的特征值，则它们被认为是相似的，而在**样本-特征-关系**表示中，两个个体可以被链接在一起(通过他们的社会联系)，即使他们在特征域中可能根本没有任何共同之处。在一个动态的世界中，用来代表个体的特征和用来代表我们联系的社会纽带也可能随着时间、空间和其他因素而演变。这种复杂性正在成为大数据应用现实的一部分，其中关键是考虑复杂(非线性、多对多)的数据关系以及不断演变的变化，从大数据集合中发现有用的模式。”*

这一段指出了在数据的早期处理阶段，每个被观测和分析的对象是彼此独立的，但是在真实世界中，个体与个体之间是有千丝万缕的联系

的，这种联系也是这个个体的重要特征，甚至在处理能力越来越强的未来，考虑这些信息随时间变化、空间和其他因素而演变的复杂关系也不是不可能。简单说来，这一段的意思就是：一个人的数据标签是高材生，另外一个人是不识字的老人，如果不考虑个体之间的联系，处理方案必然不会把他们联系在一起，但是实际上，他们可能是父子关系，这就导致他们之间的关系实际上是很大的，这说明考虑数据联系、关系在大数据方向的重要性。

- *第三部分：大数据带来的数据挖掘挑战 DATA MINING CHALLENGES WITH BIG DATA：“对于处理大数据的智能学习数据库系统[52]，关键是要扩大到异常大的数据量，并为上述 HACE 定理的特征提供处理。图 2 显示了**大数据处理框架的概念视图**，该框架从内到外包括三层，考虑了数据访问和计算(第一层)、数据隐私和领域知识(第二层)以及大数据挖掘算法(第三层)。第一层的挑战集中在**数据访问和算术计算程序**上。因为大数据通常存储在不同的位置，并且数据量可能会不断增长，所以有效的计算平台必须将分布式大规模数据存储纳入计算考虑。例如，典型的数据挖掘算法要求将所有数据加载到主内存中，然而，这正成为大数据的一个明显的技术障碍，因为跨不同位置移动数据是昂贵的(例如，受到密集的网络通信和其他输入输出成本的影响)，即使我们确实有一个超大的主内存来保存所有计算数据。第二层的挑战主要**围绕不同大数据应用的语义和领域知识**。此类信息可以为挖掘过程提供额外的好处，并为大数据访问(第一层)和挖掘算*

*法(第三层)增加技术障碍。例如，根据不同的领域应用，数据生产者和数据消费者之间的数据隐私和信息共享机制可能会有很大的不同。对于水质监测等应用来说，共享传感器网络数据可能并不令人气馁，而对于大多数(如果不是全部的话)应用来说，发布和共享移动用户的位置信息显然是不可接受的。除了上述隐私问题之外，应用程序域还可以提供额外的信息来帮助或指导大数据挖掘算法设计。例如，在购物篮交易数据中，每个交易被认为是独立的，并且所发现的知识通常通过找到高度相关的项目来表示，可能关于不同的时间和/或空间限制。另一方面，在社交网络中，用户相互联系，共享依赖结构。然后，知识由用户社区、每个组的领导者和社会影响建模等来表示。因此，理解语义和应用知识对于低级数据访问和高级挖掘算法设计都很重要。在第三层，数据挖掘挑战集中在**算法设计上，以解决大数据量、分布式数据分布以及复杂和动态数据特征带来的困难**。三级圈包含三个阶段。首先，通过数据融合技术对稀疏、异构、不确定、不完整和多源数据进行预处理。二是预处理后挖掘复杂动态的数据。第三，对局部学习和模型融合得到的全局知识进行测试，并将相关信息反馈给预处理阶段。然后，根据反馈调整模型和参数。在整个过程中，信息共享不仅是每个阶段顺利发展的承诺，也是大数据处理的目的。在下文中，我们详细阐述关于图 2 中的三层框架的挑战。”*

