

智能数据库系统:研究综述*)

车敦仁 周立柱

(清华大学计算机系 北京 100084)

摘 要 In this paper we proceed from the medium attribute of information and expound the three dimensional characters of intelligent database system, that are database feature, AI function and multimedia support. Thus we divide intelligent database systems into two generations: the first generation is simply the integration of database system with AI functions, and the second one is the incorporation of three dimensional characters, DB-feature, AI-function and multimedia-support.

关键词 Next generation database, Intelligent database, Object-Oriented database.

1 引言

今天,更先进的信息系统开发平台和与之相适应的开发方法学,是信息高速公路的基本配套设施之一,目前的信息系统开发平台的 DBMS 产品却只是处理数据,尽管最近几年受到重视和广泛研究的 KBS 较注重知识的处理,但无论 DBMS 还是 KBS 都远未能做到象人那样按信息的真正涵义直接处理信息。此外,多媒体特征是信息的基本属性之一,真正意义下的信息处理系统应该把对信息的多媒体支持纳入进来。计算机科学的一系列研究领域,特别是数据库技术,软件工程,专家系统和 AI 等,已分别从不同的角度和方面为信息处理技术和信息系统开发做出了贡献。目前,各领域技术相对成熟,为深层次的综合集成研究提供了技术条件。智能数据库系统(IDB)概念的提出,正是试图把 AI 与 DB 进行集成的结果。

关于 IDB 还没有一个比较公认的定义,现有文献对 IDB 的解释一般都过于简单、直观:即把 AI 技术应用于 DBS,以提高 DBMS 的已有功能(Activitie)的性能,并扩充新的功能。笔者曾在文[13]中把 IDB 的这种“DB+AI”式的朴素解释称之为 NIDB(Naive Intelligent DataBase),而把进一步集成了多媒体支持的 IDB 称为 TIDB(Truly intelligent Data Base)。在本文我们则把这两者划分为两代系统:前者称为第一代系统,后者称为第二代系统。

第二代智能数据库系统旨在研究和开发更为广

义的(真正的)信息管理系统,所要管理的信息应更加接近本来意义上的信息概念,而不只是限于数据信息(基于 DBMS 的 MIS)。下面拟阐述信息的四个性质,从侧面来把握信息概念的本质。

(1)信息的效应性或有有用性:是指从人类的观点看,信息被(不同的)人所接收后会产生一定的效应或发挥不同的价值。研究信息的效应性会对软件和计算机信息系统的友好用户界面之开发有指导作用。

(2)信息的内容性:是指从语言学角度看,信息总是有所指(Meaning),有其内涵。信息内涵一般是无法直接和完全表达出来的。

(3)信息的媒体(Medium)性:是指从物理的角度看,信息总是有一定的物理表现形式(Physical manifestation),即媒体性。媒体还是信息存在的载体(介质),信息不能离开媒体而存在。

(4)信息的可操作性:信息可接受的操作有:create(创建),transmit(传达),store(存储),retrieve(检索),receive(接收),copy(拷贝),process(处理),destroy(销毁)等,此外还有 transform(变换)操作,但信息变换难免会造成信息丢失。

按照面向对象观点,一个信息系统是由一集相互作用,高度并发和自治的对象组成,它们当中有些是主动的,有些是被动的,并且各自可具有不同程度的持久性。对象作为信息载体,较之数据和知识向着信息概念的本质方面更靠近了一步。

下面分别讨论第一代系统和第二代系统的概念

*)国家自然科学基金资助项目。车敦仁 博士后、主要从事新一代数据库的开发与研究。

和特点。

2 第一代系统述评

2.1 专家系统 ES

第一个专家系统 DENDRAL 在 60 年代后期间世于斯坦福大学的实验室之中^[2]。ES 较传统 AI 系统的研究有一个战略性的重点转移,强调运用知识求解特定问题。

一个专家系统是一个(或一组)程序,通过对被编码了的知识的运用和操作来完成对特定领域里较为复杂的问题的求解,被求解的问题通常必须是由领域专家依赖其专家知识才能求解的较为复杂的问题。ES 具有与传统计算机系统截然不同的下列特点:

(1)ES 的问题求解能力依赖于所拥有的专家知识。

(2)对 ES 的知识进行编码和维护应独立于(应用知识的)控制程序来进行。这样做有两个好处:知识具有一定的独立性,控制程序可单独抽取出来予以推广,从而形成专家系统外壳,可用于建立更多的 ES。

