

数据质量分析及应用

丁海龙,徐宏炳

(东南大学 计算机科学与工程学院,江苏 南京 210096)

摘要:随着信息系统的广泛应用,人们在获得海量信息的同时,越来越被数据的质量问题所困扰^[1]。自从数据库管理系统(DBMS)出现后,数据已不再是程序的附属品,转而成为一种独立的产品。在应用程序升级换代的过程中,数据不但贯穿始终,而且变得越来越宝贵。文中以税务行业为应用背景,从实践角度探讨了分析数据质量的若干途径。阐述了数据质量的定义;针对定义的各项标准,分别阐述采用的解决方法;最后对数据质量分析的未来发展进行了展望。

关键词:数据质量;数据解释;可达性

中图分类号:TP311.13

文献标识码:A

文章编号:1673-629X(2007)03-0236-03

Data Quality Analysis and Application

DING Hai-long, XU Hong-bing

(Department of Computer Science and Engineering, Southeast University, Nanjing 210096, China)

Abstract: These days, information systems have been used in almost all aspects of daily lives. Things have changed greatly after the invention of database management system (DBMS). Data is now no longer the subordination of programs. It turned out to be an independent product. It stays where it is and become more valuable when applications have been updated or changed. In this article, discuss several methods to analyse and improve the data quality in a practical way based on a tax background. In the first place, give a definition of data quality; then explain the methods according to each item of the definition. Discuss the future development in the last place.

Key words: data quality; data explanation; data accessibility

1 数据质量的定义

简而言之,数据质量反映出数据对特定应用的满足程度^[2]。数据,如同形形色色的其它各种产品一样,是为了满足人们的特定需求。在信息系统中,数据是应用程序的初始原料和最终产品,并经过应用程序的组织,提供给用户^[3]。同样的一组数据,面对不同的应用要求,可能表现出不同的数据质量。参考文献[4,5]中列出了一些行业标准,对各自行业中满足要求的数据质量做了具体的规定。

然而,数据在反映客观世界,完成信息传递功能的同时,作为一种特殊的产品,应该具备一些最基本的属性^[6]。下面将着重从数据应该满足的共性的角度来考虑数据的质量问题。

1) 数据。

此处定义的数据不单包括数据本身,还包括对数据的解释。所研究的数据主要指存储在数据库中的数

据。数据解释的作用在于阐明数据库中表和字段的约定,以及相关的业务说明。数据解释的地位相当于数据这种特殊产品的说明书,是数据不可或缺的一部分。

2) 可达性。

用可达性来衡量数据量的大小对应用的满足程度。可达性的计算方法为:可达性 = 应用能够获取的数据量/应用所要求的数据总量 * 100%。例如,为了分析1995~2005年10年间的某市地方税收的增长情况,需要得到这10年的历史数据。但2001年以前的数据没有迁移过来,所以可达性 = 5/10 * 100% = 50%。

3) 正确性。

用正确性来表示数据库中的数据与客观世界的符合程度。例如,纳税人更改了公司名称或者联系方式后,应该对纳税人基本信息表中相应记录进行更新,否则就会得到不正确的数据。正确性的计算为:表中正确的数据量/表中的记录总量 * 100%。

4) 完整性。

用完整性来表示信息的完整程度。完整性包括三个方面的内容,分别是实体完整性、引用完整性和域完

收稿日期:2006-06-22

作者简介:丁海龙(1977-),男,江苏人,硕士研究生,研究方向为数据库设计与应用;徐宏炳,教授,研究方向为数据库设计与应用、数据仓库、数据挖掘技术等。

整性。实体完整性要求一个表中的每一行必须是唯一的;引用完整性定义了一个关系数据库中不同的表的相关列的之间的引用关系;域完整性要求表的某一列的数值在该列的合法的数值范围内。完整性的计算方法为:数据集中所有满足条件(可以是上述三者之一)的数据量/集合中记录总数 * 100 %。

5) 一致性。

用一致性来衡量对于特定的规则,数据库中所有的表是否都满足这样的规则。例如:人员信息表中规定了“M”表示男性,“F”表示女性。那么可以考察所有表中表示性别的字段是否都以同样方式表达。定义一致性的计算方法为:数据库中所有满足条件(针对某个具体规则)的数据量/被考察的记录总数 * 100 %。

6) 时效性。

用时效性来考察数据的时间特性对应用的满足程度。数据从产生、发展,到消亡,有一个相对的有效期。不同类型的应用对数据的时间特性有不同的要求。通常实时应用系统中的数据有效期较短。定义时效性的计算方法为:数据集中所有尚未失效的数据量/集合中记录总数 * 100 %。

