●陈 远 罗 琳 沈祥兴

信息系统中的数据质量问题研究*

摘 要 数据质量问题是当前信息系统研究中亟待解决的关键问题,数据质量问题对信息系统的影响是多方面的。解决数据质量问题的方法有:设置专职的数据质量分析员、专门的用户联络人、元数据的支持、严格的数据规范、区分数据清理的优先级。参考文献 3。

关键词 信息系统 信息处理 数据质量 数据质量分析
分类号 TP391

ABSTRACT The authors think that data quality is a key problem to be solved in information system, because it has multiple effects on information system. To solve the problem, we should have special data analyzer, special user service staff, metadata supports, strict data specifications, etc. 3 refs.

KEY WORDS Information system. Information processing. Data quality. Analysis of data quality.
CLASS NUMBER TP391

人工智能与数据挖掘等技术使信息系统的开发迈向了一个更高阶段,但就目前应用的信息系统状况来看,仍然存在许多问题,像数据定义不一致,历史数据无法使用,数据尚未集成,数据难以共享的现象最为普遍。由于信息系统应用的成效直接取决于所处理的信息的质量和可靠性,因而从多个数据源生成清洁、完整、一致、协调的高质量数据便成了当前信息系统,尤其是基于大型数据库的信息系统研究中亟待解决的关键问题。

1 数据质量的含义

在信息系统的软件项目开发中,数据质量往往被人理解为控制错误数据的输入,即在数据输入程序中加入检查和防范机制,保证进入系统的输入都是合法的数据值。但数据质量的真正含义远不止此,从系统中数据库的局面,可以把数据质量用以下几个元素来描述。

(1)正确性。数据的值在该数据界定的值域范围之内,即保证毫无意义的数据不进入系统。例如日期中的月只有"1~12"取值范围,超出这个范围的值则不正确。

(2)准确性。数据的正确性与准确性是不同的概念,数据值可能正确,但不一定准确。例如"434131"是一个正确的邮政编码,但用它标识武汉大学所在地域则是错误的。

(3)不矛盾性。要求数据值支持现行的业务规则。如果读者某次借阅的还书日期早于借书日期,则不符合预先定义好的业务规则。

(4)一致性。同一属性字段的数据在概念模式、值域范围、格式结构上均应保持一致。假如同一份文件在系统中的不同存储地给予了不同的编号值,便造成了不一致的现象。

(5)完整性。要求数据集合中既不缺少应有的数据,也没有多余的数据。若只根据或针对某个用户或部门的数据要求来提供数据,忽略另外一些需求或相关联的数据,便会影响数据的完整性。

(6)集成性。在传统的信息系统中,各组成部分并非按照整合的模式进行设计,信息分散处理和存储,造成冗余和浪费,而数据库的主键常常不能匹配,甚至有的数据文件没有键码,因此无法在一起使用。

2 造成数据质量问题的原因

一般情况下,组织机构中的职能部门都能得到一个或多个技术部门的支持,这些技术部门应职能部门的业务要求设计建立起一些互不相关的系统。由于每个技术部门仅仅基于直接用户的狭隘要求来收集和操作数据,结果数据按照特定值被保存或复制,导致了不一致、不精确以及不完整信息等一系列问题。更为严重的是,这种情况下存在着数千个文件和数据库中不受控制的数据冗余,以及数百个系统中不受控制的程序冗余。

另一方面,数据使用者常常不能理解或不能精确地使用他们的数据。他们的确知道这些数据,但有时候却不能肯定每个数据元素的真正含义,也不

* 本文系教育部人文社会科学重点研究基地重大资助项目(2000ZDXM870002)的研究成果。

能肯定存在于数据库中的实际内容。即使同一数据项,不同的人也有不同的定义和理解,另外,采用不同的编码系统,也会引起数据形式上的不同。衡量与错误数据相关问题的责任和提炼数据的义务的指标,一直都没有得到明确。

数据欠佳的情况和数据质量的恶化并不会在一夜之间显现出来,很多是历史遗留问题,加上一些主观或客观因素造成的问题,诸如系统平台的不一致,开发工具的技术缺陷,缺乏规范的操作流程和标准,数据处理人员水平的差异等等,导致可疑数据年复一年地积累下来。

3 数据质量问题对信息系统的影响

数据质量引发的问题是多方面的,下面就一些常见的质量问题分析它们的后果。

(1)填充无意义的默认值。有些数据可能没有具体的值,例如一个职工没有参加社会保障,则其档案中相应的社会保障号码一栏就会为空。在旧式文件系统中"空"的概念就是什么都没有,这是不容许的。对这种情况,数据输入程序和输入人员往往会造出一个值来填充,他可能会选择一个无意义的数据,或有意无意填充一个特殊的数据,这样用户在查询该项数据时,产生不信任感,同时也会造成系统安全的隐患。

