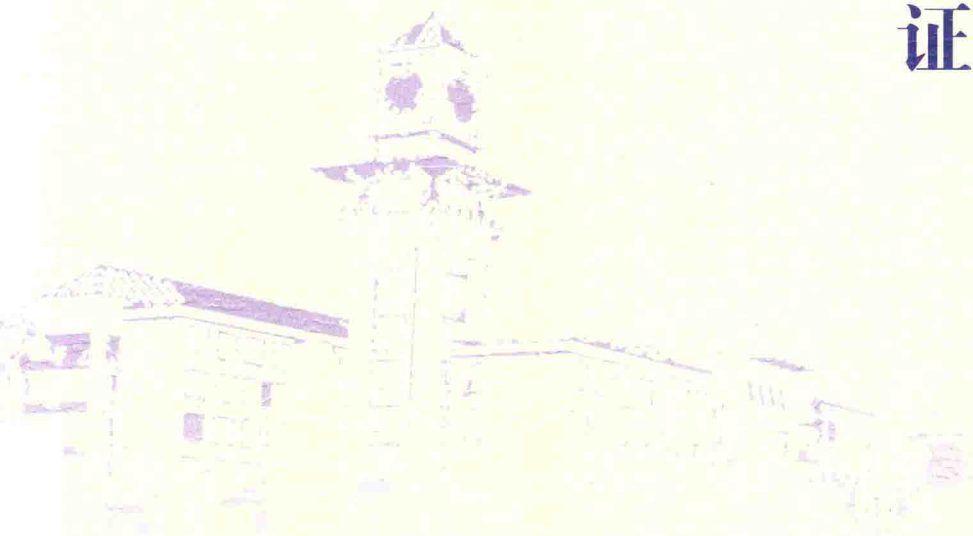
中南财经政法大学
青年学术文库

半结构化数据管理关键算法研究与实证

Large Scale Semi-structured Data Management

张 引○著



中国社会科学出版社

经政法大学
学术文库

半结构化数据管理关键算法研究与实证

Large Scale Semi-structured Data Management

张 引○著

中国社会科学出版社

图书在版编目（CIP）数据

半结构化数据管理关键算法研究与实证／张引著．—北京：中国社会科学出版社，2018.8

（中南财经政法大学青年学术文库）

ISBN 978-7-5203-2505-9

Ⅰ.①半… Ⅱ.①张… Ⅲ.①数据结构②算法分析 Ⅳ.①TP311.12

中国版本图书馆 CIP 数据核字(2018)第 103401 号

出 版 人 赵剑英
责任编辑 徐沐熙
特约编辑 汤浩然
责任校对 李 馨
责任印制 戴 宽

出 版 中国社会科学出版社
社 址 北京鼓楼西大街甲 158 号
邮 编 100720
网 址 http://www.csspw.cn
发 行 部 010-84083685
门 市 部 010-84029450
经 销 新华书店及其他书店

印刷装订 北京君升印刷有限公司
版 次 2018 年 8 月第 1 版
印 次 2018 年 8 月第 1 次印刷

开 本 710×1000 1/16
印 张 14
插 页 2
字 数 176 千字
定 价 40.00 元

凡购买中国社会科学出版社图书，如有质量问题请与本社营销中心联系调换
电话：010-84083683

本书受中南财经政法大学出版基金资助

《中南财经政法大学青年学术文库》
编辑委员会

前　言

随着互联网技术的飞速发展，传统的结构化数据模型已经无法满足人们对信息处理的要求。尤其是在云计算和物联网高速发展的今天，对管理半结构化数据、大规模信息处理等领域的研究越来越多地被关注。由于半结构化数据模型既能描述半结构化数据又能描述结构化数据，而且具有灵活易扩展的存储结构，其已被许多系统和应用作为公共数据模型，被广泛地用于异构数据量大的应用中。如今，几乎所有行业都制定了描述和共享本领域数据的半结构化数据模型应用标准。此外，由于半结构化数据模型具有易于描述结构、易于校验、易于展现等特点，许多原本是以非结构化方式进行存储的数据，也可以通过半结构化数据模型进行描述并存储。

