万维网资源质量模式挖掘技术分析

朱 焱
(西南交通大学信息科学与技术学院 成都 610031)

摘 要 基于万维网(Web)的商务智能和决策系统成功的关键是遴选并使用万维网上的高质量信息。由于Web资源具有高动态性、高自主性、数据海量、信息类型多样,以及应用要求不同等特点,造成了严峻的Web信息源质量问题。国内外已开始着手Web资源质量的研究。分析了各类基于Web的高端应用(如商务智能)对Web资源和信息的质量要求,指出了Web资源质量带来的挑战,综述了Web资源质量模式发现和评测方法的现状,深入讨论了应用数据挖掘及相关技术发现、处理Web资源质量异常的原理,指明了Web资源质量挖掘领域亟待解决的问题和需要深入研究的方向。

关键词 Web资源质量,质量模式挖掘,元数据管理,质量评测方法
中图法分类号 TP393.4 **文献标识码** A

On Web Source Quality Pattern Mining Approaches

ZHU Yan
(School of Information Science and Technology, Southwest Jiaotong University, Chengdu 610031, China)

Abstract The key to success of the Web-based information management, business intelligence and decision making systems is high quality information from the Web. However, the Web source quality is very problematic due to the peculiar characteristics of the Web, such as, dynamics and autonomy of Web sources, enormous amount and various types of Web data, multifarious quality requirements of Web applications, etc. There has been some work on Web source quality management. In this paper, the quality requirements of advanced Web-based applications (e.g. business intelligence) and the quality challenges of Web sources were analyzed. The state-of-art in Web source quality pattern discovery and evaluation was surveyed. Data mining and the related approaches for dealing with Web quality issues were investigated to reveal many still unsolved problems and to suggest several important research directions.

Keywords Web source quality, Quality pattern mining, Metadata management, Quality evaluation approaches

随着Web成为世界共享的海量信息库,商务智能发展到了一个新的层次,它集成了Web数据仓库、OLAP和Web数据挖掘等关键技术,使商务智能系统不仅在深度上满足商务组织内部的各种需求,而且在广度上能快速集成全球相关信息。掌握国内外的市场发展变化趋势,制定合理的政策,使组织在全球化的竞争机制中能"知己知彼,百战不殆"。市场权威调研机构Gartner公司在2008年出版的IT战略技术预测和展望报告[1]中将"面向Web的架构"和"商务智能"列为2009年世界十大关键技术。中国也被列为亚太地区商务智能市场增长的领军者。

1 高质量的信息是信息智能管理和决策系统成功的关键

以万维网为基础的商务智能(Web-based Business Intelligence)帮助商务组织的决策层获得高质量的集成信息,产生可执行的知识,做出合理的计划和准确的决策[2]。这类技术在关乎国家发展的重大领域,如金融、政务、电信、证券、交通运输等领域中发挥着不可估量的重要作用,今后这项技术还将渗透到各行各业,在国民经济的发展中有着广阔的应用前景。

我国光大银行早在2004年就成功实施了银行业商务智能应用。这套系统提供了强大的数据整合、即时查询、多维数据分析和统计、图形展示、数据挖掘、智能报警等功能,以及自定义查询、报表和分析的零编程和多项OLAP引擎扩展等特有功能,使光大银行可以深入分析运营数据,挖掘未知模式,使各级管理层和各级业务人员可以准确、全面地发现业务中的异常,进一步掌握客户的情况,及时了解竞争对手的发展,大大提高了业务部门的市场敏感性以及管理和决策水平[3]。另一个实例是四川移动的商务智能系统,该系统提供了流失模型的应用流程,具体分为流失状况认知(历史流失状况分析、流失预测目标用户群选择、用户流失倾向预测、预测流失用户细分)、挽留方案设计(挽留目标用户群选择、挽留方案的推荐和设置、挽留目标客户连接配置)、挽留实施与跟踪(外呼调查与管理、外呼反馈分析、挽留跟踪分析)几个功能模块,形

到稿日期:2009-09-28 返修日期:2009-12-29 本文受国家自然科学基金(60573165),教育部留学回国人员基金,西南交通大学科技发展基金项目(2007A14)资助。