(2)数据遗漏。不同的业务部门对数据有不同的需求,以便执行它们的业务操作,某些业务在系统开发时可能还未开展,或是并不重要,这些因素会造成数据遗漏,例如有关读者的年龄、种族、爱好等数据,在办理图书资料借阅流通业务时是用不到的,然而在对读者进行主动信息服务时却非常重要,这类数据的缺省将造成对读者资料分析或查询结果的障碍或曲解,使得无法准确地分析读者的信息需求,无法准确判断哪一种服务方式对读者最具吸引力,进而影响到改进图书馆服务方式的战略决策。

(3)违背业务规则的矛盾值。包括明显违背业务规则的不准确、不合逻辑的数据值,或是在根据记录中的其他字段值来生成某些字段值时发生了错误和矛盾的数据值。例如浮动利率贷款的最低利率居然高于最高利率,那么按最低利率计算贷款利息时,公司实际上蒙受了损失。又如,例如一个北京的公司被错误地赋予了一个武汉地区的邮政编码,在按地域统计公司的业务信息时,就会导致错误结论。

(4)多义数据项。表现在数据库和文件中,同一数据项(即字段)被定义为许多不同的意义,或是数据项的值被用于多个目的。例如某数据项原来定义的是日期类型,又重定义为字符串类型,这样在处理过程中就会发生冲突或混乱,造成错误,而单纯地去除重定义又可能发生数据遗漏问题。数据项的重用会给查询带来麻烦,例如值 A、B、C 可能代表业务部门类型码,如"采编部"、"流通部"、"服务部",而值 X、Y、Z 可能代表读者类型码,如"教师"、"学生"、"职工",那么在查询或对业务部门进行排序时,必须排除 X、Y、Z。如果没有弄清数据项的值域或当前使用的排除规则,错误地纳入或排除了某个值,查询结果将会出错,而且用户不会意识到。

(5)键码残缺或重复。应该关联的数据没有关联,即由于考虑不周全或出于困难没有建立联结两个对象的键码,例如,每一位新客户在银行都会分配一个惟一的账号,但很少有银行为每一位客户分配一个客户号,这样账号只能通过账户记录中的客户姓名与客户发生关系,而客户姓名缺乏惟一性,想查询账户与客户的相互关系就很困难。另一个极端是一个对象被多个键码所标记。如一个员工拥有几个员工号,因为他在几个部门中工作过,每次更换工作就分配一个新的员工号,而旧的员工号又可能被再次分配给他人,此时若要进行该员工长期的绩效、薪金、福利等分析就根本不可能。键码重复的后果也是严重的,有些事务型处理系统很少存放超过 90～180 天的历史数据,因而键码值常常被重新分配使用,这样就无法进行任何趋势分析,因为同样的键码值表示着不同时期的不同内容。

(6)非结构化的数据项及数据值。有些文件中备注型的数据项没有清晰的格式,每条记录在填写数据值时都可能以不同的秩序,例如一个地址的数据项有时按邮编、省份、街道顺序填写,而有时却按街道、省份、邮编的顺序,由此很容易被分解为单独属性的内容就无法用 SQL 语句来查询或分析。还有的现象是一个数据项被用于多种目的,即不该关联的数据却联系在一起,例如两个不同的业务系统都向同一个数据文件写入数据,这种相互混杂的对象使得用户无法进行判断。

4 解决数据质量问题的方法

数据质量的保障可借鉴企业生产的"全面质量"管理思想,对数据进行全员参与、全方位、全过程的质量控制,即从人、制度、技术等角度来制定相应措施。下面列举的是其中一些切实可行的方法。

(1)专职的数据质量分析员。尽管与信息系统相关的每一个人都有确保数据质量的义务,但必须有专职的人员去监督和担负起数据质量保障的职责,这就是数据质量员。数据质量员应发现和报告那些同数据质量有关的问题,他可以通过数据质量分析工具进行查询,以鉴别出数据质量问题;并着手调查这些问

题,估计解决问题所需的费用和牵涉的部门及利益相关者,区分出问题的轻重缓急,以向数据管理部门上报;确保那些必须解决的数据质量问题已有专人负责。并且数据质量分析员要加入到解决问题的研究和实施过程中,以提高和改善数据处理流程的质量。

