云时代的流式大数据挖掘服务平台：基于元建模的视角

朱小栋　著

科学出版社

云时代的流式大数据挖掘服务平台：基于元建模的视角

朱小栋 著

科学出版社
北京

内 容 简 介

在云时代，大数据蕴涵的知识和规律为人类社会创造了前所未有的重大价值。流式大数据挖掘平台是实施流式大数据挖掘的软件服务平台，是处理流式大数据的数据挖掘系统。构建智能、高效和快速流式大数据挖掘平台，满足人们对数据的高吞吐低延迟、计算程序的动态扩展、知识的共享交换与集成的要求，是当前大数据研究的迫切要求和焦点之一。本书内容分两篇：

第一篇是理论篇，运用形式化方法提出“元”理念和“元”理论，进而提出面向流式大数据挖掘平台的元数据和元建模的概念。同时，提出预测模型标记语言的扩展理论，该理论所设计的扩展预测模型标记语言可以应用于流式大数据挖掘平台。

第二篇是建模篇，围绕流式大数据挖掘服务平台，提出流式大数据挖掘服务平台的数据管理理论和算法管理理论。

书中应用多种形式化方法，从理论高度回答了流式大数据和流式大数据挖掘的本质是什么的问题。

本书可供相关领域的研究人员参考，也可以作为高等院校信息技术专业高年级本科生和研究生的教材。

图书在版编目(CIP)数据

云时代的流式大数据挖掘服务平台：基于元建模的视角/朱小栋著.
—北京：科学出版社，2015.8
ISBN 978-7-03-045388-1

Ⅰ. ①云… Ⅱ. ①朱… Ⅲ. ①数据采集—研究 Ⅳ. ①TP274

中国版本图书馆 CIP 数据核字(2015)第 193610 号

责任编辑：王 哲 王迎春 / 责任校对：郭瑞芝
责任印制：张 倩 / 责任设计：迷底书装

科学出版社出版
北京东黄城根北街16号
邮政编码：100717
http://www.sciencep.com

北京源海印刷有限责任公司印刷

科学出版社发行 各地新华书店经销

*

2015 年 8 月第 一 版 开本：720×1000 1/16
2015 年 8 月第一次印刷 印张：9
字数：163 000

定价：48.00 元

（如有印装质量问题，我社负责调换）

作者简介

朱小栋，1981 年 8 月生，安徽太湖人，现居上海，上海理工大学管理学院信息管理系副教授，研究生导师。2009 年毕业于南京航空航天大学计算机应用技术专业，获工学博士学位。目前研究方向包括国际电子商务、数据挖掘、软件工程。公开发表科研论文 60 余篇，出版高等院校教材 1 部。主持教育部人文社会科学青年基金项目 1 项，教育部博士点青年基金项目 1 项，教育部重点实验室开放课题 1 项，上海市教育委员会科研创新基金项目 1 项，其他纵横向课题多项。拥有软件著作权 4 项，指导大学生参加市级以上竞赛获奖 20 余次，已培养硕士研究生 10 名。曾获中国机械工业科学进步奖二等奖 1 项，上海市教育委员会教学成果奖二等奖 1 项。

序

流式大数据广泛存在于我们身边。互联网技术的广泛应用，使得世界市场二分化为传统实体市场和互联网虚拟市场。电子商务就是在互联网虚拟市场所进行的交易活动。 二十多年来，电子商务从萌芽到快速崛起，发展速度惊人。在电子商务交易过程中，产生了大量的流式大数据。当电子商务需要进一步拓展市场的时候，分析电子商务网站的大量用户点击流式大数据，挖掘客户的消费习惯和消费特征，就可以发现用户偏好并给出商家个性化推荐。

2014 年 12 月 31 日，上海外滩发生的踩踏事件给人们带来惨痛的教训：人们不禁思考能否对群体中大量的位置和轨迹数据加以计算，预测导致踩踏发生的人流对冲事件，并作出预警和干预。位置和轨迹数据也是从具有 GPS 定位功能的移动设备采集的流式大数据。

在城市 PM2.5 排放的批量大数据中，应用数据挖掘技术寻找规律，可以预测雾霾天气。对大数据进行流式计算已成为亟待解决的问题，谁抢得大数据计算的制高点，谁就能够赢得市场先机。

2014 年 6 月 9 日，习近平主席在中国科学院第十七次院士大会、中国工程院第十二次院士大会上的讲话中提到，由于大数据、云计算、移动互联网等新一代信息技术同机器人技术相互融合的步伐加快，我国将成为机器人的最大市场，这就需要我们审时度势、全盘考虑、抓紧谋划、扎实推进；2015 年 3 月 5 日，李克强总理在政府工作报告中提出，要制定“互联网+”行动计划，推动移动互联网、云计算、大数据、物联网等与现代制造业结合，促进电子商务、工业互联网和互联网金融的健康发展，引导互联网企业拓展国际市场。这些都显示出大数据的研究与应用已经上升到国家战略。

