

一种支持科技信息资源共享的数据服务模型

汤怀美¹, 张 程², 杨冬菊²

(1 山东科技大学 信息科学与工程学院, 山东 青岛 266510; 2 中国科学院计算技术研究所, 北京 100080)

摘 要: 提出一种用于科技信息资源共享的数据服务模型, 支持以服务的形式发布并共享各类科技数据, 特别是在服务层面对数据的容量、冗余度、异常情况, 以及完整性等数据本身的质量信息加以描述, 并给出了这些质量信息的获取和更新方法. 在服务层面为用户提供数据特征和数据质量的参考信息, 有利于用户对科技信息资源的选择和使用, 能够实现科技信息资源的有效共享, 通过在全国科技信息服务网项目中的应用, 初步验证了其效果.

关键词: 数据质量; 数据服务模型; 科技信息; 资源共享

中图分类号: TP311

文献标识码: A

文章编号: 1000-7180(2011)08-0166-03

A Data Service Model of Supporting Science and Technology Information Sharing

TANG Huai-mei¹, ZHANG Cheng², YANG Dong-ju²

(1 The College of Information Science and Engineering,

Shandong University of Science and Technology, Qingdao 266510, China;

2 Institute of Computing Technology, Chinese Academy of Science, Beijing 100080, China)

Abstract: This paper presents a data service model of sharing science and technology information. The model supports publishing and sharing scientific data in the form of services especially describing the quality of the data itself such as data capacity, data redundancy, data anomalies and data integrity and also gives the methods to access and update quality information. The model provides reference information of data characteristics and quality information for users in the service level and benefits users of technology selection and use. It can realize effective sharing of science and technology information resources. Its effectiveness is verified initially by the application in National Science and Technology Information Service.

Key words: quality of data; data services model; science and technology information; resources sharing

1 引言

科技信息资源是指具有价值和使用价值与社会活动相关的各种科技、贸易、生产方面的资料、消息等^[1]. 在实际应用中, 科技信息共享需求突出. 然而, 目前科技信息资源共享还面临诸多困难, 主要表现在: (1) 各地科技信息自主建设标准不统一、数据规范性差; (2) 共享需求多变, 缺少灵活装配资源的支撑机制; (3) 在发布科技信息时, 缺少反映科技信息资源本身的“特征”和“质量”的辅助信息. 近年来兴起的数据服务技术为解决异构资源共享和集成问题

带来新的方法. Oracle 公司推出的 AquaLogic^[2] 数据服务平台和 WSO2 的 Data Services Server 可以将各种异构的数据库资源和关系数据库, Excel 文件, Google Spreadsheet 以及 CSV 等格式的原始数据封装为数据服务. 在数据质量方面, 文献[3]提出将资源质量描述纳入到 REST API 和 RESTful 服务的发布, 让服务消费者指定上下文、查询服务质量和资源质量, 从而得到更加满意的结果, 文献[4]提出在描述数据服务时加入数据质量信息, 来提升数据服务选取与使用效果. 虽然有了上述的一些研究成果, 但是, 目前人们还是注重从整体上对数据服务

的计算能力、业务功能等进行刻画,而忽视了对数据质量信息的管理和使用,对如何正确获取数据质量信息探讨的也较少,这阻碍了消费者对数据服务的选择和使用.本文提出了用于发布科技信息的IDS(Science and Technology Information Data Service)模型.该模型以服务的形式对数据进行封装,并在传统服务描述基础上增加了数据质量信息的描述,同时给出了这些质量信息的更新方法,为用户在服务层面提供了数据特征和数据质量的参考信息,解决了“发布科技信息时缺少反映科技信息资源本身的特征和质量的辅助信息”这一困难.

2 科技信息资源数据服务建模

2.1 IDS模型描述

IDS模型包含发布科技信息所需要的四项基本内容即:元数据(Meta Data)、数据质量(Quality of Data)、数据环境(Data Environment)和数据安全(Data Security),各属性又包含了多个二级属性. IDS模型结构图如图1所示.

