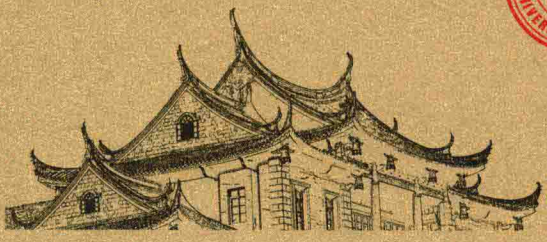
厦门大学南强丛书【第六辑】



Xiamendaxue Nanqiang Congshu

事件挖掘的理论算法及应用

李涛 徐建 张亮◎著

厦门大学出版社 XIAMEN UNIVERSITY PRESS
国家一级出版社
全国百佳图书出版单位

厦门大学南强丛书

【第六辑】

事件挖掘的理论算法及应用

李涛 徐建 张亮◎著

厦门大学出版社 XIAMEN UNIVERSITY PRESS
国家一级出版社
全国百佳图书出版单位

图书在版编目(CIP)数据

事件挖掘的理论算法及应用/李涛,徐建,张亮著. —厦门:厦门大学出版社,2016.3
(厦门大学南强丛书. 第6辑)
ISBN 978-7-5615-5980-2

Ⅰ.①事… Ⅱ.①李…②徐…③张… Ⅲ.①数据采集 Ⅳ.①TP274

中国版本图书馆CIP数据核字(2016)第054935号

出 版 人 蒋东明
责任编辑 郑 丹
装帧设计 李夏凌
责任印制 许克华

出版发行 厦门大学出版社
社 址 厦门市软件园二期望海路39号
邮政编码 361008
总 编 办 0592-2182177 0592-2181253(传真)
营销中心 0592-2184458 0592-2181365
网 址 http://www.xmupress.com
邮 箱 xmupress@126.com
印 刷 厦门集大印刷厂印刷

开本 720mm×1000mm 1/16
印张 19
插页 4
字数 316千字
版次 2016年3月第1版
印次 2016年3月第1次印刷
定价 59.00元

厦门大学出版社
微信二维码

厦门大学出版社
微博二维码

作者简介

李涛，2004 年 7 月毕业于美国罗彻斯特大学（University of Rochester），获计算机科学博士学位。现为美国佛罗里达国际大学计算机学院正教授（Full Professor）、厦门大学闽江学者讲座教授、南京理工大学讲座教授、南京邮电大学计算机学院特聘教授，2016 年入选国家“千人计划”特聘专家。长期从事数据挖掘、信息检索、大数据分析等方面的研究工作，在基于矩阵分解的数据挖掘和学习、智能推荐系统、音乐信息检索、系统日志挖掘等研究方向上做出了有影响力的工作，在国际著名会议及期刊上已经发表超过 200 篇文章。是数据挖掘和知识发现的国际权威期刊 ACM Transactions on Knowledge Discovery from Data (ACM TKDD)、IEEE Transactions on Knowledge and Data Engineering (IEEE TKDE) 和 Knowledge and Information Systems (KAIS) 的副主编。于 2006 年获得美国国家自然科学基金委颁发的“杰出青年教授奖”（NSF CAREER Award，2006–2010），并多次获得“IBM 学院研究奖”（IBM Faculty Research Awards）和 2010 年“IBM 大规模数据分析创新奖”（2010 IBM Scalable Data Analytics Innovation Award）。

徐建，2007 年 1 月毕业于南京理工大学计算机学院，获得计算机科学博士学位，留校任教，目前为南京理工大学计算机学院软件工程系书记、副教授。长期从事数据挖掘、软件自愈及其在系统管理中的应用研究，在异常检测、系统日志挖掘、性能数据分析等方面做了深入的工作，在国内外重要学术期刊和会议上发表论文 50 余篇，获得授权专利 4 项，登记软件著作权 10 余项。主持国家自然科学基金项目 1 项、教育部高等学校博士点基金项目 1 项、南京理工大学自主科研计划项目 2 项、企业公司委托课题 5 项，作为主要成员参加过部委的基础科研项目 3 项、国家自然科学基金项目 2 项、江苏省“973 计划”项目 1 项。