朱 焱(1965—),女,博士,教授,主要研究方向为数据挖掘、Web资源质量管理、Web工程、数据仓库系统,E-mail:yzhu@swjtu.edu.cn。

成一个闭环[4]。当进行外部客户管理调查时，通过这个闭环来指导离网模型的选取。在2005年3、4两个月，四川移动应用该功能分别成功挽留了2000和5000个客户。

再以国外Web商务智能应用的成功案例为例。O'Reilly图书出版公司研究开发了一种名为amaBooks的软件工具[5]。该程序每隔3个小时从Amazon公司的网站提取上兆的HTML文本，从中抽取O'Reilly及其它同行所出版书籍的数据并整合到数据仓库中。基于这些数据，O'Reilly公司能够分析当前市场的图书价格规律和畅销书籍的变化情况，分析各出版商的市场份额和图书选题趋势，并且可以在几天之内得出新的趋势分析结果。这无疑使该公司在行业中具有了强大的竞争优势。国外相类似的应用还有财务管理系统Yodlee和Umonitor。

上述例子表明，高质量信息是各类智能信息系统的系统成功实现并可靠运转的先决条件。然而Meta Group的一项调查[6]显示，41%的数据仓库项目无法成功，主要原因是糟糕的数据源质量导致了错误的决策，这种现象被称为"垃圾进，垃圾出"。而Gartner Group在2004年的报告[7]中明确指出，财富1000强公司中25%的关键数据是脏数据，这种现象还将持续存在。不同于传统的数据源(数据文件、数据库等)，基于Web的信息源具有数据海量、高动态性、高自主性、信息类型多样、应用要求不同等特点，这些特点加剧了Web信息源质量问题的严峻性。

因此，要保证基于Web的信息智能管理与决策系统的成功，关键是发现Web信息源的质量模式，分析质量异常对系统的影响，确定并遴选出高质量的Web信息源，并将它们作为智能信息系统的可靠外部数据资源。因此，研究并应用适合的数据挖掘技术捕捉信息质量异常、发现质量隐患是新颖而重要的研究课题。该研究领域被称为质量异常(隐患)模式挖掘。

本文分析了Web资源特点带来的质量问题，综述了国内外Web资源质量评测和管理方法的现状，深入讨论了应用数据挖掘及相关技术发现、处理Web资源质量异常的原理，总结出开展Web资源质量异常挖掘技术研究面临的难点和解决难点的技术方案，为实现高质量的Web信息智能系统应用提供支持。

2 Web信息源的特性分析

Web信息源有4大特性：

(1)Web信息源的高动态性

Web信息源具有高动态性，具体体现在：其信息内容不断地新陈代谢；网站的结构与版面经常修改以满足用户不断变化的需求。这是传统信息源不具备的特点。

(2)Web信息的高自主性

Web信息的自主性表现在：在Web领域，缺少全球性的质量管理机构和统一的质量检查标准，从而导致Web信息源中包含着许多错误的、不完整的、不一致的、冗余的数据或者模糊的、不确切的事实，甚至由于不良的表达(例如：缺少单位或时间标记)，正确的数据也会变得难以使用，集成不同Web信息源间的数据变得十分艰难。而由于制定了完善的规范或开发了配套的工具，在传统的出版物(如书籍、报纸、杂志等)和传统的信息源(如数据库等)中可以对信息进行统一和仔细的评价、检查和校正。

(3)Web信息的海量性和信息类型的多样性

随着Web技术的发展，它已成为全球信息量最大、类型最齐全的信息资源。据wiki百科介绍，2008年7月Google可以搜索到的不重复网页数已达到1兆(万亿)页，每天还以新增几十亿页索引的速度增长，而Web静动态网页的实际数量可能是Google可索引网页的近10倍。此外，Web信息类型十分丰富，常用的有文本、图形图像、音频、视频、动画等，这些信息的数据格式千差万别。这些特点给Web信息提取、净化、集成、转换等工作带来了巨大的困难。