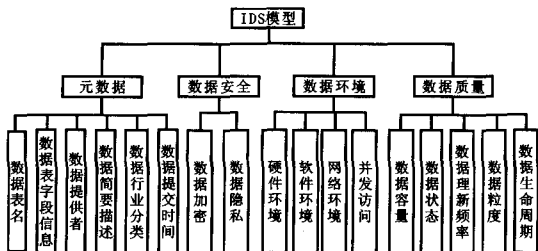


图1 IDS模型结构图

• 元数据(Meta Data)

元数据用来封装规范化科技信息的数据表名(tableName),数据表字段信息(columnInfo),提供者(source),简要描述(description),行业分类(professionCategory),提交时间(archiveTime)等内容并提供数据的存储、访问接口.该属性方便了使用者对科技信息的了解和操纵.

• 数据质量(Quality of Data)

数据质量是科技信息资源有效共享的前提,科技信息质量越高,用户利用率就越高,决策失误率就越低.根据科技信息资源特点,IDS模型中定义数据质量(Quality of Data)指标包括:数据容量(Data Capacity)、数据更新频率(Update Frequency)、数据状态(Data Status)、粒度(Granularity)、数据生命周期(Data Lifecycle)几方面,这些指标能够反映科技信息资源的质量情况,督促数据提供者加强数据质量管理,有利于用户对数据进行正确选择和

使用.

• 数据环境(Data Environment)

数据环境描述了发布科技信息资源所需要的运行环境(Runtime Environment),包括硬件环境(内存、CPU、硬盘等)、软件环境(操作系统、数据库等)、网络环境(网络带宽、数据传输速度等)和并发访问数的要求.数据环境属性能够帮助数据发布者对科技信息资源进行合理有效地发布.

• 数据安全(Data Security)

数据安全指出了哪些科技信息需要加密、隐私以及如何进行加密和隐私等.数据安全属性能够有效防止科技信息资源被非法窃取和篡改以及资源被无偿或侵权使用等,保证网络环境下科技信息资源共享的安全.

2.2 IDS模型属性值的获取及更新方法

IDS模型的属性按值的变化情况可分为动态属性和静态属性两类,在实际应用中需要获取它们的值.数据环境、数据安全以及数据质量中的粒度,生命周期等属性值不会随着科技信息的更新而变化,将它们称为静态属性,可以根据科技信息的发布要求和使用情况获取它们的值,其余属性称为动态属性.本节重点介绍动态属性值的获取和更新方法.

2.2.1 动态属性值获取

• 初始数据重复率、重复数据控制率获取

重复数据包括完全重复数据和近似重复数据两种.对初始数据进行重复数据检测,获得重复数据量并计算初始数据重复率.删除重复数据后再次计算重复数据率,如果小于规定的阈值结束处理,否则继续处理直到小于阈值.处理完成将相关值写入IDS模型.本文借鉴文献[5]中方法的思想进行重复数据检测,然后根据数据丰富量确定删除哪条重复数据.数据丰富量定义如下:

$$C_R = \sum_{i=1}^n b_i * w_i \quad (1)$$

式中, w_i 是数据记录中第*i*个字段的权重($0 < w_i < 1$), b_i 是数据记录中第*i*个字段的字段丰富量($0 < b_i < 1$), n 是该条记录中的字段个数.字段值越丰富, b_i 值越大,值为空的字段 b_i 值为0.数据重复率计算公式为

$$\text{数据重复率} = \frac{\text{重复数据量}}{\text{总数据量}} * 100\% \quad (2)$$

• 初始数据异常率、异常数据控制率获取

数据异常是指数据中包含乱码、连续较长的空字符以及回车符、html标签等特殊字符的现象.消

除数据异常首先要进行数据样本的挑选,根据指定规则将数据划分成层,使用随机抽样方法从每层抽取数据样本集,获取样本集中的数据异常现象并将其作为检测样本.在每一层利用检测样本对数据进行异常检测,统计异常数据量并计算数据异常率.删除或修复异常数据后,重新进行计算,如果数据异常率大于规定的阈值继续处理直到小于阈值,否则结束处理.处理完成,将相关值写入 IDS 模型.数据异常率计算公式为