朱小栋副教授的研究适应了国家发展大数据的战略需求，本书在阐述大数据基本理论的基础上，通过分析基于扩展预测模型标记语言的流式大数据挖掘服务平台，提出了流式大数据挖掘服务平台的元数据体系结构、流式大数据挖掘的数据建模理论和流式大数据挖掘服务平台的算法管理模型，为云环境下快速、高效和智能的流式大数据计算服务模型构建问题和智能化发展提供了理论和方法上的指导。本书的研究成果在电子商务、社交网络、智能交通、环境监测、天体与地壳运动等领域有着广泛的应用前景。

作为一个年轻学者，能够捕捉当前经济发展的热点是非常可贵的。大数据本身是一种新的科学手段，虽然目前还不成熟，但已经开始受到广泛的关注。希望本书作者能够进一步深入这方面的研究，也希望广大读者能够通过本书了解大数据，从科学的角度来理解大数据并在实际工作中加以应用。

杨坚争

2015年4月

前 言

流式大数据是一个随着时间推移不断出现的无限项目序列。与海量大数据相比，流式大数据的特征表现为突发性、实时性、无序性、无限性。物联网、移动社交媒体、无线传感器网络以及电子商务的发展，使得流式大数据的存在形式非常丰富。例如，用于环境和生态监测的传感器网络会持续产生流式大数据，移动社交网络中的各个终端用户的动作会产生流式大数据，电子商务环境中持续的用户交易行为在服务器上会持续产生流式大数据，以及交通路网中用传感器采集的持续车辆流量信息也表现为流式大数据的形式。

大数据蕴涵大知识，大知识为人类社会创造前所未有的重大价值。大数据挖掘模式分为批量大数据挖掘和流式大数据挖掘两种形态。对流式大数据进行挖掘，及时发现大数据中隐藏的规律已成为流式大数据领域的迫切需求。然而，传统的先存储后计算的批量大数据挖掘理念不适用于流式大数据挖掘的环境，流式大数据挖掘对计算机和服务器的带宽、存储空间、处理器、采集设备能源供应提出了新的要求，给计算系统的可伸缩性、系统容错、状态一致性、负载均衡、数据吞吐量带来前所未有的挑战。如何构建高吞吐率、低延迟且持续可靠运行的流式大数据挖掘系统是当前亟待解决的问题，但是目前实践成果与研究经验相对较少。

在云计算环境下，云服务是被用来作为服务提供使用的云计算产品。凭借云计算提供的庞大存储能力、极大的带宽和庞大的处理器，流式大数据挖掘以云服务的方式呈现，是未来流式大数据挖掘的必然趋势。

Berners-Lee 于 2001 年在 *Nature* 上发表文章提出语义 Web 的概念。下一代网络——语义 Web 的出现和发展，为实现智能、高效和快速的流式大数据挖掘系统提供了解决思路。语义 Web 的目标是让 Web 上的信息能够被机器理解，从而实现 Web 信息的自动处理。在语义 Web 环境中，信息的语义能够很好地加以定义，并使人机更好地协同工作。语义 Web 的支撑技术建立在一系列技术标准和规范之上，其中描述逻辑是语义 Web 的逻辑基础，也是语义 Web 具有智能的基础。本书以此为切入点，探索语义 Web 技术与流式大数据挖掘服务的结合，提出构建流式大数据挖掘服务平台相关的理论模型，满足智能、高效和快速流式大数据挖掘服务对数据的高吞吐率、低延迟、计算程序的动态扩展、知识的共享交换与集成的要求。

本书是教育部人文社会科学青年基金项目（No.12YJC870037）、教育部高等学校博士学科点基金项目（No.20123120120004）和 2014 年度计算智能与信号处

理教育部重点实验室（安徽大学）开放课题的研究成果，沪江基金研究基地专项“电子商务智库”（No.14008）给予了资助。本书参考引用了众多国内外数据挖掘研究领域专家学者的文献资料，在此对他们的工作表示衷心的感谢。

由于作者水平有限，书中难免存在不足之处，敬请广大读者批评指正。

作 者

2015 年 3 月

注 释 表

缩略词	全称	注释
AMF-DSMS	algorithms management framework for data streams mining system	面向流式大数据挖掘服务平台的算法管理框架
CRISP-DM	cross-industry standard process for data mining	跨行业数据挖掘标准流程
CWM	common warehouse metamodel	公共仓库元模型
DL4PMML	description logic for predictive model markup language	用于构建扩展预测模型标记语言（EPMML）的支撑描述逻辑
DMG	data mining group	数据挖掘联盟，制定了数据挖掘领域的元数据标准（预测模型标记语言）
DSMSF	data streams mining system framework	流式大数据挖掘服务平台框架
EPMML	extended predictive model markup language	扩展预测模型标记语言
ETL	extract-transform-load	用来描述数据从数据源经过萃取、转置、加载到目的端的过程
GRIP	graphical route information panel	图形式路径信息情报板
MDA	model driven architecture	模型驱动架构
MOF	meta object facility	元对象设施，是一个用来定义、构造、管理、交换和集成软件系统中元数据的模型驱动的分布式对象框架
OMG	object management group	对象管理组织，该组织制定过很多行业标准，例如，CORBA、UML、CWM 和 MDA 等
OWL	web ontology language	Web 本体语言，是由 W3C 制定的用于描述语义 Web 上本体论的语言
PM-DSM	process model for data streams mining	流式大数据挖掘过程模型
PMML	predictive model markup language	预测模型标记语言，是数据挖掘领域的元数据标准
XMI	XML metadata interchange	XML 元数据交换，提供了元数据交换的标准方法