(4)Web信息应用的多样性

Web信息源的用户群体数量和类型急剧增加，各类用户群体在技术背景、使用Web信息的目的和方式上都有很大区别[8]。例如，使用Web股票证券信息的用户对可信度和时效性方面的质量要求很高；基于Web的零售企业收集某一地区的人口统计信息，以便有针对性地制定营销方案，他们对信息的可信度、完整性和集成度有很高要求。显然，Web信息应用的多样性导致了对Web信息源质量的多方位要求。

3 Web资源质量模式挖掘及相关技术分析

信息质量管理一直是信息系统领域的核心问题，受到了广泛的关注[9-11]。万维网是新型信息源，具有特殊性。近年来国内外开始了对Web资源质量问题的研究，目前深入的研究成果还十分少。本文根据现有的研究成果，将Web资源质量挖掘及相关技术的研究现状归纳为以下4个方面。

3.1 数据质量模式挖掘

传统的数据质量或软件质量管理技术偏重于定性分析，具有很大的主观性。研发好的质量度量方法，(半)自动化地捕捉数据质量问题，客观地确定质量级别是学者们长期探索的问题。随着数据挖掘技术的发展，数据质量模式挖掘开始受到关注。数据质量模式挖掘旨在数据集合中发现质量方面的异常模式，文献[12-14]应用数据挖掘技术进行了这方面的尝试。Luebbers等人[12]致力于发展一个域驱动的方法，以完成传统数据集的数据质量核查工作，他们应用了数据挖掘和其他相关技术，发现真实数据与预测值之间的差异，以检查出每一数据项的偏差。在QUIS数据库系统中[13]，作者建立了符合ISO9001的质量管理体系，应用离群点检测、决策规则、统计等数据挖掘方法进行数据库数据质量管理。而Hipp等[14]讨论了应用数据挖掘技术检测、量化、解释和修正数据质量的理论化方法，并研究了应用关联规则检查数据库事务的例子。

当前数据质量模式挖掘的研究多数集中在传统的数据类型上，如传统数据集[12]、数据库[13,14]，以及信用卡欺诈检测等[15]。

将质量模式挖掘技术深化到Web信息质量管理领域是一个刚刚兴起的重要研究课题。在该领域中，主要以经典数据挖掘理论和技术的研究为基础，设计开发针对Web资源质量的新型技术，扩展或组合经典方法，挖掘质量数据集合中的小众模式(各种质量异常)，旨在自动发现未知的质量异常。WCOND-Mine项目[16]聚焦于Web内容离群点的挖掘算法研究，目的是标示竞争对手和商业趋势，提高搜索结果的质量。而Agichtein等[17]通过对百科问答式网站上用户产生文

本中的错误进行标识，辅助投票机制，对不同质量级别的文本内容进行分类。

由于 Web 资源是新型信息源，具有动态性、信息类型多样、数据维度高等特性，Web 资源质量模式挖掘尚有 5 个主要问题亟待解决。

(1)原始数据质量定义问题。万维网提供大量信息供阅读和进一步的信息处理，这些信息可视为 Web 原始数据。质量模式挖掘技术并非应用于原始数据之上，而是以原始数据的质量数据为操作对象，因此需要建立一个良构的、集成的、优质的 Web 质量元数据集合，该集合中的元数据需要从多个方面定义，如数据内在的质量(正确性)、数据表达质量(缺少度量单位)和数据应用质量(容易获取)等。

(2)质量数据的度量和获取问题。在 Web 原始数据基础上获取并量化质量元数据是一个困难的过程，因为质量数据或是零散的、模糊的、隐含在原始数据中，或是不完整的，甚至完全缺失的，需要研发有针对性的技术和工具。例如，网页上有大量产品介绍内容，但其中的文字错误率需要通过统计和计算才能得到。再如，产品售价的货币单位常常缺失，面向国际的电子商务网站没有提供人民币与国际主要币种的现时兑换比价或是缺少指向中国银行相关网址的链接。

