SSMPF

数据索引与数据组织模型及其应用

胡运发 著

复旦大學出版社

SSMPF 本书已获上

数据索引与数据组织模型及其应用

胡运发 著

复旦大學出版社

图书在版编目(CIP)数据

数据索引与数据组织模型及其应用/胡运发著. —上海:复旦大学出版社,2012.7
ISBN 978-7-309-08694-2

Ⅰ. 数… Ⅱ. 胡… Ⅲ. 索引组织 Ⅳ. TP311.12

中国版本图书馆 CIP 数据核字(2012)第 007068 号

数据索引与数据组织模型及其应用
胡运发 著
责任编辑/黄 乐

复旦大学出版社有限公司出版发行
上海市国权路 579 号 邮编:200433
网址:fupnet@fudanpress.com http://www.fudanpress.com
门市零售:86-21-65642857 团体订购:86-21-65118853
外埠邮购:86-21-65109143
大丰市科星印刷有限责任公司

开本 787×1092 1/16 印张 13.75 字数 335 千
2012 年 7 月第 1 版第 1 次印刷

ISBN 978-7-309-08694-2/T·442
定价:30.00 元

自序

本书属于数据与知识工程范畴的学术专著，是我和我的学生们过去20多年教学科研成果的总结与提升。本书一个显著特点是面向实际应用，特别是面向海量数据的处理。我们从数据组织的观点出发来处理数据索引问题，提出一种新型的数据组织的商空间模型——互关联后继索引模型。从粒子计算理论或数学变换角度论证该模型的优越于现有经典索引模型的多种特性、保序性、保假性、压缩性等。同时导出多种优越能力：①log级的快速查询能力；②原文生成能力；③高度的压缩能力；④随机查询能力等。

本书以互关联后继索引模型为中心，与多种领域的核心问题相结合，说明该模型如何在数据压缩、全文数据库、关系数据库、Web数据库、演绎数据库、逻辑推理(知识库)、文本信息隐藏等领域产生创新性的影响，并由此导出一系列新成果。对于从事关系数据库、Web数据库、事务库、演绎数据库、知识库、逻辑语言、搜索引擎、数据压缩、文本信息隐藏、数据挖掘等领域的研究人员、工程技术人员、高等教育的教师与研究生有重要的参考价值或借鉴作用。每一章都曾经或即将成为一个研究课题，如融会贯通、举一反三，会发掘更多的研究领域，如数字云文本的分类聚类、决策分析、算法复杂性与程序复杂性的关系等具有深刻理论价值与广泛应用前景的领域。总之一句话，本书的成果对未来仍具有巨大的展望空间。出版本书的目的是为了未来对此书感兴趣的青年朋友，希望能在他们的手中进一步开花结果。我对他(她)们最想说的一句话是：创新性科研成果不是急功近利的产物，而是知识、实验、兴趣、长期思考与不懈坚持的结果。

20世纪80年代中期以后，我开始研究数据与知识工程。最早做逻辑语言、知识库的教

学与研究，后来做演绎数据库和全文数据库的教学与研究，再后来做事务库挖掘、Web 数据库和搜索引擎的教学与研究。在长期教学科研中发现有一个核心问题，那就是对于海量数据来说有一个数据的组织问题或者索引问题。一般说来，不同的领域有不同的数据组织方式或不同的索引方式。数据库需要 B+ 树索引支持，全文数据库需要倒排表或 Pat 数组索引，Web 数据库需要某种树形索引的支持。有一些问题一直困扰着我，例如，这些索引之间有什么本质差异？能否在一定的框架上将它们统一起来？为什么关系数据库的发展会如此快速、应用会如此广泛，而理念更加先进的演绎数据库或知识库却停留在襁褓之中？如此等等，这说明需要一种新的探索。人们经常说一句非常概括性的话是“关系数据库 = 文件系统 + 索引”，我想问题也许出在索引之中，因为演绎数据库或知识库几乎谈不上有什么有效的索引。我正是从索引开始我的探索之路。