目　录

第二篇　建模篇

第一篇　理论篇

第一篇 理 论 篇

第1章 绪 论

我们获得的知识越多，未知的知识就会更多，因而，知识的扩充永无止境。

——统计学家Rao

21世纪的第二个十年，物联网技术、移动互联网技术、社交媒体技术、电子商务技术和云计算技术等新兴信息技术和应用模式快速发展壮大，伴随而来的是全球数据量急剧增加，推动人类社会迈入大数据时代。

1.1 大数据的概念

查阅维基百科[①]，可以找到大数据的概念：大数据，或称巨量数据、海量数据，是指在可接受的容忍时间之内，大小超出常用的理论、方法、技术、软件工具、数据库管理软件等捕获、存储、处理数据能力的数据。

1.1.1 大数据的特征

大数据呈现的特征如下：

1）数量大（volume）

大数据到底有多大？在计算机科学与技术领域，bit 是最小的数据存储单位，用于存放一位二进制数 0 或者 1。通常，以字节（byte，B）作为数据基本单位：1B = 8bit。对数据大小的描述经历了 B、KB（kilo byte）、MB（mega byte）、GB（giga byte）、TB（tera byte）、PB（peta byte）、EB（exa byte）、ZB（zetta byte）、YB（yotta byte）……的发展过程。按照进率 2^{10}=1024 来计算：

1B = 8bit

1KB = 2^{10} B = 1024B

1MB = 2^{10} KB = 2^{20} B = 1048576B

1GB = 2^{10} MB = 2^{30} B = 1073741824B

1TB = 2^{10} GB = 2^{40} B = 1099511627776B

① Wikipedia. Big data[EB/OL]. 2015-01-08. http://en.wikipedia.org/wiki/Big_data[2015-03-12].

$1PB = 2^{10}TB = 2^{50}B = 1125899906842624B$

$1EB = 2^{10}PB = 2^{60}B = 1152921504606846976B$

$1ZB = 2^{10}EB = 2^{70}B = 1180591620717411303424B$

$1YB = 2^{10}ZB = 2^{80}B = 1208925819614629174706176B$

2009 年左右，500G 的硬盘是很好的个人计算机（personal computer, PC）配置。2012 年，普通 PC 的硬盘可以达到 1TB 级别。然而，当数据上升到 1PB 的时候，则需要 1024 块 1TB 的硬盘。数据量还将以每两年 3 倍的速度增加，这一速度超过了摩尔定律的增长速度[1]①。这样的数据量过于庞大，以至于不能用传统工具存储。成熟的分布式技术、计算机网络技术成为解决流式大数据存储和分析的基础技术。

2）速度快（velocity）

一方面，大数据表现为产生速度快、传播速度快，呈现鲜明的流式特征。另一方面，在处理流式大数据时，要求数据及时快速地得到处理，故而对数据的处理分析能力提出更高的要求。

3）多样性（variety）

数据种类繁多，结构化、半结构化、非结构化的数据并存。即便是结构化的数据，也呈现异构的现象。例如，关系表数据可以用 Oracle、Microsoft、IBM 等不同公司的数据库管理软件存储，也可以用 XML 等标记语言标记。同时，半结构化、非结构化的数据所占的比例不断增加。

4）价值大（value）

当数据规模达到一定的程度时，大数据中隐含的知识、规律的价值凸显出来，有必要采取有效的数据挖掘技术，找出这些知识、规律，推动企业和社会的进步。谷歌、亚马逊和脸书这三家互联网巨头积累了大规模的数据资产，谷歌为全世界的公开网页建立了庞大的索引；亚马逊沉淀了大量的商业信息，拥有互联网上庞大的商品数据库；脸书积累了全世界庞大的人际关系数据库。这些数据的商业价值巨大。

1.1.2　大数据的分类

从数据的流式特征强弱角度，大数据可以分为批量大数据和流式大数据两种形态。大数据计算主要有批量计算和流式计算两种形态[2-5]。

① 摩尔定律，由英特尔公司创始人之一 Moore 提出，其内容是：当价格不变时，集成电路上可容纳的元器件的数目，大约每隔 18~24 个月便会增加一倍，性能也将提升一倍。换言之，每一美元所能买到的计算机性能将每隔 18~24 个月翻一倍以上。这一定律揭示了信息技术进步的速度，可用于观测或推测，而不是一个自然法则。