看来第三部分是要介绍大数据处理三层框架，以及每层框架的主要挑

战：

1）数据访问和算术计算程序；

2）不同大数据应用的语义和领域知识；

3）算法设计，以解决大数据量、分布式数据分布以及复杂和动态数据特征带来的困难；

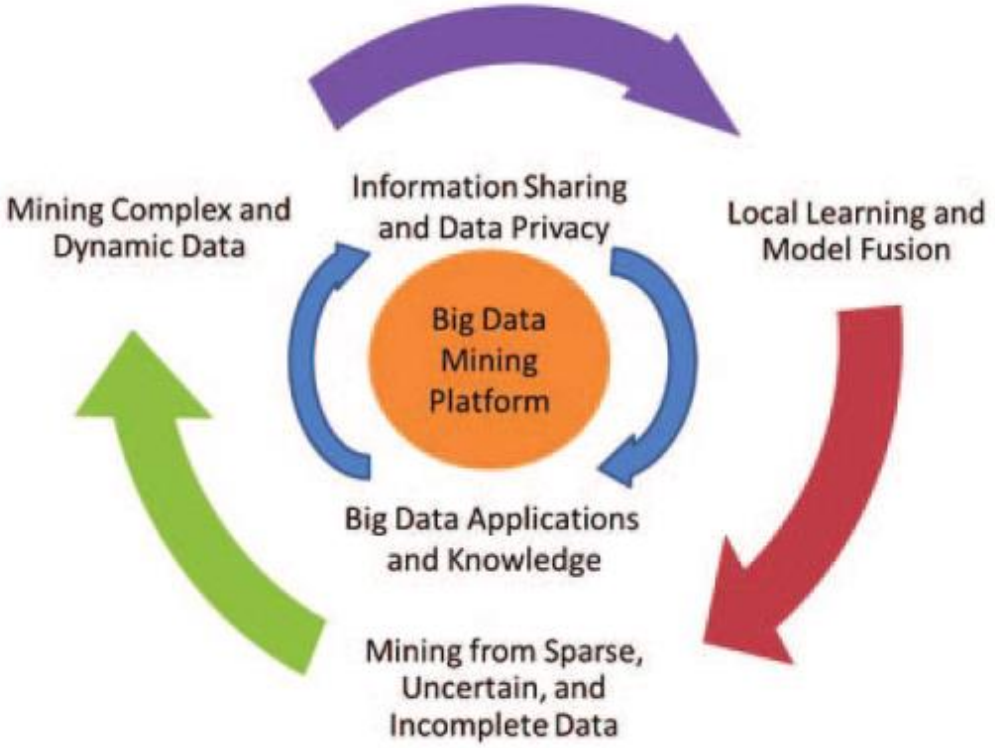


图 2 大数据处理框架:研究挑战形成了一个三层结构，并围绕“大数据挖掘平台”(第一层)展开，该平台侧重于低级数据访问和计算。信息共享和隐私方面的挑战，以及大数据应用领域和知识形成第二层，该层集中于高级语义、应用领域知识和用户隐私问题。最外面的圆圈显示了对实际挖掘算法的三级挑战。

✧ *第 3.1 节：第一层:大数据挖掘平台 Tier I: Big Data Mining Platform:“在典型的数据挖掘系统中，挖掘过程需要计算密集型*

*计算单元来进行数据分析和比较。因此，需要一个计算平台来有效地访问至少两种类型的资源:**数据和计算处理器**。对于**小规模**的数据挖掘任务，一台包含硬盘和中央处理器的台式计算机就足够了来实现数据挖掘的目标。事实上，许多数据挖掘算法都是为这类问题设置而设计的。对于**中等规模**的数据挖掘任务，数据通常很大(并且可能是分布式的)，无法放入主内存。常见的解决方案是依靠并行计算[43]、[33]或集体挖掘[12]对不同来源的数据进行采样和聚合，然后使用并行计算编程(如消息传递接口)来执行挖掘过程。对于**大数据**挖掘，由于数据规模远远超出了单台个人计算机的处理能力，典型的大数据处理框架将依赖于具有高性能计算平台的集群计算机，通过在大量计算节点(即集群)上运行一些并行编程工具(如 MapReduce 或企业控制语言(ECL))来部署数据挖掘任务。软件组件的作用是确保单个数据挖掘任务(如从具有数十亿条记录的数据库中查找查询的最佳匹配)被分成许多小任务，每个小任务运行在一个或多个计算节点上。例如，截至本文撰写之时，部署在田纳西州橡树岭国家实验室的世界上最强大的超级计算机泰坦包含 18，688 个节点，每个节点都有一个 16 核 CPU。这样一个融合了硬件和软件组件的大数据系统，如果没有主要行业股东的支持，几乎是不可用的。事实上，几十年来，公司一直在基于存储在关系数据库中的事务数据做出业务决策。大数据挖掘提供了超越传统关系数据库、依赖结构化程度较低的数据的机会:可以挖掘有用信息的博客、社交媒体、电子邮件、传感器*

和照片。主要的商业智能公司，如 IBM、Oracle、Teradata 等，都有自己的产品来帮助客户获取和组织这些不同的数据源，并与客户的现有数据进行协调，以找到新的见解并利用隐藏的关系。”

这一段指出大数据处理框架第一层需要进行的数据挖掘的等级和需要的条件，对于小规模的数据挖掘任务，一台台式计算机足矣；对于中等规模的数据挖掘任务，依靠并行计算或集体挖掘对不同来源的数据进行采样和聚合，然后使用并行计算编程(如消息传递接口)来执行挖掘过程即可；

但是对于大数据挖掘，就需要具有高性能计算平台的集群计算机，通过在大量计算节点(即集群)上运行一些并行编程工具(如 MapReduce 或企业控制语言(ECL))来部署数据挖掘任务。此时，软件组件的功能就是确保单个数据挖掘任务(如从具有数十亿条记录的数据库中查找查询的最佳匹配)被分成许多小任务，每个小任务运行在一个或多个计算节点上。这也就与我们课程中学习和实践到的 Hadoop 和 MapReduce 的相关知识和实验任务契合了。

✧ *第 3.2 节：第二层:大数据语义和应用知识 Tier II: Big Data Semantics and Application Knowledge：“大数据中的语义和应用知识指的是与法规、策略、用户知识和领域信息相关的许多方面。这一层最重要的两个问题包括*

1)数据共享和隐私；

2)领域和应用知识。

前者为解决如何维护、访问和共享数据的问题提供了答案；而后

者侧重于回答诸如“底层应用程序是什么？”和“用户打算从数据中发现什么知识或模式？”

3.2.1 信息共享和数据隐私信息共享是所有涉及多方的系统的最终目标[24]。虽然共享的动机很明显，但现实世界的一个担忧是，大数据应用程序与敏感信息相关，如银行交易和医疗记录。简单的数据交换或传输不能解决隐私问题[19]，[25]，[42]。例如，知道人们的位置和他们的偏好，可以实现各种有用的基于位置的服务，但是随着时间的推移公开披露个人的位置/移动会对隐私产生严重的后果。为了保护隐私，有两种常见的方法：

1)限制对数据的访问，例如向数据条目添加认证或访问控制，以便敏感信息只能由有限的用户组访问；

2)匿名化数据字段，以便敏感信息不能被精确定位到单个记录[15]。

对于第一种方法，常见的挑战是设计安全的认证或访问控制机制，以便敏感信息不会被未经授权的个人误操作。对于数据匿名化，主要目标是向数据注入随机性，以确保多个隐私目标。比如最常见的 k-匿名隐私措施就是保证数据库中的每个个体都必须与 k 不可区分？1 其他。常见的匿名化方法是使用抑制、泛化、扰动和置换来生成数据的修改版本，实际上，这是一些不确定的数据。基于数据公告的信息共享方法的一个主要好处是，一旦匿名，数据可以在不同方之间自由共享，而不涉及限制性访问控制。这自然导致了另一个研究领域，即隐私保护数据挖掘[30]，其中多方，

