未命名

《信息素养实践》期末论文设计

|  |
| --- |
| **大数据核心技术概论** |

作者姓名：王亚红

所在学院：软件学院

专业（系）：软件工程

班级（届）：2017级软件2班

二〇一八 年 一 月 十一 日

目录

[中文摘要 1](#_Toc503522187)

[中文关键词 1](#_Toc503522188)

[1 大数据简介 2](#_Toc503522189)

[1.1 引言 2](#_Toc503522190)

[2.大数据概括 3](#_Toc503522191)

[2.1 大数据的概念和特点 3](#_Toc503522192)

[2.2 大数据国内外发展历程 4](#_Toc503522193)

[2.3 生活中的大数据 5](#_Toc503522194)

[2.3.1商业中的大数据 5](#_Toc503522195)

[2.3.2体育竞技中的大数据 6](#_Toc503522196)

[2.3.3日常生活中的大数据 6](#_Toc503522197)

[3大数据处理流程 7](#_Toc503522198)

[3.1 数据获取 7](#_Toc503522199)

[3.2 数据集成 8](#_Toc503522200)

[3.3 数据分析 8](#_Toc503522201)

[3.4 数据解释 9](#_Toc503522202)

[4关键技术分析 9](#_Toc503522203)

[4.1 云计算 10](#_Toc503522204)

[4.2 大数据处理工具 10](#_Toc503522205)

[4.3 大数据时代的挑战 11](#_Toc503522206)

[5结论 12](#_Toc503522207)

[参考文献 13](#_Toc503522208)

[英文摘要 14](#_Toc503522209)

[英文关键字 14](#_Toc503522210)

大数据核心技术概论

软件学院 软件工程专业

作者 王亚红

中文摘要：大数据是信息技术和计算方法迅速发展的必然产物。商业数据、科学数据和网页数据等初始数据的种类日益增多，规模呈爆炸式增长，大数据时代已经到来，大数据已成为继物力和人力资源之后的又一重要资源，将在社会发展过程中发挥不可替代的作用，大数据必将带来一场生活、工作和思维的大变革。本文简要介绍了大数据的相关概念和特点，国内外大数据的发展历程，从非关系数据库、大数据处理工具和隐私保护三个方面阐述了大数据相关技术发展水平，然后分析大数据带来的挑战，最后总结全文并对大数据进行了展望。

中文关键词：大数据，大数据技术，隐私保护，挑战

# 1 大数据简介

### 1.1 引言

大数据是信息化时代迅速发展的产物和必然结果。互联网成为数据获取的主要渠道，云计算改变数据存储访问方式，物联网成为大数据搜集的来源，社交网络把人类真是世界映射到网络，智能终端的普及给大数据带来的丰富鲜活的数据。

大数据时代下，数据规模越来越大，增长速度越来越快。2010年，全球数据量跨入了ZB时代，据数据数据公司（IDC）公司研究报告，未来10年全球数据量将以40+%的速度增长，2020年全球数据量将达到35ZB，为2009年（0.8ZB）的44倍。此外，数据类型繁多，包括结构化、半结构化和非结构化数据，甚至包括非完整和错误的数据。现代互联网上半结构化和非结构化数据所占的比例已达95%以上。

大数据时代的到来，撼动了世界的方方面面，。从商业、科技、医疗卫生到政府、教育以及社会的其他各个领域。开展大数据技术与应用研究，是时代发展的必然要求，具有无可估量的社会经济价值和巨大的科学意义。目前，大数据已被应用于金融、零售、电影等行业，并取得了巨大的成功。大数据及其相关处理技术可转化为巨大的社会经济价值，被誉为“未来的石油”。大数据技术和应用一方面对社会、经济和科技的发展带来了重要机遇，另一方面也对数据的获取、存储、传输、计算以及应用提出了全新的挑战。

在对现有的大数据研究资料中进行阅读和总结的基础上，本文简要介绍了大数据的概念、特点、国内外的发展历程，在此基础上，从非关系数据库、大数据处理工具和隐私保护三个方面阐述了大数据相关技术发展水平，然后分析了大数据带来的机遇与挑战，最后总结全文并对大数据处理进行了展望。