因此，如何对大规模半结构化数据进行有效的管理，在学术界是一个重要的理论研究课题，同时在工业界又是一项具有广阔应用前景的技术。本书以 XML 为代表，探讨了大规模半结构化数据管理中的关键问题——模式提取、节点编码、索引与查询处理等研究课题。本书主要内容如下：

（1）针对现有基于正则表达式的模式提取方法的不足之处，本书根据 XML Schema 规范中元素内容模型的特点，提出了 XTree 算法，该算法可以快速、准确地并发提取多个大规模（GB 级）XML

文档的结构。该算法与其他基于正则表达式的算法最显著的区别在于，XTree 对于元素内容模型的提取加入了对元素内容模型是否有序的区分，降低了算法的时间复杂度和空间复杂度。

（2）针对现有半结构化数据节点编码方案的不足之处，本研究提出了 D2 编码方案，该算法在静态编码和动态编码中都表现出良好的性能，而且易于二进制串行化和反串行化，具有较高的实用价值。和其他半结构化数据节点编码方案相比，D2 编码最显著的特点在于，突破了传统的以整数作为层标识的限制，采用二进制真分数作为层标识，由于真分数的取值区间是无穷的，所以可以保证在任意位置插入节点都存在有效的编码。

（3）本书综合考虑了目前已有的关系型数据库和大规模半结构化数据的索引技术的优缺点，提出一套完善的索引方案——D2 - Index 索引策略，能够支持高效的查询处理。它并不只是使用了一种单一的索引技术，而是参考和借鉴了多种技术，如节点编码索引、结构索引和倒排索引等。D2 - Index 索引策略的最显著之处在于，它的索引文件包括了主索引、路径辅助索引和值辅助索引，这三种索引都采用分块存储的方式提高索引的查找和修改效率。此外，由于是基于 D2 编码方案的，所以 D2 - Index 索引策略可以有效地支持节点的动态更新。

（4）根据目前对于大规模半结构化数据查询处理的研究，本书提出一种以 D2 - Index 索引策略为基础，基于 XPath 表达式的 CAS 查询处理。这种查询处理最大的特点在于，将输入的合法 CAS 语句拆分为多个 BXCAS 语句，再对拆分的语句按顺序进行处理，根据 D2 - Index 策略中的路径和值辅助索引，获取符合查询条件的节点的 D2 物理编码，再从主索引中获取其在源数据中的位置信息，最终以异步的方式输出结果。

目　录

第一章

半结构化数据的应用背景

随着互联网的普及和广泛应用，互联网逐渐成为人们信息交换和资源共享的重要途径。与此同时，随着互联网上信息的规模与日俱增，而且种类日益繁多，尤其是云计算技术的突起，大规模异构数据的存储、转换逐渐成为关注的热点。近 10 年，以 XML（eXtensible Markup Language，可扩展标记语言）[①] 为代表的半结构化数据，在处理数据的存储及转化方面表现出了巨大的优势，半结构化数据已经逐渐成为互联网上表现、传输和转化数据的常用解决方案。但是，半结构化数据具有独特的树形（图）结构特征，这和传统的结构化数据（关系数据模型）有着巨大的差别。因此，如何管理和处理大规模的半结构化数据，是当前研究的热点问题。

本章第一节将主要介绍本书的研究背景；第二节将介绍本书的主要研究内容以及阐明本书的研究意义——研究大规模半结构化数据管理的必要性；第三节将介绍本书的组织结构。

---

① World Wide Web Consortium. Extensible Markup Language (XML) [EB/OL]. Available at: http://www.w3.org/XML/.

第一节　研究背景

半结构化数据，是相对于非结构化数据而言，具有一定的结构性，但是又不如结构化数据，具有严格的理论模型。比较典型的半结构化数据有 OEM（Object exchange Model，对象交换模型）[①]、OIM（Model for Object Integration，对象集成模型）[②]、XML 等。半结构化数据的特点是数据结构的不规则或者不完整[③]，主要表现为数据的模式不固定、结构不明显、模式的信息量大、模式变化快、模式和数据统一存储等。半结构化数据的来源一般分为两种：一种是直接来自半结构化数据源，如 Web 数据、电子文档、电子邮件等，这些都是典型的半结构化数据；另一种是以公共数据模型，在异构数据环境中被引入，用来处理信息的传输、转换和存储[④]。通常人们习惯把半结构化数据称作非结构化数据。但是，并不是所有的非结构化数据都可以用半结构化数据模型进行存储，例如图片、音频、视频等信息就不能。表 1—1 分析了结构化、半结构化和非结构化数据的区别。

---

① C. M. Eastman, Y. - S. Jeong, R. Sacks, I. Kaner, "Exchange Model and Exchange Object Concepts for Implementation of National BIM Standards", Journal of Computing in Civil Engineering, 2010, Vol. 24, No. 1, pp. 25 - 34.