每个持有一些敏感数据，试图实现一个共同的数据挖掘目标，而不共享数据中的任何敏感信息。实际上，这种隐私保护挖掘目标可以通过两种方法来解决，包括

1)使用特殊的通信协议，例如姚的协议[54]，来请求整个数据集的分布，而不是请求每个记录的实际值，或者

2)设计特殊的数据挖掘方法来从匿名化的数据中导出知识(这与不确定数据挖掘方法本质上类似)。”

3.2.1作者提出大数据应用程序与敏感信息(如银行交易和医疗记录)相关，简单的数据交换或传输不能解决隐私问题。对“数据共享和隐私”中的第二个方法，作者提出了使用特殊的通信协议来请求整个数据集的分布，而不是请求每个记录的实际值，或者设计特殊的数据挖掘方法来从匿名化的数据中导出知识(这与不确定数据挖掘方法本质上类似)。

3.2.2领域和应用知识 领域和应用知识[28]为设计大数据挖掘算法和系统提供重要信息。在一个简单的例子中，领域知识可以帮助识别建模基础数据的正确特征(例如，在诊断二型糖尿病时，血糖水平显然是比体重更好的特征)。领域和应用知识还可以通过使用大数据分析技术来帮助设计可实现的业务目标。例如，股票市场数据是一个典型的领域，它在每一秒钟内不断产生大量的信息，如出价、买入和卖出。市场不断演变，受到国内外新闻、政府报道、自然灾害等不同因素的影响。一个吸引人的大数据挖掘任务是设计一个大数据挖掘系统来预测未来一两分钟内的市场走势。

这样的系统，即使预测精度比随机猜测稍好，也会给开发者带来显著的商业价值[9]。没有正确的领域知识，找到有效的矩阵/度量来描述市场运动是一个明显的挑战，这种知识通常超出了数据挖掘者的想象，尽管最近的一些研究表明，使用社交网络(如推特)是一个挑战有可能以良好的准确性预测股票市场的上升/下降趋势[7]。

3.2.2作者通过设计预测股票市场未来一两分钟走势的大数据挖掘系统的例子（例子指出没有正确的领域知识，找到有效的矩阵/度量来描述市场运动是河很难的，进而强有力地指出领域和应用知识在大数据应用方面的重要价值：可以帮助识别建模基础数据的正确特征）

✧ *3.3第三层:大数据挖掘算法3.3 Tier III: Big Data Mining Algorithms*

*3.3.1 多信息源的局部学习和模型融合 Local Learning and Model Fusion for Multiple Information Sources:“由于大数据应用程序具有自治源和分散控制的特点，将分布式数据源聚合到一个集中的站点进行挖掘会因为潜在的传输成本和隐私问题而受到系统性的限制。另一方面，虽然我们总是可以在每个分布式站点进行挖掘活动，但对每个站点收集的数据的偏见往往会导致有偏见的决策或模型，就像大象和盲人的情况一样。在这种情况下，**大数据挖掘系统必须启用信息交换和融合机制**，以确保所有分布式站点(或信息源)能够共同努力实现全局优化目标。模型挖掘和相关性是确保从多个信息源中发现的模型或模式能够被整合*

*以满足全局挖掘目标的关键步骤。更具体地说，**全局挖掘**可以在数据、模型和知识层面上以两步(**局部挖掘和全局关联**)过程为特征。在数据级别，每个本地站点可以基于本地数据源计算数据统计，并在站点之间交换统计，以实现全局数据分布视图。在模型或模式级别，每个站点都可以针对本地化数据执行本地挖掘活动，以发现本地模式。通过在多个来源之间交换模式，可以通过聚集所有站点的模式来合成新的全球模式[50]。在知识层面，模型相关性分析研究从不同数据源生成的模型之间的相关性，以确定数据源之间的相关性，以及如何基于从自治源构建的模型形成准确的决策。”*

3.3.1 提出了多信息源的局部学习和模型融合，简单说来就是：由于数据源往往是分布式的，要将多个站点集中起来进行挖掘会有传输成本和隐私问题的困难，所以在各个分布式站点进行挖掘活动，然后在信息交换和融合机制下实现全局目标的优化，具体体现在数据级别（各本地分布站点交换数据统计）和知识层面（通过数据源相关性）。

*3.3.2 从稀疏、不确定和不完整的数据中挖掘 Mining from Sparse, Uncertain, and Incomplete Data：“备用、不确定和不完整的数据是大数据应用程序的定义特征。由于稀疏，数据点的数量太少，无法得出可靠的结论。这通常是数据维度问题的一个复杂问题，在这种情况下，高维空间(如超过 1000 个维度)中的数据不会显示出明显的趋势或分布。对于大多数机器学习和数据挖掘算法，**高维备用数据会显著降低从数据中导出的模型的可靠性**。*