(3)ES 具有对问题求解结果进行解释的能力,使用户可对之评价从而提高用户对系统的信任程度。

(4)ES 中的知识一般是逻辑上分层的,推理过程是在元知识的指导下对比它低一层的知识进行操作来完成。

2.2 演绎数据库 DDB

DB 与 AI 是计算机科学中非常重要的两个独立的研究领域,研究目标和基本原理各不相同。DBMS 追求的目标是,满足现代信息处理社会对日益增长的海量信息进行高效维护和管理需要;而 AI 研究则追求“使计算机具有象人一样的行为”。与此同时,AI 研究者还指望借此获得对人类智能本质的更深入的认识^[3]。AI 研究在策略上往往瞄准某一特定的应用。与此不同,DB 研究则注重探索数据管理的较通用的方法。由于应用的纷繁多样,因而 AI 研究所取得的成果五彩缤纷,单就已开发的表示方法,如语义网、框架和脚本等就各不相同;DB 研究所获得的成果是相对集中的通用数据模型,如关系数据模型 RDM,层次数据模型 HDM 和网状数据模型 NDM 等,以及早已标准化了的查询语言 SQL。于是,DB 中的少量通用数据模型与 AI 中的不同知识表示机制形成了鲜明的对比。

从概念上,ILDB 呈现图 1 所示的两层功能结构:

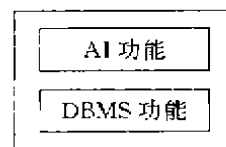


图 1 ILDB 的概念结构

把 AI 功能集成到 DBMS 之中有两种基本途径^[4]:一种是基于逻辑的方法,另一种是专家系统方法。第一种途径得到演绎数据库 DDB(Deductive DB),第二种途径得到专家数据库系统 EDBS。

演绎数据库系统是这样一种 DBMS,由于在其中引入了逻辑推理,从而使之除了具有传统 DBMS 的全部功能和特征之外,还可进行演绎推理,即可以从库中显式表示的数据推演出未曾直接存入的新数据或新信息,这些新数据或新信息实际上被隐含在显式表示的数据之中。自然地,从概念上将 DDB 系统的数据库划分为外延数据库(即传统意义下的 DB)和内涵数据库(即未被显式存储的隐含数据库)。

显然,上述定义并未要求 DDB 必须基于某一种特定的 DBMS(如 RDBMS),但是由于演绎推理与 RDM 基于相同的数学背景,即一阶逻辑^[5],关系数据库的数据检索(关系演算)事实上可以被看作逻辑演绎的一种特例,因而演绎功能就成为对关系数据模型的一种极自然的扩充,于是基于 RDBMS 扩充演绎推理也就占据了至今为止 DDB 研究中的主流。

2.3 专家数据库系统 EDBS

专家数据库系统是一个(或多个)专家系统和一个(或多个)DBMS 的组合系统,目的在于满足那些需要在知识指导下进行共享信息处理。ES 负责推理,DBMS 负责共享与保护。

EDBS 与 DDB,特别是与基于关系模型的 DDB,有所不同,DDB 是一个同质(Homogeneous)系统,EDBS 则是一个由多个不同组件形成的协作型系统。在 EDBS 中由于存在多系统协作,必然就存在紧耦合型和松耦合型系统之分。EDBS 的一般性功能结构示于图 2。

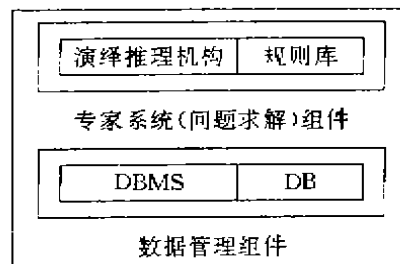


图 2 EDBS 的一般性功能结构

2.4 知识(库)系统 KBS

知识库系统或基于知识的系统(简称知识系统)

是由来自 AI 和 DB 的研究者共同从一个新的视角对未来信息系统的目标予以研究的结果^[1]。下面拟初步给出知识库、知识库系统和知识库管理系统的概念。

知识(库)系统 KBS 是任何这样一个由多个资源集成的系统,资源可包括硬件、软件、拟人或人,其中有一个最为重要的分离资源就是知识库 KB 资源,并由知识库管理系统 KBMS 负责对知识库的统一管理与维护。

知识库 KB 是一个可被共享的,拥有大量简单事实和通用规则的一个会集(Collection)。任何一个知识系统的问题求解能力都极大地依赖于对知识库中知识的运用(启发式推理)。