严格来说数据的组织和数据索引是两个不同的问题。数据组织说的是如何将原始数据重新加以组织，然后被有效地利用，但组织过程中，不能改变数据本身的语义和结构性质。数据的索引虽然需要对数据加以组织，但可以简化数据，也不要求一定要保持数据的结构性质，只要能帮助更快地查询到需要的内容就行。例如，一本书的目录，一座图书馆的书目索引，它们都是索引，但它们不能代替一本书内容本身，也不能代替图书馆所有藏书中的信息。在一定条件下，能否将数据索引看成数据组织呢？初看起来，如果把数据组织当作索引，会不会把索引弄得过分复杂或过分庞大呢？如果处理得当，或许可以破解这个难题。沿着这样的思路，我们提出互关联后继索引模型，其基本思想是用组织后数据空间（在下篇，我们要从理论上说明那就是商空间）坐标代替数据本身，用商空间中区间表达商空间粒子符号，用商空间的另外一些坐标来表达符号间的结构信息——后继关系。这样一来，不但没有增加数据组织的负担，反而有可能减少数据组织的复杂性，数据索引就等价于数据组织本身了。按这条技术路线提出的索引模型称为互关联后继索引模型，命名的原因是直观的，符号 a 的后继可能是 b，而 b 的后继可能是 a，符号间的后继是互相关联的。

我选择解决这个问题的突破点是全文数据索引。选择这一点纯属偶然，那是因为在 1994 年，施伯乐教授让我有机会参加上海图书馆一个全文数据库的课题，出乎我预料的是全文数据库的索引问题也是一个困难的课题。要想找到又快又小的索引，并非易事。2003 年北大方正给我提出的关系数据与全文数据的协同查询难题，更激发我研究的兴趣。

从这项技术的研发历史来看，形成互关联后继索引模型中相关的一些技术，的确经历了较为漫长的时间。1995 年我和我的学生实现了倒排表索引 golomb 压缩模型，开启了我研究索引模型的工作；1999 年，我提出一种改进的索引方式，当时叫做 Σ^2 索引模型，它是对

倒排表的一种改进，计算机实验的效果高于倒排表。这使我兴奋不已，因为毕竟按我的直觉，事情有了良好的开始。2001 年我又提出互关联后继树索引模型，它是对Σ^2 索引模型的改进。在我的一部分学生的支持下，不间断地实验与改进，一直做到 2008 年。通过逐步改进，反复试验，不间断的求新变异，直到 2008 年，差不多花去十年时间，我们才分别提出后继编码大小有序、后继字符有序、后继字符与后继编码大小双有序的互关联后继树、倒互关联后继索引、直接后继树与间接后继树等多种索引模型，并在原文生成、查询功能、查询效率和索引大小等方面取得优于现有索引模型的结果。

与此同时，我们也开始互关联后继索引模型的应用研究，例如，把互关联后继索引模型用于时间序列分析、关系数据库与全文数据库的协同查询；用于全文与 XML 数据库联合查询；用于当时较为热门的数据挖掘以及结构索引的研究，甚至数据压缩和基因数据分析等方面的研究。2006 年，我们曾和中国电信公司合作进行搜索引擎系统研究。理论分析和实验数据表明，在大多数情况下，互关联后继索引均取得明显的效果。当然，这是一个循序渐进的过程，而且是我和我的学生们共同努力的结果。这些成果体现在已经发表的诸多论文之中，详情可参看本书附属的参考文献。我把其中少量基础性成果选编（如代表性互关联后继索引的生成与查询算法、少数领域的应用）作为上篇，目的有三个方面：①真实地再现我们的科研过程；②此中包含大量相关知识背景和互关联后继索引模型的基本概念、性质和基本算法、相关的实验对比图表和一些初步的应用，便于理解本书的其他的所有结果；③体现我和我的学生们之间教学相长的过程，一般说来由我提出相关领域和相关算法思想，而由我的学生们加以实现和验证。但有些时候，学生们也提出某些创新性的构想，如倒互关联后继、更好的互关联后继的创建、查询算法等。

上篇中包含的基本成果，是本书下篇进一步深化的基础。可以说没有上篇就没有下篇，本书主要内容虽然是下篇，但上篇仍然是我进行理论上思考与提炼的基础。如果不把上篇的成果在一定程度上加以展现，那么下篇就显得是无源之水，一些理论方法的提出也显得有点唐突。更加重要的是，如果说下篇有一点理论上的升华的话，那也是来源于上篇科研过程的经验积累和不停顿思索。没有上篇经历的过程，那么我的脑子里就没有问题；即使有了问题，也没有解决问题思路和解决问题途径。

上篇反映的成果，大多是基于直觉的，属于方法上的改进，并没有上升到理论的高度。我们称新的索引模型为互关联后继索引模型，也是基于直觉的。在下篇我们仍然沿用这样的称谓，目的是保持概念的连贯性。其实，互关联后继索引模型是数据的索引模型，但它同时也是一种数据的组织模型。在本书中，凡是提到互关联后继索引模型的地方，都是指具