常见的方法是采用**降维或特征选择**[48]来降低数据维数，或者小心地包含额外的样本来缓解数据稀缺，例如数据挖掘中的通用无监督学习方法。不确定数据是一种特殊类型的数据现实，其中每个数据字段不再是确定性的，而是服从一些随机/误差分布。这主要与数据读取和收集不准确的领域特定应用相关。比如 GPS 设备产生的数据本身就具有不确定性，主要是因为设备的技术壁垒将数据的精度限制在一定的水平(比如 1 米)。因此，每个记录位置由平均值加上方差来表示，以指示预期误差。对于与数据隐私相关的应用[36]，用户可能会有意将随机性/错误注入数据以保持匿名。这类似于这样一种情况，即一个人可能不愿意让你知道他/她的确切收入，但可以提供一个大致的范围，如[120，160k]。**对于不确定数据，主要的挑战是每个数据项被表示为样本分布而不是单个值，因此大多数现有的数据挖掘算法不能直接应用。常见的解决方案是考虑数据分布来估计模型参数。**例如，错误感知数据挖掘[49]利用每个单个数据项的平均值和方差值来构建朴素贝叶斯分类模型。类似的方法也被应用于决策树或数据库查询。**数据不完整**是指某些样本的数据字段值缺失。丢失的值可能是由不同的实际情况引起的，例如传感器节点的故障，或者有意跳过一些值的一些系统策略(例如，丢弃一些传感器节点读数以节省传输功率)。虽然**大多数现代数据挖掘算法都有内置的解决方案来处理缺失值(如忽略缺失值的数据字段)**，但**数据插补**是一个既定的研究领域，旨在估算缺失值以产生改进的模型(与从原始数据构建的模

型相比)。为此，存在许多插补方法[20]，主要方法是根据给定实例的观察值，填充最常见的观察值或建立学习模型来预测每个数据字段的可能值。”

在 3.3.2，作者提出三大数据带来困难，分别是 Sparse, Uncertain, and Incomplete Data，作者指出稀疏的高维数据带来困难，会显著降低从数据中导出的模型的可靠性，常见的方法是采用降维或特征选择来降低数据维数，或者加入额外的样本来缓解数据稀缺，例如数据挖掘中的通用无监督学习方法；此外，作者指出不确定数据也会带来困难，主要的困难是每个数据项被表示为样本分布而不是单个值，因此大多数现有的数据挖掘算法不能直接应用。常见的解决方案是考虑数据分布来估计模型参数。最后，作者还指出数据不完整（是指某些样本的数据字段值缺失）带来困难。解决方法是大多数现代数据挖掘算法都有内置的解决方案来处理缺失值(如忽略缺失值的数据字段)，或是旨在估算缺失值以产生改进的模型数据插补方法（大部分方法是根据给定实例的观察值，填充最常见的观察值或建立学习模型来预测每个数据字段的可能值。）

*3.3.3 挖掘复杂动态数据 Mining Complex and Dynamic Data:“复杂数据的快速增长及其在数量和性质上的变化推动了大数据的崛起[6]。发布在万维网服务器、互联网主干网、社交网络、通信网络和运输网络等上的文档都以复杂的数据为特征。**虽然数据下复杂的依赖结构增加了我们学习系统的难度，但它们也提供了简单的数据表示无法实现的令人兴奋的机会**。例如，研究人员已经成*

功地使用著名的社交网站 Twitter 来检测地震和重大社交活动等事件，速度接近实时，准确性非常高。此外，通过总结用户提交给搜索引擎的查询，现在有可能建立一个早期预警系统来检测快速传播的流感爆发[23]。对大数据应用程序来说，利用复杂数据是一项重大挑战，因为**复杂网络中的任何两方都有可能通过社交联系对彼此感兴趣。这种连接与网络中的节点数量成二次关系，因此一百万个节点的网络可能会受到一万亿个连接的影响。**对于像脸书这样的大型社交网站来说，活跃用户的数量已经达到 10 亿，分析这样一个庞大的网络对大数据挖掘来说是一个巨大的挑战。**如果我们把日常的用户行为/交互考虑进去，难度会更惊人。**受上述挑战的启发，已经开发了许多数据挖掘方法来从具有复杂关系和动态变化量的大数据中找到感兴趣的知识。例如，寻找社区并追踪它们动态演化的关系对于理解和管理复杂系统是必不可少的[3]，[10]。发现社交网络中的离群点[8]是识别垃圾邮件发送者并为我们的社会提供安全网络环境的第一步。如果只是面对海量的结构化数据，用户可以简单的通过购买更多的存储或者提高存储效率来解决问题。然而，大数据的复杂性表现在许多方面，包括复杂的异构数据类型、数据中复杂的内在语义关联以及数据之间复杂的关系网络。也就是说，**大数据的价值在于它的复杂性。**复杂的异构数据类型。**在大数据中，数据类型包括结构化数据、非结构化数据和半结构化数据等。**具体来说，有表格数据(关系数据库)、文本、超文本、图像、音频和视频数据等。现有的数据模

*型包括键值存储、大表克隆、文档数据库和图形数据库，它们按照这些数据模型的复杂程度的升序排列。传统的数据模型无法处理大数据环境中的复杂数据。目前，没有公认的有效和高效的数据模型来处理大数据。数据中复杂的内在语义关联。网络上的新闻、推特上的评论、Flicker 上的图片和 YouTube 上的视频剪辑可能会同时讨论一个学术获奖事件。毫无疑问，这些数据中有很强的语义关联。从“文本-图像-视频”数据中挖掘复杂的语义关联将极大地有助于提高搜索引擎或推荐系统等应用系统的性能。然而，**在大数据环境下，如何高效地描述语义特征和建立语义关联模型来弥补各种异构数据源之间的语义鸿沟是一个巨大的挑战。**数据中的复杂关系网络。在大数据环境中，个体之间存在关系。在互联网上，个人是网页，网页通过超链接相互链接形成一个复杂的网络。个人之间也存在社交关系，形成复杂的社交网络，如来自脸书、推特、领英和其他社交媒体的大关系数据[5]、[13]、[56]，包括通话详细记录(CDR)、设备和传感器信息[1]、[44]、全球定位系统和地理编码地图数据、通过管理文件传输协议传输的大量图像文件、网络文本和点击流数据[2]、科学信息、电子邮件[31]等。为了处理复杂的关系网络，新兴的研究工作已经开始解决结构与进化、群体与互动以及信息与交流等问题。大数据的出现也催生了用于实时数据密集型处理的新计算机架构，例如运行在高性能集群上的开源 Apache Hadoop 项目。大数据(包括事务和交互数据集)的规模或复杂性超出了在合理成本内捕获、管理和处*