# 2.大数据概括

### 2.1 大数据的概念和特点

最先经历信息爆炸的学科，如天文学和基因学，创造出了“大数据”这个概念。大数据并非一个确切的概念，尚未有一个确切、统一的定义。目前关于大数据的几个主流定义有：研究机构Garther的定义，大数据是指需要新处理模式才能具有更强的决策力、洞察发现力和流程优化能力的海量、高增长率和多样化的信息资产。维基百科的定义，大数据指的是所涉及的治疗量规模巨大到无法通过目前主流软件工具，在合理时间内达到撷取、管理、处理并整理成为帮助企业经营决策目的的资讯，麦肯锡的定义，大数据是指无法在一定时间内用传统数据库软件工具对其内容进行采集、存储、管理和分析的数据集合。IDC的定义，大数据一般会涉及两种或两种以上的数据形式，他要收集超过100TB的数据，并且是高速、实时数据流，或者是从小数据开始，但数据每年会增长60%以上。大数据科学家John Rauser提到一个简单的定义，大数据就是超过了任何一个计算机处理能力的庞大信息量。

目前大数据的特点有3V和4V两种说法，2014年5月1日美国白宫发布的关于大数据与隐私的两份报告中仍采用3V特点：规模性（Volume）、多样性（Variety）和快速性（Velocity）。所谓规模性，是指数据的存储量大和计算量大；多样性是指要处理的数据来源多、格式多，其中数据类型除包含结构化数据外，还有半结构化和非结构化数据；快速性是指数据的增长速度快、处理速度要求快。4V特点是在3V特点上增加一个V，但关于第4V并未形成统一的观点，IDC认为大数据应当具有价值性（Value），大数据的价值往往呈现稀疏性的特点，而IBM则认为大数据具有真实性（Veracity）的特点。【EB】

### 2.2 大数据国内外发展历程

大数据时代的到来，引起了国内外学术界、工业界和政府的广泛关注。Nature早在2008年就推出了Big Data专刊，从互联网技术、超级计算、生物医学等方面专门探讨了大数据的研究。2011年2月，Science推出专刊“Dealing with Data”，主要讨论了科学研究重大数据的相关问题并说明大数据对于科学研究的重要性。2011年6月，全球知名咨询公司麦肯锡（McKinsey）发布了一份关于大数据的详细报告“Big data: The next frontier for innovation, competition, and productivity”，详细分析了大数据的影响、关键技术和应用领域。2012年3月份美国奥巴马政府发布了“大数据研究和发展倡议”正式启动“大数据发展计划”，旨在提高和改进人们从海量数据中获取信息的能力。2012年5月，日本重启ITC站撸计划，重点关注“大数据应用，将大数据定位为战略领域之一”。2013年初，英国商业、创新和技能部对大数据的投资达1.89亿英镑，。国务大臣戴维威利茨说，政府将在计算基础设施方面投入巨资，加强数据采集和分析，从而在数据革命中占得先机。2013年2月，法国政府发布《数据化路线图》，列出了5项将会大力支持的战略型高新技术，其中一项就是大数据。2013年6月，日本安倍内阁正式公布新IT战略---“创新最尖端IT国家宣言”，“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。【EB】

与国外相比，国内大数据研究起步稍晚，企业使用数据挖掘技术尚不普遍，但近几年出现了蓬勃发展的态势。2012年5月，香山科学会议组织了以“大数据科学与工程---一门新兴的交叉学科”为主题的回忆，深入讨论了大数据的理论与工程数据研究、应用方向。2012年6月9日，中国计算机学会常务理事会决定成立大数据专家委员会CCF，2012年10月，成立了首个专门研究数据应用和发展的学术咨询组织---中国通信学会大数据专家委员会，推动了我国大数据的科研与发展。2012年11月，首届数据科学与信息产业大会在北京国际数学研究中心召开，在中国大数据相关会议中具有里程碑的意义。2013年12月6日，“2013年中国智慧政府发展年会”在北京外受宾馆隆重举行，大会以“大数据时代的政府在线服务”，对新时期我国政府网站的发展具有重要推动作用。2014年5月15日，上海市宣布自今年起推动各级政府部门将数据开放，并鼓励社会对其进行加工和运用。2014年5月20日至23日，第六届中国云计算大会于北京国家会议中心拉开帷幕，此次会议通过技术专场、产品发布和培训课程等方式，深度剖析云计算与大数据的核心技术。中国学术界、工业界以及政府对大数据的广泛关注，必将促进我国大数据的快速发展。