张亮，2010 年毕业于东南大学信息科学与工程学院，获得电路与系统博士学位，同年加入华为网络研究部，从事网络相关创新研究。致力于将数据挖掘技术与网络结合，提高网络自动化水平。在异常检测、故障根因分析、SDN/NFV 等方面做了深入的工作，获得授权专利 12 项。

总　序

厦　门　大　学　校　长
“厦门大学南强丛书”编委会主任

厦门大学是由著名爱国华侨领袖陈嘉庚先生于1921年创办的，有着厚重的文化底蕴和光荣的传统，是中国近代教育史上第一所由华侨出资创办的高等学府。陈嘉庚先生所处的年代，是中国社会最贫穷、最落后、饱受外侮和欺凌的年代。陈嘉庚先生非常想改变这种状况，他明确提出：中国要变化，关键要提高国人素质，要提高国人素质，关键是要办好教育。基于教育救国的理念，陈嘉庚先生毅然个人倾资创办厦门大学，并明确提出要把厦大建成“南方之强”。陈嘉庚先生以此作为厦大的奋斗目标，蕴涵着他对厦门大学的殷切期望，代表着一代又一代厦门大学师生的志向。

1991年，在厦门大学建校70周年之际，厦门大学出版社出版了首辑“厦门大学南强丛书”，共15部优秀的学术专著，影响极佳，广受赞誉，为70周年校庆献上了一份厚礼。此后，逢五逢十校庆，“厦门大学南强丛书”又相继出版数辑，使得“厦门大学南强丛书”成为厦大的一个学术品牌。值此建校95周年之际，我们再次遴选一批优秀著作出版，这正是全校师生的愿望。入选这批“厦门大学南强丛书”的著作多为本校优势学科、特色学科的前沿研究成果。作者中有院士、资深教授，有全国重点学科的学术带头人，有新近在学界崭露头角的新秀，他们都在各自的学术领域中受到瞩目。这批学术著作的出版，为厦门大学95周年校庆增添了浓郁的学术风采。

至此，“厦门大学南强丛书”已出版了六辑。可以说，每一辑都从一个侧面反映了厦大学人奋斗的足迹和努力的成果，丛书的每一部著作都是厦大发展与进步的一个见证，都是厦大人探索未知、追

求真理、为民谋利、为国争光精神的一种体现。我想这样的一种精神一定会一辑又一辑地传承下去。

大学出版社对大学的教学科研可以起到很重要的推动作用，可以促进它所在大学的整体学术水平的提升。在95年前，厦门大学就把“研究高深学术，养成专门人才，阐扬世界文化”作为自己的三大任务。厦门大学出版社作为厦门大学的有机组成部分，它的目标与大学的发展目标是相一致的。学校一直把出版社作为教学科研的一个重要的支撑条件，在努力提高它的学术出版水平和影响力的过程中，真正使出版社成为厦门大学的一个窗口。“厦门大学南强丛书”的出版汇聚了著作者及厦门大学出版社全体同仁的心血与汗水，为实现厦门大学“两个百年”的奋斗目标做出了一份特有的贡献，我要借此机会表示我由衷的感谢。我不仅期望“厦门大学南强丛书”在国内学术界产生反响，而且更希望其影响被及海外，在世界各地都能看到它的身影。这是我，也是全校师生的共同心愿。

2016年3月

前　言

笔者长期从事数据挖掘研究和教学工作，经历了从最初数据挖掘基础研究的兴起到如今数据挖掘应用百花齐放的时代变迁，深刻体会到研究和应用两者间不可分割的联系：数据挖掘研究源于真实世界中的实际应用需求，用具体的应用数据作为驱动，以方法、工具和系统作为支撑，最终将发现的知识和信息运用到实践中去，从而提供量化的、合理的、可行的，并且能够产生巨大价值的信息。数据挖掘是理论技术和实际应用的完美结合，所以数据挖掘践行者们需要时刻坚定“应用是检验研究的最高标准”这样的理念。