7) 相互关系。

绝大多数的应用都会要求访问一定范围内的数据。为了支持特定的应用,可达性是数据应该满足的首要特性。正确性是数据质量的根本属性。完整性、一致性和时效性,从几个方面对正确性进行反映。完整性从数据数值的合法性角度考察数据的正确性;一致性从数据对应用逻辑的符合程度去考察;时效性从数据这样一种特殊产品的生命周期来考虑。数据质量几个特性之间的关系如图 1 所示。

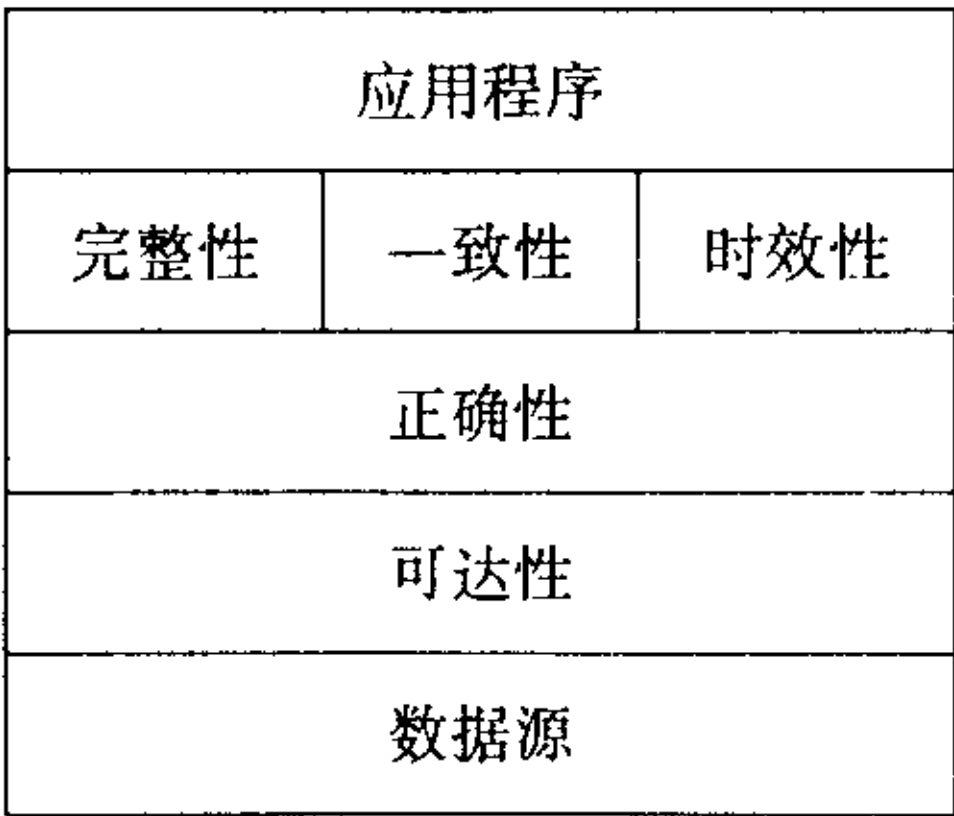


图 1 数据质量各特性间的关系

2 实例分析

某市地税局从 1994 年建立第一个税收系统至今,其信息化建设取得丰硕成果。但由于缺少统一规划,导致信息系统建设质量不高,数据质量严重下降,目前已经严重制约其信息化的进一步发展。依据数据质量的几个方面的指标,对其进行了改进数据质量的一些工作。

1) 可达性分析。

统计分析性质的应用,例如分析企业的诚信记录,或者对下一年的税收指标进行预测等,需要访问历史数据,并且需要访问其它业务系统,如:稽查系统中的处罚情况,征管系统中的税收增长数据等。由于业务系统中表的数据量很大(往往占用 1G 以上的存储空间),该税务局把超过三年的数据以文件备份的方式迁移出了当前数据库,导致该部分历史数据无法访问。此外,各个时期开发的业务系统使用的数据库也各不相同,数据互访困难。

为了改善可达性,首先把备份文件导入到一个独立的数据库中,然后建立该数据库到当前业务数据库的连接。同时创建了包括历史数据和当前数据的统一视图,使得历史数据可以被平滑访问。为了达到各个异构的数据库相互访问的目的,设定可以被全局共享的 ODS(Operational Data Stores)。通过统一的规则,把数据从各个分散的系统抽取到一起。在应用规则的同时,还达到数据清洗的目的。