有数据组织功能的数据索引。

我和我的学生们在2008年之前的有代表性的成果涉及15个方面:①后继字符有序互关联后继索引模型创建与查询算法(申展);②互关联后继索引模型区间查询算法(王政华);③后继编号有序与字符有序(双排序)的互关联后继索引模型创建与查询算法(袁天宇);④双排序互关联后继树的正反向二分查询算法(李卓尔);⑤双排序互关联后继索引二分加验证查询算法(杨茹);⑥互关联后继索引的编码优化(曹小冲);⑦基于互关联后继树的文本压缩(申小霞);⑧基于互关联后继的XML与全文协同查询(王竟原);⑨基于互关联后继模型的搜索引擎(王正刚);⑩基于互关联后继树的时间序列分析(曾海泉);⑪基于互关联后继树的结构索引ISTR(k)(范颖捷);⑫基于间接后继的频繁项挖掘算法(马海兵);⑬基于后继路径的频繁项挖掘(李卓尔);⑭基于互关联后继的关系库与全文的协同查询(王竟原);⑮基于互关联后继的生物基因挖掘算法(陈祎)。以上括号内的作者,均是我指导下的硕士生或博士生。应当说,这份名单仅是其中的一部分,名单中还应包括朱立、张锦、周水庚、陶晓鹏、孙敬宇、刘永丹、周逸群、唐洋运、马科、颜文伟、杨传耀、杨啸天、匡月、王鑫印、蔡喁等。上述所列名单(包括成果)也仅是我们成果的一部分,大约只占全部成果的一半左右。这些成果概括起来有两个方面:第一方面是针对序列的数据(如全文)或可交换的集合数据(如事务库),研究如何创建和查询各种互关联后继索引;第二方面是将互关联后继索引模型应用到不同的领域。第一方面是基础性的,包括概念、方法思想、算法和实验验证。我从其中选择基础性的几篇,通过整理、合并与修改构成三章。第二方面的成果也选择其中三篇:XML数据库、数据压缩与搜索引擎。两者合起来构成本书的上篇。未选入本书成果,也具有一定的重要性,例如,“基于间接后继树的频繁项挖掘算法”,首次提出间接后继的概念和用于挖掘的方法;同样,“基于后继路径的频繁项挖掘”,实质上提出了隐式后继的概念与方法。在本书的下篇,我准备在这样技术方向上做出更多的改进。“基于互关联后继的时间序列分析”、“基于互关联后继树的结构索引ISTR(k)”、“基于互关联后继的生物基因挖掘算法”等都给出不同领域的研究方向,值得作出更多关注或研究。

上篇的内容如下:第一章,第一后继字符有序的互关联后继索引的创建与区间查询算法。第一后继字符有序的互关联后继索引是相对简单的互关联后继索引模型,特点是只保证单一的后继字符有序,不要求后继编码有序。本章介绍相关的概念,创建算法与多种查询算法。该模型算法简单,但查询算法效率远高于倒排表索引,在原文存储于外存条件下,还高于Pat数组索引模型。本章给出相关的复杂性分析和实验结果。第二章,双排序互关联后继索引创建与查询算法。双排序互关联后继索引是相对复杂的互关联后继索引模型,

特点是要求同时保证第一后继字符和第二后继编码同时有序，我们有时简称双有序互关联后继索引模型。本章介绍相关的概念，创建算法与多种查询算法。在所有的索引模型中，该索引查询效率最高，本章后面给出相关的复杂性分析和实验结果。第三章，互关联后继树索引模型的编码优化方法。这一章主要研究一种压缩互关联后继树索引的方法。方法是采用与申农熵互补的后继熵。实验证明该方法对索引有明显的压缩效果，远小于倒排表或 Pat 数组索引。第四章，基于互关联后继的文本压缩。利用互关联后继索引一些特性：如快速创建、快速查询、独有的任意最长串或其子串匹配能力与随机查询能力等，产生一种优于 LZW 的文本压缩算法，给出相关算法复杂性分析和相关的实验结果。第五章，基于后继模式树的 XML 索引模型。树形结构和全文结构伴随出现的 XML 文档索引问题是一个公认的难题。本章采用倒向互关联后继树与全文联合索引模型，使得 XML 与全文的协同查询索引模型统一且不需要语义转换。实验表明：XML 与全文的协同查询的效率要大大高于基于常规索引模型的 SQL Server 2005 的查询效率。第六章，基于互关联后继树的搜索引擎。本章立足于互关联后继树全文模型，设计并实现互关联后继树搜索引擎的若干关键技术，主要包括基于互关联后继模型本体语义词索引与汉字混合序列的索引创建、更新与查询算法；基于互关联后继的汉字切分算法等核心技术以及包括匹配度计算，全文与关系数据库的协同查询，搜索引擎排序技术等。在匹配度计算方面，提出了两个通用的公式来表示其计算方式，分析和实验表明，其在完全匹配和部分匹配上都工作得很好。在排序方面，提出了动态划分的多权值部分排序算法，减少了互关联后继树返回结果的排序时间，使其在平均情况下为线性时间复杂度。