事件挖掘是当下数据挖掘领域的研究热点之一，在真实世界中有着广泛的实际应用需求。计算系统管理是事件挖掘应用的一个主要领域，计算系统仅能从系统产生的事件这一角度进行观测。“事件”通常包含发生时间和相关联的系统状态，系统的事件是时序的，事件的出现通常涉及系统状态的改变。事件通常以日志的方式进行存储，例如业务事务日志、股票交易日志、传感器日志、计算系统日志、http 请求、数据库查询和网络流量数据等，这些事件数据描述了系统状态和随着时间变化的系统行为。面向计算系统的事件挖掘致力于利用以事件为单位加以组织的经验来改善系统自身的性能，或者提高系统管理效率，而所获得的“经验”通常是以数据的形式存在的，要获取并利用事件相关的经验就不可避免地要对大规模历史日志数据、流数据进行分析和挖掘，因此，事件挖掘已逐渐成为计算机系统分析技术的源泉之一。此外，随着计算系统复杂性的日益增加，利用事件挖掘技术来分析计算机系统的要求越来越广泛，越来越迫切，从而使得事件挖掘在计算系统管理领域的重要性越发显著。

事件挖掘应用的另外一个重要领域是网络社交媒体。在网络社区中，事件被定义为“发生在某一特定时间和地点，具有一定影响力的事情”。为了更好地解决网络背景下事件的提取问题，美国国防高级研究计划局

(The Defense Advanced Research Projects Agency,DARPA)提出了“话题检测与跟踪”(Topic Detection and Tracking),即 TDT 任务。但随着研究的深入,发现 TDT 的相关任务已不能适应时代要求,主要原因是仅仅提取话题是远远不够的,更迫切的是需要挖掘事件内部隐含的信息,以及事件发生前的潜在现象以达到预测事件发生、发展和认识事件演化过程的作用,因而新闻事件挖掘、微博事件挖掘、面向社会公共安全的视频事件挖掘等有着广阔的应用前景。

笔者在计算系统异常检测、性能保持、系统自愈、社交媒体分析等方面有着多年的事件挖掘研究积累,因此,经常将研究工作的相关内容和经验制作成课堂讲义在数据挖掘课程上进行讲解和交流。每当走进教室,看到很多学生将从我的主页上下载的课程讲义打印并装订成册,一种强烈的责任感驱使我下定决心将这些积累整理成书。然而,各种繁杂事务的处理耗费了我不少精力,使得写作过程断断续续。但无论如何,这本拙作还是完成了。

本书所采用的编写思路是:以挖掘事件揭示有用的事件模式作为主题,在客观地介绍事件挖掘的基本理论和思想方法的基础上,重点阐述多种类型的事件挖掘算法,并用实验数据和分析结果阐述这些事件挖掘算法的性能,以及各自的应用场景,提倡多种方法兼收并蓄;同时,关注这些事件挖掘方法在各个领域,如系统管理、twitter 等方面的应用,并给出应用案例。

本书的目标群体是研究人员、系统管理人员和对了解事件挖掘的研究现状感兴趣的研究生。当然,本书也能充当高级课程的教科书。事件挖掘方法的学习是有难度的,主要是因为它涉及多个交叉领域,需要熟练地掌握多个领域的知识;相关的文献资料分散在很多种出版物上,例如 ACM (Association for Computing Machinery)的知识发现和数据挖掘国际会议 SIGKDD(Special Interest Group on Knowledge Discovery and Date Mining),IEEE(Institute of Electrical and Electronics Engineers)的数据挖掘国际会议 ICDM(International Conference on Date Mining)、网络运行和管理会议 NOMS(Network Operations and Management Symposium)、网络和服务管理国际会议 CNSM(International Conference on Network and Service Management),IFIP/IEEE 的一体化网络和服务管理国际会议 IM (Symposium on Integrated Network and Service Management)。我们希