理这些数据的常规技术能力和时间限制。在大数据环境下，复杂数据的实时处理是一项极具挑战性的任务。”

3.3.3节的信息量很大,这段话提出的最精彩的一句话我认为是“That is to say, the value of Big Data is in its complexity.（也就是说，大数据的价值在于它的复杂性）”。作者指出了一个很深层的观点：数据复杂的依赖结构增加了学习的难度，但是复杂也有复杂的巨大好处，作为一个对大数据只是很浅层次了解的学生，这句话无疑是很有启发性的,作者举出通过搜索引擎的查询历史可以建立早期预警系统，比如近期某一地区的人都开始搜索“喉咙痛了怎么办、感冒了怎么办”之类的问题，那么就可能可以预警该地区最近有流感爆发，此外，复杂网络中的社交联系也是很有趣的，一方面社交联系与网络中的节点数量成两次关系，这对大数据挖掘来说有是巨大的挑战，如果要是把用户行为/交互加入数据中，难度更是惊人，出于对上面问题的思考,现在已有很多从具有复杂关系和动态变化量的大数据中找到感兴趣的数据挖掘方法。

此外，“文本-图像-视频”数据中挖掘复杂的语义关联也是一个契机（极大地有助于提高搜索引擎或推荐系统等应用系统的性能。),如何高效地描述语义特征和建立语义关联模型来弥补各种异构数据源之间的语义鸿沟就有了巨大的意义。新兴的研究工作也已开始致力于解决这些问题，最后，大数据还生了用于实时数据密集型处理的新计算机架构，例如运行在高性能集群上的开源Apache Hadoop项目。

✧ *第四部分：研究计划和项目 RESEARCH INITIATIVES AND PROJECTS：*

“为了应对大数据挑战并“抓住新的数据驱动解决方案提供的机会”，美国国家科学基金会(NSF)在奥巴马总统政府的大数据倡议下，于 2012 年宣布了大数据征集。这样一个联邦倡议已经产生了许多获奖项目，以研究大数据管理的基础(由华盛顿大学领导)、基于基因组学的海量数据计算的分析方法(由布朗大学领导)、可能高达 50 万维的高维数据集的大规模机器学习技术(由卡内基梅隆大学领导)、大规模科学文献的社会分析(由罗格斯大学领导)以及其他几个项目。这些项目寻求开发方法、算法、框架和研究基础设施，使我们能够将海量数据降低到人类可管理和可解释的规模。其他国家如国家自然科学基金(NSFC)在大数据研究上也在追赶国家拨款。与此同时，自 2009 年以来，作者在以下涉及大数据组件的国家项目中处于领先地位:。在生物网络中整合和挖掘来自多个来源的生物数据，由美国国家科学基金会赞助，中期资助号 CCF-0905337，2009 年 10 月 1 日-2013 年 9 月 30 日。问题和意义。我们整合并挖掘了来自多个来源的生物数据，以破译和利用生物网络的结构，为生物系统的功能提供新的见解。我们讨论整合和挖掘生物网络的理论基础以及当前和未来的使能技术。我们扩展和整合了信息网络的信息获取、传输和处理的技术和方法。我们开发了基于语义的数据集成方法、从挖掘的数据中自动生成假设的方法，以及用于评估模拟结果和优化模型的自动可扩展分析工具。。大数据快速响应。大数据流实时分类，澳大利亚研究委员会(ARC)主办，批准号:DP130102748，2013 年 1 月 1 日-2015 年

12 月 31 日。问题和意义。我们建议构建一个基于流的大数据分析框架，用于快速响应和实时决策。主要挑战和研究问题包括:设计大数据采样机制，将大数据量减少到可管理的处理规模；-根据大数据流构建预测模型。这种模型可以自适应地调整以数据的动态变化，以及准确预测数据未来的趋势；和知识索引框架，以确保大数据应用程序的实时数据监控和分类。带通配符和长度约束的模式匹配与挖掘，国家自然科学基金，资助号:60828005(第 1 阶段，2009 年 1 月 1 日-2010 年 12 月 31 日)和 61229301(第 2 阶段，2013 年 1 月 1 日-2016 年 12 月 31 日)。问题和意义。我们对模式匹配、带通配符的模式挖掘和应用问题进行了如下系统研究:-探索匹配和挖掘问题的 NP-hard 复杂性，-带通配符的多模式匹配，-近似模式匹配和挖掘，以及我们的研究在无处不在的个性化信息处理和生物信息学上的应用。多种异构数据源集成和挖掘的关键技术，由中国国家高技术研究与发展计划(863 计划)资助，批准号 2012AA011005，2012 年 1 月 1 日-2014 年 12 月 31 日。问题和意义。我们对多源、海量、动态信息的可用性和统计规律进行了研究，包括基于信息抽取的跨媒体搜索、抽样、不确定信息查询、跨领域、跨平台的信息聚合。为了突破传统数据挖掘方法的局限性，我们研究了复杂线内数据的异构信息发现和挖掘、数据流挖掘、海量多源数据的多粒度知识发现、海量知识的分布规律、海量知识的质量融合。社会网络中的群体影响和互动，由中国国家基础研究 973 计划资助，批准号 2013CB329604，2013 年 1

月 1 日-2017 年 12 月 31 日。问题和意义。我们研究了社会网络中的群体影响和互动，包括:运用群体影响和信息扩散模型，运用动态博弈理论研究社会网络中的群体互动规则；研究受群体情绪影响的社会网络下的互动个体选择和效应评估，分析个体和群体之间的情绪互动和影响；建立社会网络群体的互动影响模型及其计算方法，揭示社会网络的互动影响效应和演化。”