知识库管理系统 KBMS 是知识库 KB 的管理者,就如同 DBS 中 DBMS 的作用一样;本质上讲,KBMS 是一个软件系统,提供对大规模、共享知识库的管理设施。KBMS 在功能上应具有 DBMS 功能的某些典型特色,如并发访问、知识库分布、错误恢复、安全和可靠性等,再加上 AI 系统的某些特有机件,如演绎推理、带回溯的搜索、解释、答问、动态人机交互、自动分类、一致性维护等^[1]。

尽管 KBMS 与 DBMS 相呼应,但二者仍有本质的区别:

1) 在 DBMS 中,实例知识("Ground" Knowledge)和类属知识被截然分开,且分别保存在数据库和数据库模式之中;而在 KBMS 中,这两类知识间没有明显相应的区分;

2) DBMS 在功能上集中于满足最终用户的需求,KBMS 则重在满足 KB 设计人员(即知识工程师)的需求;

3) KBMS 需要考虑提供各种推理机制(如演绎与溯因(Abductive)推理等)及其相应的解释机制,DBMS 则不予考虑这些问题;

4) 设计一个 DBMS 的第一步是选择合适的数据库模型,而设计 KBMS 的第一步则是选择合适的知识表示框架。

2.5 主动数据库 ADB

传统 DBMS 只有在应用发出显式请求时,才对库中的数据施以相应的操作(仅此而已!)作为对应用请求的响应。从这个意义上来讲,传统 DBS 均是被动型的^[4]。

主动数据库系统 ADB,可以被定义成这样一个 DBS,它能够自动地响应(无需用户干预!)发生于系统内部或外部的事件^[5,6]。

从概念上来讲,传统 DBS 是较“纯”的数据库(只存储数据),对数据所能进行的处理仅限于根据应用请求完成必要的查询和增、删、改操作。主动数据库则除了存储“纯”数据之外,还试图将更多的行为(与数据库状态相关的动作)和数据一道存入库中,可以在对库中数据进行操作时,亦能因此触发某些已“存储”的行为的执行,这便是主动数据库中的主动性的全部本质(当然,行为的“存储”可以是物理的,也可以是逻辑的)。

由于主动性的引入,ADB 极大地挖掘了传统 DBMS 的功能潜力,可望更大地拓宽 DBMS 的功能范畴。主动性可用于更有效地进行 DBMS 自身的完整性维护、安全存取控制、导出数据(如视图 views)管理;异常处理、监控与报警、状态开关切换、系统性能评测、实时数据库与协作 DB 的实现;特别是各种推理机制的支持,额外地使 ADB 亦获得了 IDB 的美誉。除了这些功能上的扩充外,由于更多的行为已存入库中,应用开发遂被极大地简化。实际上在理想情况下,应用开发变成了对已存储行为和配置的一次配置,犹如应用所需的各处理模块均已齐备,只需组织一个总控程序一样。

ADB 着重追求的是如何获得主动性,并不强调 DBMS 应该基于什么样的数据模型。最早在数据库中涉足主动性功能的当属 CODASYL 的“ON”条件语句,但 ADB 试图把行为存入库中的愿望正好吻合了面向对象模型的特点。

基于关系模型的 ADB 是“通过预先在数据库之外嵌入一些可由某些事件引发执行的规则的办法来实现”其主动性的,很难(如果不是无法做到的话!)结合数据模型来实现主动性,唯有对象模型能够在数据的模型一级上支持数据与行为的集成管理,因而对象模型成了新一代的 ADB 的天然选择(这与 DDB 把关系模型做为其天然选择的理由一样充分)。

3 第二代系统

从功能上来讲,第一代系统是二维(即 DB 特征和 AI 功能)的。第二代系统则明确地引入了第三维功能,即多媒体支持,这是两代系统之间的根本区别。此外,第二代系统对第二维功能(AI 功能)的支持强度也更大。我们可认为第一代系统是相对“弱智”的,如 DDB 只在 DBS 中引入了演绎推理,ADB 只引入了产生式推理。第二代系统的所谓“高智”也是相对而言的,而且两代系统间的边界比较模糊。