望本书能够为不熟悉事件挖掘的读者提供一个好的起点，使得他们能更容易地接受事件挖掘，同时能够为在本领域工作的读者提供全面的参考。

尽管本书的章节大部分都是自包含的，能够按照任意的顺序读这些章节，但是我们还是将它们进行了分组和排序，从而为读者提供一种关于事件挖掘的结构指引。特别地，第 1 章之后的章节被划分为如下三个部分。

第 1 部分：事件生成和系统监控

第 2 章　事件生成：从日志到事件

第 3 章　优化系统监控配置

第 2 部分：模式发现和摘要

第 4 章　事件模式挖掘

第 5 章　时滞挖掘

第 6 章　日志事件摘要

第 3 部分：应用

第 7 章　系统管理中数据驱动的应用

第 8 章　面向 Twitter 流的社交媒体事件摘要

由于个人水平和能力所限，没能做到对本书中所涉及的每一个细节都十分精通。此外，推托为客观的因素，我无法在断断续续的仓促时间内集中精力完成繁多内容的学习理解和组织整理。因此，越是在接近完成本书时，越感诚惶诚恐。所以，恳请读者对书中不妥或者谬误之处给予批评指正，并将意见反馈给我，这真是求之不得的万幸。

作者

2015 年 12 月

目　录

第2部分 模式发现和摘要

第3部分 应用

第1章 引 言

1.1 数据驱动型系统管理

诸如计算系统、物理系统、业务系统和社会系统在内的许多系统都仅能从系统产生的事件的角度间接进行观测。事件被定义为一种现实世界系统状态的体现，并且通常涉及系统状态的改变。本质上，事件是时序的，且经常以日志的方式进行存储，例如业务事务日志、股票交易日志、传感器日志、计算系统日志、HTTP 请求、数据库查询和网络流量数据等。这些事件捕获了随着时间变化的系统状态和系统行为以及它们间的时序关系。

大型复杂系统通常由大量异构组件构成，难以监测、管理和维护。例如，包括云在内的现代分布式计算系统正随着异构软硬件数目的增加而变得越加复杂。不断增加的复杂性也使得以系统失效率为关键指标的系统可靠性不断下降，因此为企业级用户提供一个高性能、可靠、可扩展、可管理的系统并不是那么容易的事。传统的系统管理方法绝大部分都是利用领域专家的知识，通过一个知识获取过程将领域知识转化成系统运行保障的规则、策略和依赖模型。然而，这一过程不仅工作量巨大，代价高昂，易于出错，而且不能满足持续、快速变化的复杂系统环境的需求。例如，对一个大中型企业而言，据估计，其 30%～70%的信息技术资源被用作系统维护。因此，企业迫切需要运用自动化的、高效的方法完成对复杂系统的监控和管理。

为了达成自动化和高效的系统管理目标，IBM(International Business Machines Corporation)作为自主计算(autonomic computing，AC)的创始人试图去构建一个能在尽可能少的人为干预下管理自身的自主系统。自主计算的思想使得科学界和工业界意识到了自动系统管理，并帮助他们引

入了更加复杂和自动化的过程以达到提高生产率以及保证服务和产品质量的目的。为了达到自主计算系统的目标，潜在地假设我们有能力去定义和管理知识库，并能随着环境改变去不断适应。为了具备自我管理能力，系统需要能够自动地监控、刻画和理解系统的行为及其动态性，挖掘事件以发现有用的模式，从历史日志和事件数据中获取所需的知识。

图 1.1 描述了一体化的数据驱动型系统管理框架的体系结构。该框架的主要组件包括：

- 日志数据组织：计算系统中各个组件、主机以及在应用中植入的监控器提供了从计算系统中采集日志数据的能力。日志解析器/适配器，以及事件生成过程可以实现数据采集、数据整合，以及将来自多个异构数据源的数据转换为历史数据的功能。
- 实时分析：实时分析组件负责实时地处理新产生的日志数据，并根据离线分析获得的知识完成在线的管理操作。典型的实时分析技术主要有异常检测、问题确定和故障诊断等。
- 离线分析：离线分析组件负责从历史日志数据中获得知识（例如关联性和依赖性知识），并构建知识库。典型的离线分析技术主要有时序模式发现和摘要。