第四部分介绍了国内外集团对于大数据技术的扶持，并介绍了作者在涉及大数据组件的国家项目中处于领先地位的部分。

✧ *第五部分 相关工作 RELATED WORK*

第 5.1 节大数据挖掘平台(第一层) Big Data Mining Platforms (Tier I)：“由于分布式环境中涉及的应用数据具有多源、海量、异构和动态的特点，大数据最重要的特点之一是对千兆字节(PB)甚至千兆字节(EB)级别的数据进行计算，计算过程复杂。因此，利用并行计算基础设施、相应的编程语言支持和软件模型来高效地分析和挖掘分布式数据是大数据处理从“数量”向“质量”转变的关键目标，目前，大数据处理主要依靠 MapReduce 等并行编程模型，以及为公众提供大数据服务的云计算平台。MapReduce 是一个面向批处理的并行计算模型。与关系数据库相比，性能仍有一定差距。随着 MapReduce 并行编程被应用于许多机器学习和数据挖掘算法，提高 MapReduce 的性能和增强大规模数据处理的实时性受到了极大的关注。数据挖掘算法通常需要扫描训练数据以获得统计量来求解或优化模型参数。频繁访问大规模数据需要密

集的计算。为了提高算法效率，Chu 等人提出了一种通用的并行编程方法，该方法基于多核处理器上简单的 MapReduce 编程模型，适用于大量的机器学习算法。在该框架中实现了 10 种经典的数据挖掘算法，包括局部加权线性回归、k-均值、逻辑回归、朴素贝叶斯、线性支持向量机、自变量分析、高斯判别分析、期望最大化和反向传播神经网络[14]。通过对这些经典机器学习算法的分析，我们认为算法学习过程中的计算操作可以转化为对大量训练数据集的求和操作。求和操作可以在不同的子集上独立执行，并且可以在 MapReduce 编程平台上轻松执行。因此，一个大规模的数据集可以被分成几个子集，并分配给多个映射器节点。然后，可以在映射器节点上执行各种求和操作来收集中间结果。最后，通过合并归约节点上的求和并行执行学习算法。Ranger 等人【39】提出了一种基于 MapReduce 的应用编程接口 Phoenix，支持多核和多处理器系统环境下的并行编程，实现了 k-Means、主成分分析和线性回归三种数据挖掘算法。Gillick 等人[22]改进了 Hadoop 中 MapReduce 的实现机制，评估了 MapReduce 框架下单通道学习、迭代学习和基于查询的学习算法的性能，研究了并行学习算法中涉及的计算节点之间的数据共享、分布式数据存储，进而表明 MapReduce 机制适用于大规模数据通过在中型集群上测试一系列标准数据挖掘任务来进行挖掘。Papadimitriou 和 Sun [38]提出了一个分布式协作聚合(DisCo)框架，使用了实用的分布式数据预处理和协作聚合技术。在开源 MapReduce 项目 Hadoop 上的实现表

明，DisCo 具有良好的可扩展性，能够处理和分析海量数据集(数百 GB)。为了改善传统分析软件扩展性差、Hadoop 系统分析能力差的问题，Das 等人【16】进行了 R(开源统计分析软件)与 Hadoop 集成的研究。深度集成将数据计算推向并行处理，这为 Hadoop 提供了强大的深度分析能力。Wegener 等人【47】实现了 Weka(开源机器学习和数据挖掘软件工具)和 MapReduce 的集成。标准的 Weka 工具只能在单台机器上运行，内存限制为 1gb。算法并行化后，Weka 通过利用并行计算在 MapReduce 集群上处理超过 100 GB 的数据，突破了限制，提高了性能。Ghoting 等人【21】提出 Hadoop-ML，在其上开发者可以在语言运行时环境下，在程序块上轻松构建任务并行或数据并行的机器学习和数据挖掘算法。”

第五部分的第 5.1 节作者指出，大数据最重要的特点之一是对兆字节级别的数据进行计算，继而计算过程复杂，所以引出需要并行计算基础设施、相应的编程语言支持和软件模型。目前，大数据主要依靠的是 MapReduce 等并行编程模型，以及为公众提供大数据服务的云计算平台。此外随着大数据越受关注，人们开始致力于提高 MapReduce 的性能和增强大规模数据处理的实时性。

接下来作者介绍了人们在提高 MapReduce 的性能和增强大规模数据处理的实时性方向所做的努力：Chu 等人提出了一种通用的并行编程方法，来提高算法效率，该方法基于多核处理器上简单的 MapReduce 编程模型，思想是“算法学习过程中的计算操作可以转化为对大量训练数据集的求和操作。求和操作可以在不同的子集上独立执行，并且

可以在 MapReduce 编程平台上轻松执行。因此，一个大规模的数据集可以被分成几个子集，并分配给多个映射器节点。然后，可以在映射器节点上执行各种求和操作来收集中间结果。最后，通过合并归约节点上的求和并行执行学习算法。”此外作者还介绍了 Ranger 等人、Gillick 等人、Papadimitriou 和 Sun、Das 等人、Wegener 等人、Ghoting 等人的工作。

- *在 5.2 节，大数据语义和应用知识(第二层) Big Data Semantics and Application Knowledge(Tier II)：“在海量数据的隐私保护方面，叶等人[55]提出了一种多层粗糙集模型，该模型能够准确描述不同级别的泛化所产生的粒度变化，为衡量数据在匿名化过程中的有效性标准提供了理论基础，并设计了一种平衡隐私和数据效用的动态机制，以解决分类的最优泛化/细化顺序。最近一篇关于大数据保密性保护的论文[4]总结了多种保护公开发布数据的方法，包括聚合(如 kanonymity、I-diversity 等)。)，抑制(即，删除敏感值)，数据交换(即，切换敏感数据记录的值以防止用户匹配)，添加随机噪声，或者简单地用从模拟分布合成生成的值替换具有高泄露风险的整个原始数据值。对于涉及大数据和巨大数据量的应用程序，通常情况下数据会物理地分布在不同的位置，这意味着用户不再物理地拥有其数据的存储。为了进行大数据挖掘，拥有一个高效且有效的数据访问机制是至关重要的，尤其是对于打算雇佣第三方(如数据挖掘者或数据审计者)来处理其数据的用户。在这种情况下，用户的隐私限制可能包括 1)没有本地数*