② 张伟业、贺飞、顾明：《基于 OIM 数据对象模型的数据交换系统研究》，《计算机应用研究》2005 年第 11 期。

③ Serge Abiteboul, "Querying Semi - structured Data", In: Foto Afrati, Phokion Kolaities ed. Lecture Notes in Computer Science 1186, Database Theory - ICDT'97. New York: Springer - Verlag, 1997, pp. 1 - 18.

④ 陈滢、王能斌：《半结构化数据查询的处理和优化》，《软件学报》1999 年第 8 期。

表 1—1　　结构化、半结构化和非结构化数据的区别

类型	数据模型	数据与结构的关系	典型的存储与管理方式	典型代表
结构化	二维表	先有结构再有数据	关系型数据库	关系型数据
半结构化	树、图	先有数据再有结构	XML 文档 原生 XML 数据库	OEM OIM XML
非结构化	无	只有数据	内容管理系统	图片 声音 视频

随着互联网技术的飞速发展，传统的结构化数据已经无法满足人们对信息处理的要求。尤其是在云计算和物联网高速发展的今天，对管理半结构化和非结构化数据、大规模信息处理等领域的研究越来越被关注，尤其是对管理半结构化数据的研究。由于半结构化数据模型既能描述半结构化数据又能描述结构化数据，且具有灵活易扩展的存储结构，其已被许多系统和应用作为公共数据模型，被广泛地用于异构数据量大的使用场景中。如今，许多行业都制定了表示和共享本领域数据的半结构化数据模型应用标准。此外，由于半结构化数据模型具有易于描述结构、易于校验、易于展现等特点，许多以非结构化为传统存储方式的数据，也采用半结构化数据模型进行存储。以下就是几种具有代表性的基于半结构化数据模型的应用标准：

（1）XHTML（eXtensible HyperText Markup Language，可扩展超文本置标语言），正在逐渐取代 HTML（HyperText Markup Language，超文本置标语言），成为 Web 页面的标准编写语言。①

① World Wide Web Consortium. XHTML2 Working Group Home Page [EB/OL]. Available at: http://www.w3.org/MarkUp/.

（2）RSS（Really Simple Syndication，简易信息聚合），是一种文件格式，用来描述和同步网站内容，被网站用于发布内容并使其更易于被读者获取。①

（3）SOAP（Simple Object Access Protocol，简单对象访问协议），是一种基于 XML 的消息框架，被用于在 Web 上交换信息。其最大的优点是可扩展，并且可以独立于平台、操作系统、目标模型和编程语言独立实现。②

（4）WSDL（Web Service Description Language，Web 服务描述语言），是一种使用 XML 描述 Web 服务接口的，其定义了 Web 服务可以传输的消息的类型。③

（5）KML（Keyhole Markup Language，Keyhole 标记语言），是 Google（谷歌）旗下的 Keyhole 公司开发的，用来描述和表达地理标记的语言。KML 应用于 Google Earth（谷歌地球）和 Google Map（谷歌地球）等软件中，用于显示地理数据，很多其他的 GIS（Geographic Information System，地理信息系统）相关企业也追随 Google 开始采用此标准进行地理数据的交换。④

（6）ODF（Open Document Format，开放文档格式）和 OOXML（Office Open XML，由微软为其产品 Office 2007 开发的技术规范），这两种格式都是采用基于 XML 的文件格式，用来存储和转换纯文本、电子数据表格和图表等传统的非结构化数据。⑤

（7）其他专业领域也有许多采用非结构化数据模型进行标准化的例子，如 DocBook XML，用于编撰书籍和文档，尤其是技术

---

① Kevin Howard Goldberg. *XML, Second Edition* [M]. Peachpit Press, 2009.