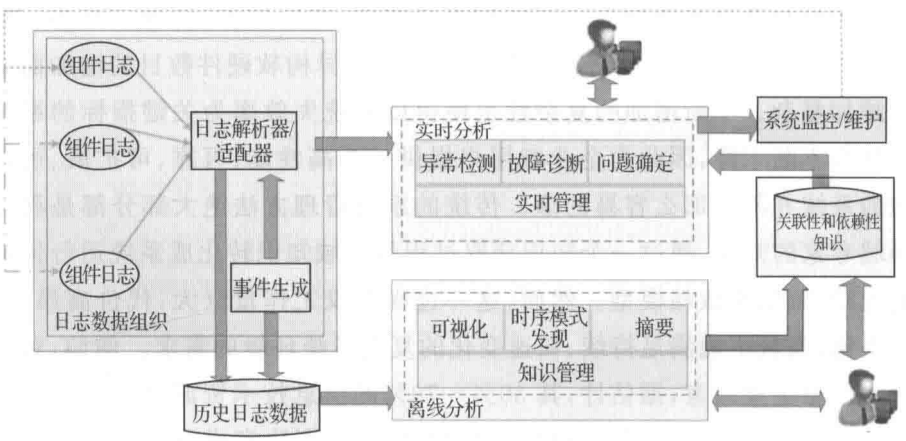


图 1.1　一体化的数据驱动型系统管理框架的体系结构

值得注意的是，系统监控和管理中涉及的任务包括管理员公告、主动数据采集、数据采集器和操作执行器部署或钝化、监控配置变更等。除了将系统管理员从闭环的管理结构中解放出来之外，数据驱动型框架能将领域专家和自主智能技术很好地协同起来，从而为系统管理建立先进和实用

的解决方案。

1.1.1 系统日志数据实例

日志文件中的数据表明每个组件的状态，记录了诸如服务的启动和停止、网络应用发现、软件配置变更和软件执行错误等在内的系统运行状态变化。本节给出了一些系统日志数据的例子。这些日志数据采集自分布式计算系统的各个组件。

- 应用层日志。应用层日志记录了应用程序的行为以及产生的消息。例如，Windows 系统和应用日志、数据库活动日志和 Linux 系统日志。
- 失效数据。失效数据包含系统和应用崩溃以及错误消息。
- 性能数据。性能数据记录了周期性时间间隔下特定组件的性能观测值，如 CPU(Central Processing Unit，中央处理器)每隔 5 min 的使用率。典型的性能指标包括 CPU 使用率、内存使用率、磁盘交换空间利用率、平均负载和平均响应时间等。
- 来自操作人员的报告。这类数据也被称作故障工作单和运维日志，包含由操作人员给出的问题描述。此外，也许还包含可能的原因和失效症状的描述。
- 请求数据。请求数据记录了在系统中处理的请求的时间、主机 ID、用户 ID 和应用等。典型的例子有 Apache 和 Microsoft IIS 的日志。
- 其他数据。其他类型的日志数据有网络流量数据、网络告警日志、程序执行追踪数据等。

1.1.2 数据驱动型系统管理的挑战

一般说来，系统管理包括根本原因分析、异常检测和故障诊断。数据驱动型系统管理的一些主要挑战有：

- 计算系统的异构特性使得管理任务复杂化。一个典型的计算系统包含多种不同类型的设备，如路由器、处理器和适配器等，以及多种不同类型的软件，如操作系统、中间件和用户应用程序等，且它们可能由不同的供应商(如微软、IBM 和思科)提供。计算系统的异构性增加了理解组件之间交互关系和依赖关系的难度。