$$\text{数据异常率} = \frac{\text{异常数据量}}{\text{总数据量}} \times 100\% \quad (3)$$

• 数据完整性比率获取

数据不完整是指数据中属性值丢失,不完整数据会引起决策错误.因此,数据完整性是数据质量的一个重要因素.通过判断字段值是否为空来检测不完整数据,然后采用文献[6]提出的方法检测不完整数据的可用性,如果数据不可用将其删除,否则,补全记录丢失值使其成为可用记录.数据完整性比率计算公式为

$$\text{数据完整性比率} = (1 - \frac{\text{不完整数据量}}{\text{总数据量}}) \times 100\% \quad (4)$$

• 数据容量及数据更新频率获取

预设合理的数据量统计机制,对每一次更新的数据量进行统计,原数据容量和本次更新数据量的和就是总数据容量.由更新数据量和数据更新周期计算数据更新频率.计算公式为

$$\text{数据更新频率} = \frac{\text{更新数据量}}{\text{更新周期}} \quad (5)$$

2.2.2 动态属性值更新

当有新数据进入数据库时需要根据上述方法重新获取属性值并更新到 IDS 模型中以保证其属性值真实可靠,更新公式为

$$\text{新数据比率} = \frac{\text{原脏数据比率} \times \text{原数据容量} + \text{新脏数据条数}}{\text{数据容量}} \quad (6)$$

3 应用实例及效果分析

目前,IDS 模型已经在全国科技信息服务网的建设中得到应用.应用 IDS 模型对各节点提交的科技信息资源进行规范化封装,不仅使科技信息达到了规范标准,而且为实现科技信息资源的有效共享发挥了应有作用.模型中封装了科技信息资源的元数据信息并提供了数据存储和访问接口,方便了用户对数据的操纵.模型中提供的初始数据和规范化

数据的数据质量信息,为用户提供了数据特征和数据质量的参考信息,这不仅可以让数据提供者了解所提供数据存在的问题,加强数据质量控制,而且方便了数据使用者对数据的选择和使用.此外,模型中提供的数据库环境和数据安全的相关说明和要求,有利于科技信息的正确发布和保护数据安全.

IDS 模型不仅在服务层面实现了科技信息资源的有效共享而且使业务层面的工作也得到了提升.从数据提供者来看,可以使用 IDS 模型进行问题自检,利用模型提供的初始数据质量信息发现数据存在问题,从而加强数据质量控制.在全国科技信息服务网建设过程中,各节点利用该模型进行问题自检,不断提高数据质量,5 个月时间,数据问题率总体下降了 10%.此外,IDS 模型也为数据管理员、数据发布者、数据使用者和管理者等带来了工作上的便利.例如,可以帮助管理者进行员工绩效考核,帮助使用者准确选择所需数据等.

4 结束语

本文针对科技信息资源的数据特点和共享需求,给出了一个具有数据质量描述的、规范化的数据服务模型的建模方法,该方法通过增加对数据容量、数据重复率、数据异常率等的描述能够对数据的质量信息进行刻画,同时给出了这些质量信息的更新方法,为生成规范化数据服务提供了保障.该模型不仅能够反映科技信息资源的特点和数据质量,督促数据提供者加强数据质量管理,而且能够实现科技信息资源的有效共享.该模型在中国科学技术部牵头的全国科技信息服务网建设中已经得到实际应用,实验和事实都证明了该数据服务模型的有效性和实用性.

参考文献:

- [1] 蒋芬. 科技信息资源建设发展现状的思考[J]. 今日科技, 2010, (4): 36-37.
- [2] Borkar V, Carey M. Query processing in the aqualogic data services platform[C]// Proceedings of the 32nd International Conference on Very Large Data Bases, Seoul, Korea; IEEE, 2006: 1037-1048.
- [3] Truong H L, Dustdar S. Context, quality and relevance: dependencies and impacts on RESTful web services design [C]// ICWE Workshops on Lightweight Composition on the Web, Vienna; ACM, 2010: 347-359.