### 2.3 生活中的大数据

2.3.1商业中的大数据

沃尔玛是最早利用大数据而受益的企业之一，曾拥有世界山最大的数据仓库系统。通过对消费者的购物行为等非结构化数据进行分析，沃尔玛成为最了解顾客购物习惯的零售商，并创造了经典商业案例。

Tesco PLC（特易购）这家超市连锁在其数据仓库中收集了700万部冰箱的数据。通过对这些数据的分析，进行更全面的监控并进行主动的维修以降低整体能耗。

梅西百货（Macy’ s ）的实时定价机制可以根据需求和库存的情况进行实时调价。该公司基于SAS的系统可以对多达7300万种货品进行实时调价。

2.3.2体育竞技中的大数据

F1赛车场可能是大数据最经典的应用场景之一，一辆辆风驰电擎的造价高达200万美元的F1赛车的设计、模拟、测试和建造完全在电脑中完成，这个流程的每个环节都将产生大量数据。虽然F1的模拟测试需要昂贵的计算机软硬件环境，但这仍然比在赛道上实测的成本要低，据悉，一辆F1赛车在赛道上实测的花费每天高达40万~60万美元。

新的国际祁连规则意味着各大F1车队需要对赛车设计进行大改，以达到新的节能指标，搭载新的V6引擎也需要心得传统系统预置匹配。莲花车队的对策是部署两个并行的IT项目支撑全新的赛车设计和测试工作。其中一个在工厂端，运行微软的Dynamics商务套件；另外一个在车队端，只要是与赛道测试相关的数据采集和分析。据悉，莲花车队将采用EMC的存储、虚拟化软件和思科的服务器搭建其大数据环境，此外据EMC市场总监Jeremy Burton向GigaOM透露，莲花车队也可能采购EMC的Atoms用于存储和管理内容，Syncplicity用于异地同步和分享文件，Data Domian用于备份和恢复。

2.3.3日常生活中的大数据

美国电业公司TXU Energy发明了一种智能电表技术。有了智能电表，公司能每隔15分钟就读一次用电数据，而不是过去的一个月一次，从而大大节省了抄表的人工费用。且由于能高频率快速采集分析用电数据（产生大数据），供电公司能根据用电高峰和低谷时段制定不同的电价。该公司甚至打出了这样的宣传口号“亲，晚上再洗衣服洗碗吧，晚上用电不要钱”。实际上，智能电表和大数据应用真正让分时动态定价成为可能，而且对于TXU Energy和用户来说是一个双赢的结果。

又如Prada试衣间的大数据。传统奢侈品牌Prada正在向大数据时代迈进，其在纽约及一些旗舰店里开始了大数据时代行动。在纽约旗舰店里，每件衣服上都有RFID码。每当顾客拿起衣服进试衣间时，这件衣服上的RFID会被自动识别，试衣间里的屏幕会自动播放模特穿着这件衣服走台步的视频、人一看见模特就会下意识里认为自己穿上衣服就会是那样，不由自主地会认可手中所拿衣服。

# 3大数据处理流程

大数据数据规模巨大，数据种类繁多，处理方式多种多样，但大数据也有自己的基本处理流程，即数据获取、数据集成、数据分析和数据解释。

### 3.1 数据获取

多源异构数据的获取是数据处理的基本前提。随着互联网、物联网、云计算的快速发展，互联网已成为获取数据的主要渠道，物联网成为大数据搜集来源，社交网络把人类真实世界映射到网络，以博客、微博为代表的社交网络和以手机、平板电脑为代表的移动智能终端的快速发展，每天都会产生大量的数据。云计算改变了传统的数据存储方式，为大数据的存储和计算提供了基础平台和技术支撑。当前时代下，数据获取方式众多，其中最主要的有如下几种数据获取技术：传感器技术、Web2.0技术、条形码技术、RFID技术和移动终端技术、智能可穿戴设备（谷歌眼镜、智能手表等）。