- 大型复杂计算系统经常在系统失效、系统扰动，甚至是正常运行时出现一些出乎意料的行为。这类系统的复杂性极大地增加了系统管理员理解系统的难度，或者说系统管理员难以确定系统处于何种状态下是需要进行管理的。
- 当前的计算系统是动态的，随着软件和硬件数量的不断增加而快速改变。快速的改变降低了系统可靠性，同时增加了理解系统行为的难度。
- 正确地理解和解释从日志数据中得到的模式是一个巨大的挑战。在系统管理应用中，许多日志数据以时序事件方式生成。采用数据挖掘方法分析时序事件往往集中在发现频繁或者感兴趣的模式上，而所发现的模式可能仅出现在整个日志数据中的一小段时间内。发现事件之间的时序关系对于监控和管理复杂系统是极其重要的。

1.2 本书概览

事件挖掘是一系列从历史事件和日志数据中自动地和高效地获取有价值的知识的技术，其在数据驱动型系统管理领域中扮演了重要角色。本书的目的在于阐述各种事件挖掘方法，及其在计算系统管理领域的应用。特别地，通过提出和改进这些方法来应对 1.1.2 节中提到的挑战。此外，不同的章节研究讨论了数据驱动型框架中的不同组件。面向的读者群体主要包括对基于事件挖掘的系统管理感兴趣的科研人员、系统管理人员以及研究生。当然，本书也可以作为系统管理领域高级课程的教科书。学习事件挖掘是一件有挑战性的事，主要是因为事件挖掘是一个交叉领域，需要熟悉多个研究领域，且相关的文献资料也很分散。我们希望本书能够成为不熟悉事件挖掘的读者的一个好的起点，同时为本领域的研究人员提供一个全面的参考，从而使得事件挖掘变得更加简单易懂。

尽管本书的章节大部分都是自包含的，读者能够按照任意的顺序阅读这些章节。但是我们还是将它们进行了分组和排序，从而为读者提供一种关于事件挖掘的结构指引。特别地，在第 1 章之后的章节被划分为了如下三个部分。

第 1 部分：事件生成和系统监控

第 2 章　事件生成:从日志到事件

第 3 章　优化系统监控配置

第 2 部分:模式发现和摘要

第 4 章　事件模式挖掘

第 5 章　时滞挖掘

第 6 章　日志事件摘要

第 3 部分:应用

第 7 章　系统管理中数据驱动的应用

第 8 章　面向 Twitter 流的社交媒体事件摘要

1.3　本书内容

1.3.1　第 1 部分:事件生成和系统监控

一旦发现系统告警,进行细致的告警分析就需要具备丰富的针对特定类型系统的知识和经验。系统管理员通常不得不对大量的历史系统日志进行分析。日志文件中的数据描绘了每个组件的状态,并记录了系统运行时的变更,例如服务的启动和停止、网络应用发现,软件配置变更和软件执行错误等。系统管理员利用这些数据来理解过去的系统行为,诊断告警的根本原因。系统日志数据的几个特点,如数据报告中格式和内容各不相同且相对较短的文本消息、数据表示中的时序特征、语义信息的匮乏以及词汇量的巨大规模使得非常有必要进行自动化分析。第 2 章主要研究一些能将具有不同格式和内容的日志数据转换为一种标准形式的方法,这个标准形式不仅确保了多个相似字段上的一致性,而且提升了在多个日志文件之间建立相关性的能力。此外,数据组织架构应该除了能适应以标准格式和内容呈现的数据,还应该能适应当前存在的异构数据源。这一章首先回顾了将原始文本型的系统日志预处理为离散的系统事件的三种不同类型的方法,即日志解析器、分类和基于聚类的方法,并阐述了它们的缺点,然后提出了两种基于聚类的从日志数据中生成系统事件的新方法。

自主系统管理和问题检测通常都是通过系统监控软件实现的,如 IBM 的 Tivoli 监控系统和 HP 的 OpenView 系统。大量的研究工作致力