(下转第 172 页)

4 实验结果对比分析

表 1 说明了本文使用小数据原则与大数据原则方法进行启发式规模约束聚类所得到的评价结果。由表中可以清楚地看到,本文采用小数据原则与大数据原则进行规模约束,聚类性能得到了提高,得到了优化,将小数据原则与大数据原则规模约束条件加入到传统的或现有的随机规模约束的聚类算法中,仍然无需确定精确的聚类规模,UCI 数据集上的实验结果表明聚类性能有了显著提升。

使用 UCI 中四个数据集并且通过放宽精确规模约束到范围规模约束来研究本文所提方法的灵活性,这意味着除了指定特定的规模给各个组类,大致的规模范围同样可以在本文所提方法中进行应用,这在现实应用中非常重要。本文指定各组类的样本数量不超过 75,对数据采用 1:1 模式与随机模式进行比较结果如表 2 所示。

表 2 规模约束放宽后的案例研究结果

数据集类型	算法	准确度	平均信息量	ARI	NMI
Wine	规模约束放宽	0.846 7	0.255 9	0.641 6	0.675 0
	1:1 模式放宽	0.930 1	0.198 9	0.763 7	0.792 8

5 结束语

本文针对规模约束环节来进行改进,以强化聚类算法的性能并以此来优化数据挖掘的过程:对各类规模约束后所含实例对象数量不一致将降低聚类算法的性能的具体问题,采用一种新的模式对各类进行了规模约束,并将这个问题转化为了线性优化问题,对线性优化问题进行了最优处理和求解,从而改进规模约束问题的解;通过 UCI 标准数据集上的实验,说明这个算法与随机模式相比具有更好的聚类精度,也就是说使当规模约束适当放宽后,聚类性能也可得到明显提升。

参考文献:

- [1] Wagstaff K, Cardie C, Rogers S, et al. Constrained K-means clustering with background knowledge[C]// Proceedings of ICML. San Francisco, CA, USA: ACM, 2001:167-173.
- [2] Zhu Shunzhi, Wang Dingding, Li Tao. Data clustering with size constraints[J]. Knowledge-Based Systems, 2010:883-889.
- [3] Basu S, Banerjee A, Mooney R J. Active semi-supervision for pair-wise constrained clustering [C]// Proceedings of SIAM Data Mining. Florida, USA: ACM, 2004:111-119.
- [4] 钟洪,夏利民. 基于互信息约束聚类的图像语义标注[J]. 中国图形图像学报, 2009, 14(6):25-28.
- [5] 李敏强,李智. 基于约束聚类的一种概念学习方法[J]. 系统工程学报, 2004, 19(5):32-36.
- [6] Banerjee A, Ghosh J. Scalable clustering algorithms with balancing constraints [J]. Data Mining Knowledge Discovery, 2006:31-34.
- [7] Zhong S, Ghosh J. A Unied framework for model-based clustering [J]. Journal of Machine Learning Research, 2003:12-15.
- [8] Massatfa H. An algorithm to maximize the agreement between partitions [J]. Journal of Classification, 1992: 22-24.
- [9] Hubert L, Arabie P. Comparing partitions [J]. Journal of Classification, 1985.
- [10] Studholme C, Hill D, Hawkes D J. An overlap invariant entropy measure of 3D medical image alignment [J]. Pattern Recognition, 1999.

作者简介:

朱顺彪 男,(1973-),博士,副教授.研究方向为信息系统、数据挖掘、GIS应用研究。

洪文兴 男,(1981-),博士,助理教授.研究方向为推荐系统与数据挖掘。

(上接第 168 页)

- [4] Wang R Y. A product perspective on total data quality management[J]. Communications of the ACM, 1998, 41(2):58-65.
- [5] 陈伟,王昊,朱文明. 一种提高相似重复记录检测精度的方法[J]. 计算机应用与软件, 2006, 23(10):29-30.
- [6] 陈伟,丁秋林. 数据清理中不完整数据的清理方法[J].

微型机与应用, 2005, (2):44-45.

作者简介:

汤怀美 女,(1984-),硕士研究生.研究方向为信息资源管理。