在 2008 年前的成果的基础上，我产生下面的想法，这些看起来不同数据对象(全文、事务库、关系库、知识库)的数据组织方式能够在多大程度上统一起来呢？为此，我在 2004 年提出一项自然科学基金项目“适合多种数据类型的索引模型的研究”的申请，以及 2009 年我参加国家自然科学基金重点课题“文本内容安全研究”，均得到国家自然科学基金委员会的批准，这给了我很大的激励。我要进一步研究的问题涉及：①互关联后继索引有理论基础吗？如果有，其理论基础是什么？②互关联后继模型有其他现有的索引模型有什么关系，究竟哪一种模型更好一点？③作为一种数据的组织方式，互关联后继索引有熵吗？其熵有多大？与申农熵有什么关系？我差不多花了 3 至 4 年的时间，回答了上述问题。但是，还有一些问题，如④互关联后继索引模型能否用于事务库、关系数据库、演绎数据库等领域的数据组织？利用互关联后继索引模型能否提高数据挖掘、关系查询、关系演算效率？这些问题一直萦绕在我的脑海中，经常是日有所思、夜有所梦，有时竟回到思考的原点，无果而终。

但由于思考多了，有时也会突发奇想，然后坐下来，深入地分析与对比，竟有所斩获。有的经过检验，并不成立，于是被抛弃了；有的虽有优点，但并不完善。总之在思考中，我深感其中的乐趣，“思中有乐，乐能忘疲”，多思必有果。

其实，在2008年以前的成果中我们已经涉及两类数据类型，一类是严格有序的，如全文数据等；另一类不是严格有序的，如像事务库那样的集合数据。在树形数据结构中，同一路径中的节点数据，有的不可以交换，有的却可以交换等。我们已经提出用间接后继的思想来处理事务库数据、概念格挖掘问题。不过，当时有点就事论事，只是后来联想到要解决关系数据库或演绎数据库的数据组织问题时，忽然觉得眼前一亮：它们不也是集合型数据吗？能否用间接后继的思想来解决相关难题呢？经过反复思考终于发现，如用间接后继的思路，看待关系与演绎数据的难题，主要问题就已经解决，剩下的问题只是一些技术性难题而已。主要技术性难点有两点：一是我们虽然提出间接后继的思路，但方法不够精致，需要提出一种更为精致的间接后继索引结构；二是关系库演绎库数据更为复杂，要找到适应这种复杂情况的途径。后来发现，这种复杂性仅是表面现象，当我提出复合项的概念时，一切就迎刃而解了。把关系库中复合项与事务库中的简单项对应起来，处理事务库的方法同样可处理关系库等复杂情形。等这个问题有明确结论以后，解决第5个问题就相对容易一些，即⑤互关联后继索引模型是否能够成为知识库等领域的数据组织模型？利用互关联后继索引模型能否提高知识推理的效率？因为，解决关系数据库的经验使我马上有了联想：既然互关联间接后继的思想能解决集合类型的数据组织问题，当然也能解决集合类型数据的逻辑演算问题；但是，严格有序的逻辑程序是否能用严格有序的互关联后继索引加以组织呢？我的脑海中有了肯定的答复，在克服一些技术障碍之后，这个问题基本上也有了解决方案。

下篇就是2008年以后一段时间研究的结果，共分七章。主要内容包括：第七章，互关联后继索引的商空间理论。本章主要研究互关联后继树的商空间变换。互关联后继索引实质上是一种从数据的原空间向数据的商空间变换，商空间变换具有保序性和保假性以及其他一些优良的性质，它们是互关联后继索引模型在多方面表现突出的理论基础。第八章，互关联后继索引与其他索引模型的关系。本章研究互关联后继的数学变换方法。过去一些著名的索引模型，都是经过某种程度的数学变换出来的，只是它们没有互关联后继索引变换那么多，那么彻底。数学变换的程度能够说明诸多索引模型的关系与各自的优劣。第九章，互关联后继索引模型的熵与压缩。互关联后继索引有自己的熵，我们论证了二元互关联后继索引的熵与申农熵的互补性。提出差异熵的概念与差异熵编码方法，用实例说明差异熵编码的高度压缩能力。最终表明，互关联索引模型也是一种通用压缩模型——是一