② Airi Salminen, Frank Tompa. *Communicating with XML*. New York: Springer, 2011.

③ Kevin Howard Goldberg. *XML, Second Edition*. Peachpit Press, 2009.

④ Josie Wernecke. *The KML Handbook: Geographic Visualization for the Web*. Boston: Addison - Wesley Professional, 2008.

⑤ Kevin Howard Goldberg. *XML, Second Edition*. Peachpit Press, 2009.

性文档[①]；SVG（Scalable Vector Graphics，可缩放矢量图形），是图形图像标记语言，用于描述二维矢量图形[②]；CDA（Clinical Document Architecture，临床文档架构），旨在规定用于交换的临床文档的编码、结构和语义[③]；CML（Chemical Markup Language，化学标记语言），用于描述化学分子、化学反应、光谱等化学数据的标记语言[④]；MathML（Mathematical Markup Language，数学置标语言），用于在互联网上书写数学符号和公式的置标语言[⑤]；MusicXML（Music Extensible Markup Language 音乐扩展标记语言）是一个开放的基于 XML 的音乐符号的文件格式，用来作为乐谱信息存储和交换格式[⑥]。

在上述背景下，随着半结构化数据的普及和广泛应用，半结构化数据模型不仅用来存储半结构化数据，越来越多的领域也开始使用半结构化数据模型作为标准用于存储和描述传统的结构化和非结构化数据。由于半结构化数据不同于传统的结构化关系型数据，因此，如何高效而合理地存储、分析、索引和查询大规模半结构化数据的问题就随之产生。目前，比较主流的管理半结构化数据的方式有两种，基于传统关系数据库和原生 XML 数据库。但是，前者由于只是对关系数据库的扩展，如果要利用其管理半结构化数据，就

---

① OASIS. The DocBook Schema Working Draft [EB/OL]. Available at: http://www.oasis-open.org/docbook/specs/.

② World Wide Web Consortium. Scalable Vector Graphics (SVG) [EB/OL]. Available at: http://www.w3.org/Graphics/SVG/.

③ Health Level 7. Clinical Document Architecture [EB/OL]. Available at: http://hl7book.net/index.php?title=CDA/.

④ Murray-Rust P., Rzepa H. S. Chemical Markup Language [EB/OL]. Available at: http://cml.sourceforge.net/.

⑤ World Wide Web Consortium. W3C Math Home [EB/OL]. Available at: http://www.w3.org/Math/.

⑥ Recordare LLC. MusicXML™ [EB/OL]. Available at: http://www.musicxml.org/xml.html/.

必须将数据导入关系数据库，以关系数据库的方式进行存储和管理，因此在功能、效率和性能上均不够理想；后者管理的对象大多仅限于以数据为中心的半结构化数据，无法应对管理以内容为中心的半结构化数据的功能需求。针对上述问题，本书以大规模半结构化数据（包括以数据为中心和以文档为中心两种类型）的模式提取、节点编码、索引与查询处理关键算法为研究内容，提出一套可并发提取模式、节点编码、索引和查询处理多个同结构大规模半结构化数据的理论基础和解决方案。

第二节　研究内容及意义

一　研究内容

一般而言，数据管理的核心问题研究是构建一个完善的、可以快速有效地根据条件获取目标数据集的系统。根据大规模半结构化数据的特点，本书认为半结构化数据的模式提取、节点编码、建立索引与查询处理，是管理大规模半结构化数据的基础和关键性问题。提取半结构化数据的模式，可以解析半结构化数据的结构特点，是查询处理中构建路径查询表达式的基础。对半结构化数据进行节点编码，不仅是为半结构化数据建立索引的核心问题，而且在管理以数据为中心的半结构化数据时，还是压缩存储大规模半结构化数据的关键所在。在传统的关系型数据库中，建立索引可以提高数据的访问效率，同时对于大规模半结构化数据的管理，也可以通过建立索引，优化查询实现。无论是何种类型的数据管理，查询处理始终是研究的重点。

因此，本书以大规模半结构化数据模式提取、节点编码、索引和查询处理为基础，研究大规模半结构化数据的模式提取，大规模