2) 完整性分析。

着重描述在域完整性分析方面所做的工作,主要针对数据类型为数值型的数据。具体分为以下三个方面的内容:非法数值的分析、数据空值情况分析,以及数值分布情况的分析。

(1)非法数值的分析。

任何字段的数据都应该符合特定的数据格式以及值域范围。例如:表示身份证号码的数据应该为 15 位或者 18 位;移动电话或者固定电话号码应该符合特定的位数;表示学历的信息应该如“本科”、“硕士”、“博士”等;人的年龄应该在 0 到 120 岁之间等等。

笔者编制程序,在整个数据库的范围内对非法数据进行分析,并把分析的结果自动写到 Excel 电子表格中。得到的结果与表 1 类似(根据分析的字段性质不同,表的格式略有不同)。

表 1 非法数值的分析

表名	字段名	最小值	<0 的记录数	所占比率
纳税人投资方信息	投资金额	- 515731620	293	0.62%
社保上传数据	缴纳金额	- 6584	6933	4.67%
...

(2)数据空值的分析。

人们直观的感觉认为空值数据对辅助决策支持系统影响很大。人们探讨各种方法来解决空值数据所带来的问题。由于设计不严格,数据表中一些关键的字段可能没有约束。这些字段出现空值会对业务系统产生很严重的影响。笔者编制程序,对数据库中被认为

重要的字段进行了空值情况分析,得到的结果如表 2 所示。

表 2 数据空值的分析

表名	表中记录数	字段名称	空值数	所占比率
社保缴纳情况分户清册	9077332	税务登记证号	3201	0.035%
		税务机构代码	1205	0.013%
...

(3)数值分布情况的分析。

相当多的数据,无法通过上面的非法数值分析以及数据空值分析来发现数据的质量问题。这时可以通过分析其数值的分布情况来发现可疑值。例如:我们绘制出该税务局所辖地区上一年的个人收入情况,把该曲线与国家统计局公布的数据进行对比,可以发现该地区的个人收入申报情况是否真实。针对某个具体指标,通常先画出其散点图(如图 2 所示)。根据这样初步的信息,进一步分析数据的正确性。