据拷贝或下载，2)所有分析必须基于现有的数据存储系统部署，而不违反现有的隐私设置等。在王等人[48]中，提出了一种用于大规模数据存储(例如云计算系统)的隐私保护公共审计机制。基于公钥的机制用于启用第三方审核(TPA)，因此用户可以安全地允许第三方分析他们的数据，而不会违反安全设置或危及数据隐私。对于大多数大数据应用程序，隐私问题集中在排除第三方(如数据挖掘者)直接访问原始数据。常见的解决方案是依靠一些隐私保护方法或加密机制来保护数据。Lorch 等人[32]最近的一项研究表明，用户的“数据访问模式”也可能存在严重的数据隐私问题，并导致地理上位于同一位置的用户或具有共同兴趣的用户的披露(例如，搜索相同地图位置的两个用户可能在地理上位于同一位置)。在他们的系统，即 Shround 中，用户从服务器访问数据的模式是通过使用虚拟磁盘隐藏的。因此，它可以支持各种大数据应用，如微博搜索和社交网络查询，而不会损害用户隐私。”

在 5.2 节，作者对人们在大数据语义和应用知识方面所做的努力做出介绍：在海量数据的隐私保护方面，叶等人[55]提出了一种多层粗糙集模型，在保护公开发布数据的方法方面，“Big Privacy: Protecting Confidentiality in Big Data”这篇论文总结的很好，此外，王等人[48]中，提出了一种用于大规模数据存储(例如云计算系统)的隐私保护公共审计机制。Lorch 等人[32]最近的一项研究表明，用户的“数据访问模式”也可能存在严重的数据隐私问题。

✧ *在 5.3 节，大数据挖掘算法(第三层)Big Data Mining Algorithms*

(Tier III)：“为了适应多源、海量、动态的大数据，研究人员以多种方式扩展了现有的数据挖掘方法，包括提高单源知识发现方法的效率[11]，从多源角度设计数据挖掘机制[50]、[51]，以及研究动态数据挖掘方法和分析流数据[18]、[12]。从海量数据中发现知识的主要动机是提高单源挖掘方法的效率。在计算机硬件功能逐步完善的基础上，研究人员不断探索提高知识发现算法效率的方法，使其更好地适用于海量数据。因为海量数据通常是从不同的数据源收集的，所以海量数据的知识发现必须使用多源挖掘机制来执行。由于现实世界的数据通常以数据流或特征流的形式出现，因此需要一种成熟的机制来发现知识并掌握动态数据源中知识的演变。因此，多源数据的海量性、异构性和实时性为单源知识发现和多源数据挖掘提供了本质区别。吴等[50]、[51]、[45]提出并建立了局部模式分析理论，为多源数据挖掘中的全局知识发现奠定了基础。该理论不仅为完全搜索问题提供了解决方案，而且为寻找传统挖掘方法无法找到的全局模型提供了解决方案。数据处理的局部模式分析可以避免将不同的数据源放在一起进行集中计算。数据流广泛应用于金融分析、在线交易、医学测试等领域。静态知识发现方法不能适应动态数据流的连续性、可变性、快速性和无限性等特点，容易导致有用信息的丢失。因此，需要有效的理论和技术框架来支持数据流挖掘[18]，[57]。知识进化是现实世界系统中的普遍现象。例如，临床医生的治疗方案将根据患者的情况不断调整，如家庭经济状况、健康保险、疗程、

治疗效果和分布心血管和其他慢性流行病随着时间的推移而发生的变化。在知识发现过程中，概念漂移旨在分析数据流中由动态和上下文触发的隐含目标概念变化甚至根本变化的现象。根据不同类型的概念漂移，基于单个特征、多个特征和流特征，知识进化可以采取突变漂移、渐进漂移和数据分布漂移的形式[53]。”

在 5.3 节部分，作者针对大数据挖掘算法进行了已有的工作介绍，介绍了吴等提出并建立的局部模式分析理论。该模式为多源数据挖掘中的全局知识发现奠定了基础。该理论不仅为完全搜索问题提供了解决方案，而且为寻找传统挖掘方法无法找到的全局模型提供了解决方案。数据处理的局部模式分析可以避免将不同的数据源放在一起进行集中计算。数据流广泛应用于金融分析、在线交易、医学测试等领域。静态知识发现方法不能适应动态数据流的连续性、可变性、快速性和无限性等特点，容易导致有用信息的丢失。

✧ *第六部分对全文做了总结：“在现实应用程序和关键行业利益相关者的推动下，由国家融资机构启动，管理和挖掘大数据已被证明是一项具有挑战性但非常紧迫的任务。虽然“大数据”一词实际上与数据量有关，但我们的 HACE 定理表明，大数据的关键特征是:1)巨大的异构和多样的数据源，2)分布式和分散控制的自治性，3)数据和知识关联的复杂性和演变性。这些综合特征表明，大数据需要“大头脑”来整合数据以实现最大价值[27]。为了探索大数据，我们分析了数据、模型和系统级别的几个挑战。为了支持大数据挖掘，需要高性能计算平台，这些平台需要系统化的设计*