3.1 对多媒体的支持

首先,第二代系统欲要成为未来的信息系统的理想开发平台,那么就必须对作为信息的四个基本属性之一的媒体性(Medium)予以充分的支持。“Medium”一词有下列两方面的含义:(1)Medium描述信息被传达的方式,信息是通过被传达来揭示其内容的。(2)Medium指信息存在和传输的介质,例如计算机系统磁盘介质,打印纸介质等。今后我们将采用“媒体”和“介质”这两个不同的中文词分别代表“medium”的这两种不同的含义。针对一定的场合,信息媒体有合适与不合适之分,本来是一幅画,非要用语言去表述而不许观其物,就让人(有观赏水平者)颇感不自然;而且信息媒体的合适与不合适也因场合的不同而异。

其次,AI研究的目的是使机器模拟人的行为,同时获得对人类智能的本质的认识^[1]。机器的行为能否表现得自然,也是计算机系统能否体现出人类智能的一个标志^[2]。衡量是否自然的依据是人的感觉,对于一个计算机系统而言就是用户的感受。因此,计算机系统若能在人机界面上支持从机器到人和从人到机器的多种多样信息传达方式,就会使用户感觉自然和方便,从而表现出智能来。因此,真正的智能数据库系统必须重视信息的多媒体支持。

再次,从AI研究的内容看,图象(模式)识别、语音识别(如 Hearsay-I, -II, -III^[3])和自然语言理解等都属于多媒体处理范畴,它们同时又都是AI领域的重要研究分支。即便是按照较直观的解释(即DB与AI简单结合),也应将多媒体支持引进IDB中。

最后,从多媒体技术(特别是多媒体信息系统)的角度来讲,多媒体信息的检索(查询)不象传统数据信息的查询那样能够方便地给出严格、准确的查询,用户常常只能给出较模糊、不精确和不完整的条件来,系统则必须对这些条件进行模糊的、非精确和非完全的匹配。这一方面是因为用户常常难以对非数据媒体信息给出精确和完全的匹配条件;另一方面,就目前的技术条件看,系统内部也难于对多媒体信息进行精确和完全的表示和组织,更难于精确和完全地去匹配检索条件。要进行非精确、非完全的匹配,必须借助AI技术,特别是对基于知识的方法以及非精确性推理(可能性推理、类比推理)等的运用。

因此,新的智能数据库系统需要从DB、AI和Multimedia三维来展开。欲把DB、AI和Multimedia支持集成在一个系统之中,探讨一个统一的数据模型和适宜的体系结构至关重要。

3.2 对象模型作为第二代系统的核心数据模型的可行性

依计算词典(第三版)的定义,信息包含其内容(在计算机中为符号表示)和操作两个方面;依文[2]之解释,知识是描述和操作的统一;按传统数据类型对数据的解释,数据也包括表示和操作两个方面,这些均与面向对象观点极为吻合。此外,信息固有的媒体性只有在具有较强模型能力的对象模型中才易于被充分地体现。因此,OO模型可望把数据、知识和信息统一起来,从而开发出具有单一概念模型(Object)的信息系统。

在OOP范型下,不仅模型概念简单一致,由于对象值(对象状态)与对象操作(对象行为)是被封装起来的实体,可予以一体化地管理。

3.3 第二代系统的体系结构特点

在第二代系统的三维功能结构中,DB特征仍强调数据模型,AI重视知识表示,而多媒体支持则注重强调信息传达的各种方式(媒体性)。在这三维当中,仍应把DB的作用摆在第一位。这样,在探索新一代智能数据库的体系结构时,就应该从“寻找一个适宜的数据模型”出发。

第二代系统若能基于一个统一的Object模型,其体系结构的核心就可归结为一个精简的OODBMS(Reduced Object-Oriented DBMS, ROODBMS)——这是现阶段实现第二代系统的一种可行途径。一方面多媒体数据库已把OO选作其模型,另一方面面向对象的演绎数据库(DOODB)和面向对象的主动数据库也已经积累了若干年的经验。于是,共同基于OO模型把它们进一步集成起来也就顺理成章了。多媒体智能数据库系统MIDS/BUAA系统的总体结构正是按照这样的思想设计出来的^[4]。

“分布性”和“协同性”^[7,1]代表着未来信息系统的体系结构应该具有的另外两个特征。DCW(分布式协同工作)范型已被广泛认为代表着未来的计算范型。

DCW范型的基础是,大量持久、共享并且分布的知识和信息。在协同工作的各Agent之间,有些知识是可共享的知识,有些是非共享的(领域专业知识和经验);有些知识是协调而且完全的,更多的情形则是不协调和不完全的,甚至是相互冲突的。一个计算机系统将包括成百上千台计算机,每台机器包含一组Agents。在这样的计算环境下,问题求解将是由一系列分布的Agents共同协作,来确定在现有资源