### 3.2 数据集成

数据集成主要完成数据的抽取、清洗和存储操作。数据获取阶段的数据规模巨大、种类繁多，数据的抽取将复杂的数据类型转化为单一的或者便于处理的构型。原始数据价值密度稀疏，很大一部分数据相对于某一特定应用来说可利用价值较低或者没有价值，甚至可能包含错误的“噪声”数据，数据清洗可以保证数据的质量和可信度。数据存储为数据处理提供一个平台，一个良好的数据库，可以实现数据的存储、索引、查询，并可以有效地解决数据难以复用的问题。目前，数据库种类繁多，既包含传统的关系数据库，也包含非关系数据库。传统的关系型数据库有MySQL和Oracle等，菲关系型数据库有Redis和MongoDB等。

### 3.3 数据分析

数据分析是指从多源异构大数据中提取价值信息的过程，是大数据处理的核心。当前情况下，我们正处于数据泛滥而我们所能提取到的有用信息较少的时代，这个问题已成为世界各国尤其是发达国家重点关注的对象。面对几大的数据规模、繁多的数据类型和某些因时效性约束需要快速处理大规模数据集合，传统的数据管理技术、数据处理和分析不能直接移植于大数据时代，必须做相应的改变才能使用与大数据的分析与处理。根据大数据处理多样性的需求和不同的特征维度，典型的大数据分析模型可如图3.3.1所示进行分类。

**大数据分析模型分类**

**Table1 Classification of big data analysis model**

|  |  |
| --- | --- |
| 典型大数据计算模式 | 典型系统 |
| 大数据查询分析计算 | Hbate，Hive，Cassandra，Impala，Hana等  Hadoep MapReduce，Spark等  Storm，S4，Flume等  HaLoop，MapReduce，Spark等  Pregel，Giraph，PowerGraph，GraphX等  Dremel，Hana，Spark等 |
| 批处理计算  流式计算  迭代计算  图计算  内存计算 |

### 3.4 数据解释

数据界实时将数据处理结果展示给用户的关键技术。数据分析需要专业的数据分析知识，而数据处理的结果需要展示给普通的用户，显然用户也不一定具有数据分析的相关知识，因此，如何将结果清晰地展示给用户是一个重要问题。当前主要采用将数据处理结果通过图形、图像化或者表格化的方式，主要有宇宙星球图法、标签云法和历史流图法等。

# 4关键技术分析

大数据技术是指从各种类型的数据中快速获取有价值信息的技术。大数据领域已经涌现出大量新的技术，他们成为大数据采集、存储、管理、分析和解释的有力工具。大数据处理关键技术一般包括大数据采集、大数据存储和管理、大数据分析及挖掘、大数据解释和应用（大数据检索、大数据可视化、大数据应用、大数据安全等）。大数据处理和使用需要多学科、多技术的协同，本文仅从非关系数据库、大数据处理工具以及隐私保护三个方面进行简要介绍。

### 4.1 云计算

Google首席执行官埃里克施密特于2006年8月9日在搜索引擎大会（SES San Jose2006）首次提出“云计算”的概念。云计算是分布式计算、并行计算等传统计算机技术和网络技术发展融合的产物，云存储是建立在云计算上的一个新概念，指通过集群应用、网络技术或分布式文件系统等功能，将网络中大量各种不同类型的存储设备通过应用软件集合起来协同工作，共同对外提供数据存储和业务访问功能的一个系统。云存储是一个以数据存储和管理为核心的云计算系统。

从技术上看，大数据与云计算的关系就像一枚硬币的正反面一样密不可分。大数据必然无法用单台计算机进行处理，必须采用分布式计算架构。云计算能为大数据提供强大的存储和计算能力。

### 4.2 大数据处理工具

大数据带来的巨大商业价值驱动着一些依靠数据牟利的大公司的激烈竞争，近几年，大数据处理工具如“雨后春笋”般不断出现，从最早的Hadoop开源平台，到现在的Yahoo！的S4和Twitter的Storm等，几乎每个公司都拥有自己针对特定用途的处理平台。当前情况下，Hadoop、Storm和Spark从这些大数据处理工具中脱颖而出，网络上关于“2014数据时代Storm、Spark、Hadoop”三个大数据处理工具为代表简要介绍大数据处理工具。