(2)专门的用户联络人。用户联络人来自业务部门,是在用户中享有信誉、受用户信任的人。用户拥有数据,他们知道数据的有效范围,也知道他们将会需要的数据和它们的组织形式,用户决定数据访问权限、数据的可用性要求,并处于证实数据质量的极佳位置。设立用户联络人旨在用户与系统开发者之间建立联系,监督数据的质量。为确立数据的现实界定、数据的类型以及它们的总体安排提供帮助;对数据应该是具体的还是摘要的,及其安全性提出建议;在用户需求很多、时间有限且资料不够时,能够优先考虑向哪些需求提供帮助。

(3)元数据的支持。元数据是一种很好的"导航"(Navigation)数据。一旦原始数据被清理、转化、整合、简化乃至以各种形式被剖析,若不借助元数据将无法在数据仓库中重新定位该数据。元数据能告诉人们数据的来源、数据上一次的更新时间、数据的所有者、数据的含义、数据的转换方式以及数据值的可靠程度等。数据质量指数也可以作为元数据存储于数据库中。元数据有两种类型:业务元数据和技术元数据。业务元数据是为用户服务的,在日常基础上向用户解释他们使用的数据;技术元数据同时向用户和技术人员提供帮助,帮助他们研究像迷宫一样复杂的图表和计划(这些东西是被用来解释和维持数据仓库以及与此相关联的程序的)。值得注意的是,必须制定一定的规则,明确并分配获取和维护元数据的权利和责任,同时还要确定它的使用方式和时间,保证元数据的完整和准确。

(4)严格的数据规范。一个好的数据逻辑结构应该是没有冗余的、灵活的、简单的,并可适用于多个不同的应用,在对数据库进行操作时应消除插入异常,删除异常现象。数据库的规范化理论已经提出并应用了30多年,从1NF(First Normal Form)2NF、3NF、BCNF(Boyce/Codd Normal Form)、4NF到5NF,它们通过消除具有重复组的数据项,消除多次出现的数据,以及消除不依赖于键码的属性等步骤,提高数据库结构的规范化,以避免可能出现的会降低数据库内容可靠性的问题和异常情况。实践证明这种规范化理论是非常有效的。另外,在数据录入时建立一些规范化指南和控制,可以使用抽取、转化、加载工具(extract/transform/load,ETL),许多ETL本身具有数据清理功能。也可以使用专门的清理工具用于清理不符合标准的数据,诸如数据类型不合、数据值溢出、数据无效、与业务规则冲突以及错误的地址等等之类。现在市面上的数据清理产品能够解决大多数系统共有的带普遍性的数据质量问题。

(5)区分数据清理的优先级。每个人都希望高质量地控制所有的数据。但这是不现实的。我们可以运用各种手段尽可能地检测出系统中存在或暗藏的数据质量问题;评估其可能造成的不良后果,并进行可行性分析,包括各种解决方案的成本与功效,以及技术上的可能性;将解决方案的成本与不解决它所带来的损失进行对比,若损失大于解决方案的成本,就将该项解决方案放入"待解决"的清单中;对"待解决"清单中的各种解决方案列出优先顺序。对大多数情况而言,使用事务型数据的用户不需要更加"清洁"的数据,而那些具有深层次意义的数据必须保证其应有的质量,所以应当获得最高的清理优先级。

5 结语

为了交叉组织系统所在领域的业务功能,并由此生成整体性数据,以便给用户提供更好的信息服务,目前大多数信息系统都引入了数据仓库模式。建立基于数据仓库模式的信息系统的一个主要原因是:数据仓库能够提供比现在的在线分析工具及传统的信息系统更为清洁可靠的数据。在这种系统中,主要的数据清理工作往往依靠数据仓库的抽取程序完成。

然而,数据质量问题是永远存在的,技术部门应该制定有非常主动的质量计划,同时用户应该积极主动地分辨出数据质量问题,并向技术部门报告。解决数据质量问题不能光靠等待有效清理工具的出现,重要的是要强调管理的职能,发挥出信息系统所有干系人的主观能动性。

参考文献

1 [美]锡德·阿德尔曼,拉里萨·特佩卢克·莫斯著;薛宇,王剑锋译.数据仓库项目管理.北京:清华大学出版社,2003
2 [美]斯太尔,雷诺兹著;张靖,蒋传海译.信息系统原理.北京:机械工业出版社,2000
3 俞瑞钊,陈奇.智能决策支持系统实现技术.杭州:浙江大学出版社,2000

陈 远 武汉大学信息管理学院副教授。通讯地址:武汉大学信息管理学院。邮编430072。

罗 琳 武汉大学信息管理学院教师、博士。通讯地址同上。

沈祥兴 高级工程师,武汉大学信息管理学院实验中心主任。通讯地址同上。

(来稿时间:2003-04-22)