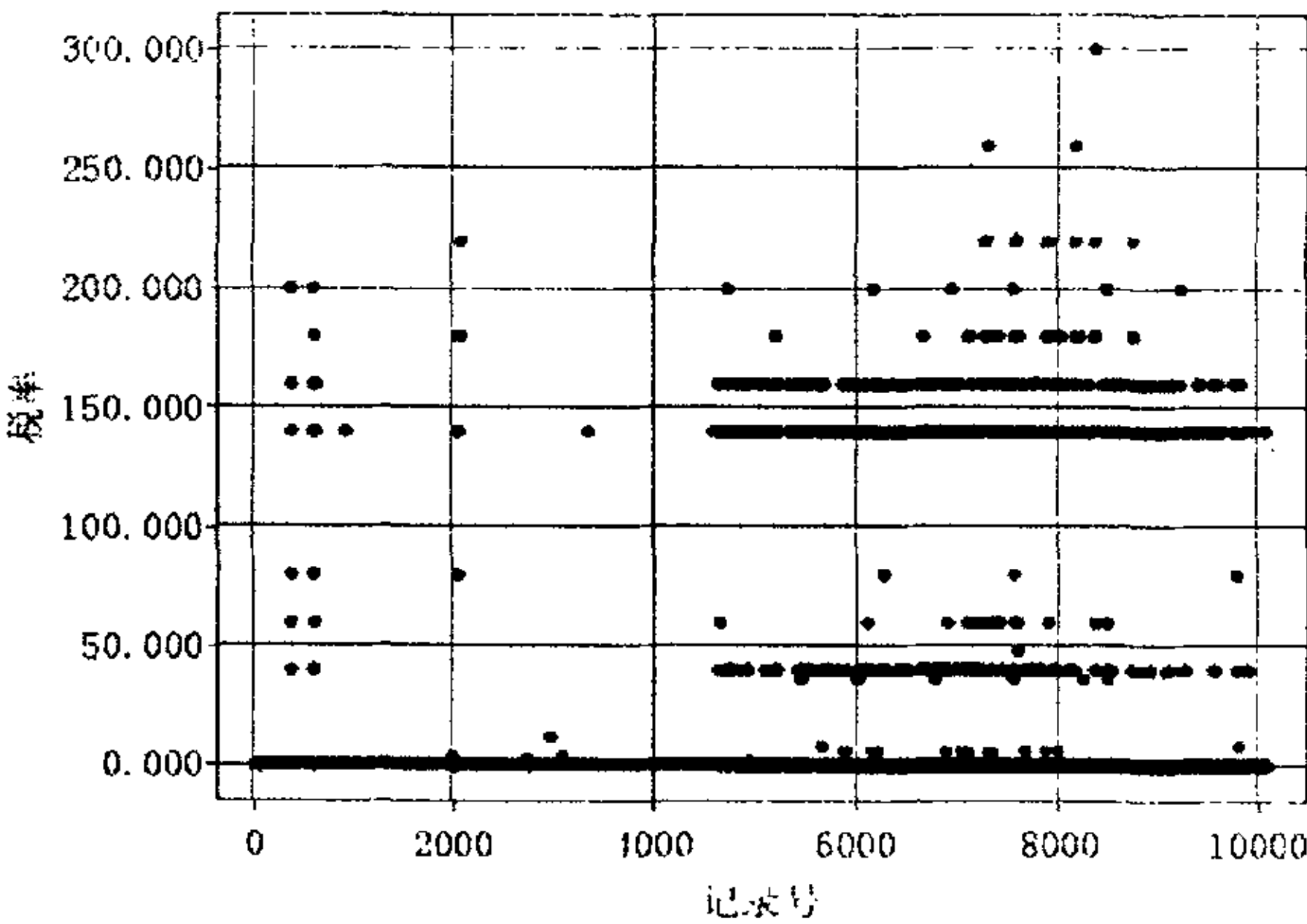


图 2 税率的散点图

3)一致性分析。

在改善数据可达性的时候,采用了 ODS 共享数据库的方式。在进行数据集成的同时,解决了很多数据一致性的问题。但是在各个分立的业务系统内部,数据不一致的现象大量存在。例如,“失踪户”和“失效户”意义相同。在相关的几个表中,这两个术语同时存在。通过与用户确认,最终用后者代替前者。利用程序对整个数据库进行扫描,从而确保了一致性。

4)时效性分析。

通常,比较容易发现对于一个确定的应用,需要哪些数据支持。但是,当一个数据库系统运行很长时间后,会产生很多废弃的表,留下大量垃圾数据。解决这

个问题,可从两个方面入手。首先对数据库中所有表的记录变化情况进行了长达六个月的连续监测。显而易见,数据量有变化的表是当前正在使用的表。对于数据量不变的表要分为两类:一类就是字典表。这类表的数据量变化很小,或者不变;另外一类就是真正被废弃的表。其次,再分析数据库中保存的 SQL 语句(例如在 Oracle 数据库中的 VS SQL 系统表)。通过对表被访问的情况进行统计,最终可以发现那些从来就没有被使用过的表。在该税务局的征收管理数据库中,一共分析了 1470 张表,发现其中的 683 张表没有被使用,其中的 219 张表还保存有少量数据。

3 结论及展望

自从人类进入工业社会以来,已经有了严格的产品质量标准。遗憾的是,对于数据这样一种特殊的产品,目前还没有统一的标准来衡量其质量。文中以地税为行业应用背景,做了一次有益的尝试。着重从数据的可达性和完整性的角度,分析了数据的质量问题。数据这种特殊产品是为应用而服务的。数据质量的高低归根结底表现为对应用的满足程度。卓有成效的数据分析应该和具体应用紧密结合^[7]。此外,为了应对信息社会的海量数据,自动化的分析工具必不可少。

参考文献:

[1] Strong D M, Lee Y W, Wang R Y. 10 Potholes in the Road to Information Quality[J]. IEEE Computer,1997,30(8):38-46.

[2] Lee Y W, Strong D M. Knowing - Why About Data Processes and Data Quality[J]. Journal of Management Information Systems,2003,20(3):13-39.

[3] Lee Y W, Pipino L, Strong D M, et al. Process - Embedded Data Integrity[J]. Journal of Database Management, 2004, 15(1):87-103.

[4] 中国科学院计算机网络信息中心科学数据库中心. 大气数据元数据标准[S]. 2003.

[5] 国际货币基金组织统计部. 国民账户统计数据质量评估框架[S]. 2003.

[6] Pipino L L, Lee Y W, Wang R Y. Data Quality Assessment [J]. Communications of The ACM, 2002,45(4):211-218.

[7] Strong D M, Lee Y W, Wang R Y. Data Quality In Context [J]. Communications of The ACM,1997,40(5):103-110.