来释放大数据的全部力量。在数据层面上，自主信息源和数据收集环境的多样性往往导致数据具有复杂的条件，例如缺失/不确定的值。在其他情况下，隐私问题、噪音和错误可能会被引入到数据中，从而产生更改的数据副本。开发安全可靠的信息共享协议是一项重大挑战。在模型级别，关键的挑战是通过组合本地发现的模式来生成全局模型，以形成统一的视图。这需要精心设计的算法来分析分布式站点之间的模型相关性，并融合来自多个来源的决策，以从大数据中获得最佳模型。在系统级别，基本的挑战是大数据挖掘框架需要考虑样本、模型和数据源之间的复杂关系，以及它们随时间和其他可能因素的变化。系统需要仔细设计，以便非结构化数据可以通过它们的复杂关系链接起来，形成有用的模式，数据量和项目关系的增长应该有助于形成合法的模式来预测趋势和未来。我们认为大数据是一种新兴趋势，所有科学和工程领域都需要大数据挖掘。借助大数据技术，我们将有望提供最相关、最准确的社会感知反馈，以更好地实时了解我们的社会。我们可以进一步鼓励公众受众参与社会和经济事件的数据生产圈。大数据时代已经到来。本工作得到了国家 863 计划(2012AA011005)、国家 973 计划(2013CB329604)、国家自然科学基金(61229301、61273297 和 61273292)、美国国家科学基金(NSF CCF-0905337)和澳大利亚研究委员会未来基金(FT100100971)的资助作者感谢匿名审稿人对改进论文的宝贵和建设性意见。”

本段是对论文的主旨进行了精辟的总结，再次强调了 HACE 定理，可

以说该定理是作者在对大数据深刻理解和大量自己和他人工作的总结下，以俯视的角度提出的理论，理论指出大数据的关键特征是:

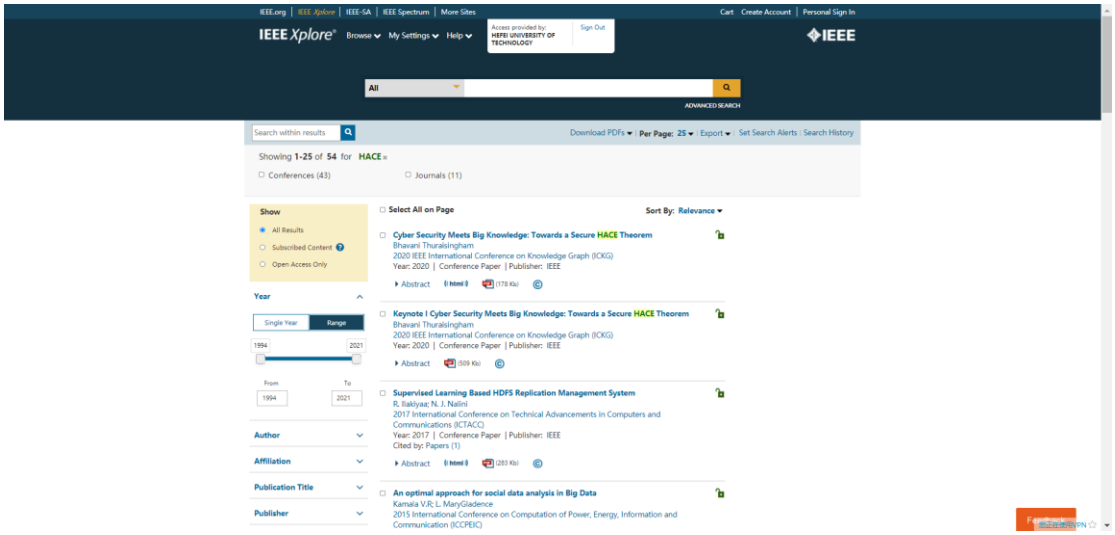
1)巨大的异构和多样的数据源

2)分布式和分散控制的自治性

3)数据和知识关联的复杂性和演变性。

这些综合特征表明，大数据需要“大头脑”来整合数据以实现最大价值。此外作者还再次简要分析了数据、模型和系统级别的几个挑战。

读完全文，我认为这篇文章之所以在 2013 年 12 月到 2015 年 6 月连续 19 个月被下载的次数均排第一，一是因为其对大数据方向相关工作做了很精辟的介绍和总结，使得后来要做大数据方向的科研工作者们可以快速对大数据已有的架构和工作有个大致的了解，继而可从 Literature Survey 中寻找自己感兴趣的数据领域及其提出的方法。，是一篇很好的学习大数据挖掘的综述论文，论文题目也完美的概括了文章的工作“Data Mining with Big Data”，二是由于其对多项工作、理论进行总结和分析，最后提出了 HACE 定理，精辟的指出大数据的关键特征是:1)巨大的异构和多样的数据源 2)分布式和分散控制的自治性 3)数据和知识关联的复杂性和演变性。这一定理对大数据特征定义是俯视性的，在大数据方面的含金量很足，在 IEEE 上检索 HACE，有 54 条相关论文，其中最新的就在 2020 年。



主要事件

年份	事件	相关论文/Reference
2014	Wei Ding等人提出了HACE定理	Wu, X.; Zhu, X.; Wu, G.; Ding, W. (2014). Data mining with big data. IEEE Transactions on Knowledge and Data Engineering. 26(1): 97-107.
2014	Saint John Walker发表文章探讨了大数据对我们生活和工作的影响，及其未来趋势	Walker, S. J. (2014). Big Data: A Revolution That Will Transform How We Live, Work, and Think. International Journal of Advertising. 33(1): 181-183.
2015	Murtaza Haider整合了从业者和学者的定义，对大数据进行了综合描述	Gandomi, A.; Haider, M.; (2015). Beyond the hype: Big data concepts, methods, and analytics. International Journal of Information Management. 35(2): 137-144.
2016	Nada Elgendy和Ahmed Elragala提出了大数据，分析和决策（B-DAD）框架	Elgendy, N.; Elragala, A.(2016). Big Data Analytics in Support of the Decision Making Process. Procedia Computer Science. 100: 1071-1084.