条件下的最佳问题解。

3.4 第二代系统的理论

我们认为,第二代系统的核心数据模型的理想选择是扩充的对象模型,其体系结构的内核可归结为 ROODBMS。因此第二代系统的理论研究最好也基于 OO 模型及其理论,但是到目前为止 OO 理论尚未成熟,已有的 OO 理论大多仅限于反映 OO 范型的某一部分特征,远未达到可用程度,如:C-logic, O-logic, F-logic^[10],概念对象代数 CA (Concept object Algebra)^[11,12],其中概念代数理论是较有前途的新一代数据库系统的理论。

4 总结

DBMS 是应现代社会对信息处理的需要而产生的,过去所开发的 DBMS 产品实际上仅仅是在处理数据,而非信息;所有的 AI-DB 系统(指第一代系统)所直接处理的仍不是信息,而是数据和知识。本文从信息具有的媒体性这一角度出发,认为真正的 IDB 不应该只是 AI-DB 的简单结合,还应该把对信息的多媒体支持引入进来,从而把 IDB 的概念向前推进了一步,并由此把 IDB 划分为第一代系统和第二代系统。下面再着重讨论一下各种第一代系统之间的区别及各自较适宜的应用。

逻辑程序 LP (Logic Program) 一般只适于有较少量的事实和相对多一点的规则,规模较小又需要逻辑推理的应用;专家系统 ES 可以看作是规模稍大一点的 LP;演绎数据库可以支持需要大量事实和相对少量的规则的应用;专家数据库系统 EDBS 对于既需要大量事实,又需要大量规则的应用较为适宜;知识库系统 KBS 在支持应用的规模上,与 EDBS 无太大差别,它与 EDBS 的主要区别是,在 KBS 中已看不见独立存在的 ES 和 DEMS,二者深度集成,融为一体;与前面介绍的诸多第一代系统相比,主动数据库系统 ADB 就如同是 IDB 社团中晚来的“移民”,人们更多强调的仍是 ADB 的数据库特色,很少将之与 AI 视为同宗,其实 ADB 中蕴藏着更大的 AI 潜力;新一代的智能数据库,无论在支持应用的规模和质量方面都应该较以往的系统有较大的改进,而且应该对 DCW 范型有很好的支持。

参 考 文 献

[1] Elke A. Rundensteiner, The Role of AI in Databases versus the Role of Database Theory

in AI: An Opinion, In Artificial Intelligence in Databases and Information Systems (DS-3), Robert A. Meersman 等 Eds, IFIP, 1990

- [2] Frederick Hayes-Roth et al., Building Expert Systems, Addison-Wesley Publishing Company, INC., 1983
- [3] D. A. Beel et al., Integrated Deductive Database System Implementation, A Systematic Study, The Computer J., Vol. 33, No. 1, 1990
- [4] U. Dayal, et al., The HiPAC Project: Combining Active Database and Timing Constraints, SIGMOD Record, Vol. 17, No. 1, 1988
- [5] N. H. Gehani, H. V. Jagadish, Ode as an Active Database: Constraints and Triggers. Proc. the 17th Intl. Conf. VLDB, 1991
- [6] Michael Stonebraker, The Integration of Rule System and Database Systems, IEEE Trans. on Knowledge and Data Engineering, Vol. 4, No. 5, 1992
- [7] Michael L. Brodie, et al., Future Artificial Intelligence Requirements for Intelligent Database Systems, Expert Database Systems, Proc. the 2nd Intl. Conf. 1989
- [8] Kamran Parsaye et al., Intelligent Databases, John Wiley & Sons, Inc., 1989
- [9] Dunren Che and Zhongfan Mai, The Overall Design of MIDS/BUAA: a Multimedia Intelligent Database System. ICYCS'93, Beijing
- [10] Weidong Yu and Bo Zhang, A Framework for Deductive Object-Oriented Database Systems, 同上
- [11] J. F. Nilsson, A Concept Object Algebra CA + $\wedge$ [:], Proc. from the 2nd European-Japanese Seminar on Information Modeling and Knowledge Bases, H. Kanggassalo et al., Eds., IOS Press, 1993
- [12] J. F. Nilsson, An Algebraic Logic for Concept Structures, Preprint of the 3rd European-Japanese Seminar on Information Modeling and Knowledge Bases, H. Kanggassalo et al., Eds., 1993
- [13] 车敦仁, 面向对象的智能数据库系统: 研究、设计与实现. 工学博士学位论文, 北京航空航天大学研究生院, 1994.