种将原文与它的索引同时压缩的通用模型。互关联直接后继模型适合有严格顺序关系的数据类型;互关联间接后继模型适合于对顺序要求不高的集合数据类型。第十章,事务库的组织与数据挖掘。本章提出一种基于互关联隐式间接后继树的数据挖掘方法。采用深度优先的次序,对树形结构进行区间编码,利用区间的包含关系判断节点间的后继关系;然后将树结构数据向有序字符序列上投影,得到一种称为 T-Istr$^+$ 的隐式间接互关联后继索引表。挖掘是在该表上进行,挖掘方法的优点是:①静态挖掘索引:只建一棵固定不变的 T-Istr$^+$ 索引表;无需像韩加威算法那样不间断创建动态频度树;②利用简单项符号的组合、区间包含判别与保假性实现快速挖掘;③本索引方法既可用于事务库的挖掘,也可用于事务库的查询。第十一章,我们还用隐式间接互关联模型进一步研究关系库、演绎数据库的索引模型。所谓隐式,就是要借助某种判定方法就能确定某节点是否是另一节点的后继关系。隐式间接互关联模型帮助我们开辟了关于事务库、关系库、演绎库查询方法的新领域。此外,我们的实验证明了 B+ 树与互关联后继树在索引级的协同查询的效率,要高于一般数据库(如 SQL Serve 等商业关系数据库)的协同查询效率(效率要高于一个数量级以上)。第十二章,逻辑程序或知识库数据的组织与索引。本章主要研究逻辑推理的索引方法,借助互关联后继索引模型,我们发现有一种叫做索引合一的方法可代替逻辑推理中的合一算法。使用索引合一的方法,产生一种与子句顺序无关,无需回溯的新的推理方法,无需回溯的特性显著地提高了推理效率。由此一种事先未曾料想的结果出现了,互关联后继索引模型不仅提高了推理效率,而且竟然为子句头的或并行、子句体的与并行、合一的与并行提供了强有力的支持、甚至提供了像部分匹配等在逻辑程序历史上未曾有过的新功能。第十三章,研究文本类型媒体的信息隐藏技术。此前,人们已在图像领域的信息隐藏技术取得重要的进展,但是在文本领域却进展缓慢。直到现在,大家公认的看法是对于文本对象来说,冗余度小、隐藏信息量小、隐藏信息容易受到攻击,因此文本媒体的信息隐藏缺乏突破性的进展。本章的技术路线是将互关联后继数据模型的数据组织技术、文本生成控制技术、Rsa 密钥技术和随机变换的几项技术紧密地结合起来,形成一种安全性强、鲁棒性好、隐蔽性高、信息隐藏量大、安全可控性强等优点的文本信息隐藏系统。本章以文本熵作为弱数字水印,以文本特征向量组作为强数字水印,以载体的数字模板做密钥代表隐藏信息。由于数字模板的随机性以及受到加密技术的保护,本项技术具有足够的安全性。互关联后继模型中索引创建算法、原文生成算法和信息隐藏算法、信息提取算法有着异曲同工之处。

尽管如此,整个下篇大部分术语与成果只是初步的,所考察的问题大多是开放的。上篇的章节,叙述的互关联后继索引模型大多比较具体化、直观化、更加面向需求,除了具体

算法细节外，比较容易理解。具体算法是为了那些关心系统实现的人们，一般的读者不要太拘泥于这些细节。上篇也有一定的局限性，有时会妨碍对一般性的理解。下篇的章节，大多比较抽象化、形式化，比较不易理解，但更具有一般性。上下篇最好互相参照。另外，互关联后继索引模型的创建，有一个发展过程，期间我们陆续发表多篇论文，本书的参考文献中基本上列出了所有相关论文的标题与出处，如能把本书与相关参考文献互相参照，相信能够帮助加深理解，并能收到举一反三的效果。理解术语、性质、算法、编码等的最好方法是用一些例子进行演练，如能发现问题，那是阅读的深化；如能提出不同见解，那是思想的升华；如能证明见解更有效，那就是智慧的创新了。

我的同事于玉教授、葛家翔教授、张成洪教授、陶晓鹏副教授、朱洁副教授、沈瑶英副教授、陈彤兵博士长期参与相关学术讨论并协助我指导研究生工作，张成洪教授还承担上篇第二章、第六章与下篇的第八章等章节的撰写工作，在此一并对他们表示感谢。

仅以本书向支持我多年研究的国家自然科学基金委员会和家人致谢！

胡运发

2011 年 8 月于复旦大学

目录

上　篇

下 篇