### 4.3 大数据时代的挑战

大数据时代，大数据已成为继物力和人力资源的另一重要资源，将在社会经济发展过程中发挥不可替代的作用：大数据及相关技术可转化为巨大的社会经济价值，被誉为“未来的新石油”。大数据在金融、零售、电影等行业取得的成功仅是其价值体现的冰山一角。然而，由于大数据的多源异构、复杂多样、变化快等特点，使得大数据时代的数据管理、数据分析以及信息安全面临着新的挑战。

大数据时代下的数据管理机制不同于传统方式。大数据时代下，数据的增长速度远远超过存储空间的增长速度，现有的数据管理方法已不能适用于海量多源异构数据在多种存储设备之间频繁密集流动，不同的应用对数据传输“时效性”和“完整性”的要求也不尽相同。因此，需要研发新型存储技术、传输交换机理，以满足大数据时代对数据的海量存储以及数据的跨层、跨城、实时和完整的传输的需求。

大数据的规模效应冲击着传统的数据分析方法。传统的数据分析方法主要是针对规模较小的结构化数据，且已形成一套行之有效的分析体系。大数据的出现将变革传统的思维方式。大数据的精髓在于我们分析信息时的三个转变，这些转变将改变我们理解和组建社会的方法。第一个转变是，在大数据时代，我们可以分析更多的数据，有时候甚至处理和某个特别现象相关的所有数据，而不是依赖于随机采样。第二个转变就是，研究数据之多，以至于我们不再热衷于追求精确度。第三个转变安因前两个转变而成，即我们不在热衷于寻找因果关系，而是寻找食物之间的相关关系。另外，由于大数据时代对数据分析实时性的要求，传统的数据分析方法不能直接移植与大数据时代，必须做相应的改变才能使用与大数据的分析和处理。

# 5结论

“大数据”的影响增加了对信息管理专家的需求，甲骨文、IBM、微软和思爱普等公司花了超过15亿美元在软件只能数据管理和分析的专业公司上。这个行业自身价值超过了1000亿美元，增长近10%。

大数据已经出现，我们生活在一个信息的社会中，有46亿全球移动电话用户，有20亿人访问互联网。基本上，人们比以往任何时候都需要有大数据或信息交互。大数据其影响力涵盖了经济、政治、文化等方面。大数据可以帮助人们开启循“数”管理的模式，也将成为“大社会”的集中体现。有人说：“三分技术，七分数据，得数据者得天下”。大数据技术的应用前景是十分光明的，当前，我们正处在全面建成小康社会征程中，工业化、信息化、城镇化、农业现代化任务很重，建设下一代信息基础设施，发展现代信息技术产业体系，健全信息安全保障体系，推进嘻嘻网络技术广泛运用，是实现四化同步发展的保证。大数据是新一代信息技术的集中反映，是一个应用驱动性很强的服务领域，是具有无穷潜力的新兴产业领域；目前，其标准和产业格局尚未形成，这是我国实现跨越式发展的宝贵机会。我们要从战略上重视大数据的开发利用，将它作为转变经济增长模式有效抓手，但要注意科学规划，切忌一哄而上。现在探讨中国互联网新的价值和他的科学价值或者产业价值的时候，会有很多机遇，努力探索这方面的工作。

# 参考文献

[1]高景德，王祥珩．交流电机的多回路理论．清华大学学报，1987，27(1)：1-8.

[2]欧阳静，《云计算---大数据时代带来全球性深刻变革》，民生视窗聚焦，2013年3月26日

[3][美]维克托.迈尔.舍恩伯格著，周涛 译，《大数据时代》，浙江人民出版社，2012年12月

[4]李永宏，《大数据与云计算》，热点透视，2013年6月

Introduction to big data core technologies

Abstract：Big data is the inevitable result of the rapid development of information technology and calculation method. Business data, scientific data and web data such as the types of the initial data are increasing, the scale of explosive growth, big data era has arrived, the big data has become the material resources and human resources after another important resources, will play an irreplaceable role in the process of social development, big data will bring a big change of life, work and thinking. This paper briefly introduces the related concepts and the characteristics of big data, the development of large data at home and abroad, from the relational database, data processing tools and privacy protection from three aspects this paper expounds the big data related technology development level, and then analyzed challenges posed by big data, finally summarizes the full text and to the big data is discussed.

Key Words: big data, big data technology, privacy protection, challenge