(3)基准测试数据集问题。Web 质量模式挖掘技术需要在基准测试数据集上进行验证和性能完善，才能应用到上述真实的数据集合上。基准测试数据集是人为按比例设置、具有确定质量高低类别的试验用数据集合，是验证 Web 质量数据挖掘技术性能的试金石。然而迄今为止，在 Web 质量挖掘领域还没有这样的基准测试集。

(4)开发或拓展数据挖掘技术问题。Web 资源中由于各种因素相互影响，质量数据模糊性很明显，需要研究新的技术，以解决不确定性条件下质量异常/隐患发现问题。此外，相比传统数据信息源，Web 信息源的动态性和自主性使得 Web 质量有波动性。如何开发相关技术，(半)自动监测 Web 信息源质量，动态跟踪质量的变化，正确反映不同时间段中 Web 信息源的质量状况是又一个难题。

(5)复合技术问题。需要组合多个学科的技术完成质量元数据量化、数据挖掘、综合判断，以遴选出高质量的 Web 信息源。

3.2 多准则质量评测研究

3.2.1 关于评测指标体系

在 Web 使用的早期，信息和图书管理领域的学者从用户浏览 Web 信息的角度提出了一系列的质量评测指标，例如，OASIS 体系[18]提出的客观、准确、来源、信息、覆盖范围 5 个评测指标。Stoker 和 Cooke 在文献[19]提出了 8 项指标：权威性、信息来源、范围和论述、文本格式、信息组织方式、技术因素、价格和可获取性、用户支持系统。文献[20]通过对 Web 用户问卷调查，总结了高质量的网络信息资源应具备的条件。

随着对 Web 质量评测的深入研究，学者们认识到 Web 信息质量管理还缺乏理论框架的指引，以后的研究工作尝试在其他领域已经成熟的质量评价体系基础上，建立 Web 资源质量的评测指标，例如，MIT 的 TDQM 项目[21]建立的软件系统质量管理评测指标被许多学者所借鉴。

Calero 和 Kabassi 根据 Web 商务应用要求建立了质量模型[22,23]；文献[24,25]研究了基于软件质量标准 ISO/DEC 9126 的 Web 信息源质量建模，给出了多类多层次的 Web 质量的评测指标。Mich 等在文献[26]中应用经典的 Ciceronian loci 作为其 2QCV3Q 评测模型的理论依据。文献[27]结合 ISO/DEC 9126 与 W3C 的技术推荐定义了 FQT4Web 质量模型。

目前，Web 质量评测指标研究还存在以下问题。

· 评价指标体系不够健全。Web 质量指标体系的研究很多，但不论是从纸质出版物的角度出发，还是以已建立的其他对象的评价体系为基础，至今还未形成一个公认的质量评测模型。

· 自动化的 Web 质量评价系统开发困难。某些指标体系较庞大，具有较好的覆盖面，但不具备可使用性和可度量性，难以开发(半)自动化的 Web 质量评价系统。

· 某些指标设计不统一。有些指标的设计是从 Web 浏览者的角度，而有些指标的设计是从 Web 设计者的角度，还有指标的设计是从自动化管理 Web 信息，在高质量信息上开发深层应用的角度。有些指标互相不独立，存在交集。显然，指标设计角度不同，指标的类型和所在的层次会有不同，评测的方法也不同，直接影响指标的量化取值和质量评价。

3.2.2 关于 Web 资源质量评测的需求

对 Web 资源进行评测有不同的用户需求。

(1)普通用户。据中国互联网信息中心(CNNIC)2009 年 7 月发布的统计报告[8]显示，中国普通用户使用率最高的 3 类网络应用是网络娱乐、信息获取和交流沟通。这类用户对 Web 资源质量的要求集中在信息内容的时新性和多样性、资源的可达性和快速性、服务方便易用等方面。需要用户制定评测准则，参与评测过程，将用户的评价结果反馈给网站，帮助网站改进质量。

(2)网站开发者。网站开发者更多地是从技术层面比较网站的质量，从功能性、效率、数据库设计、可维护性、易升级、网站安全等方面对所开发网站进行改进和优化。

(3)商务智能决策者。商务智能决策者代表着各种商务组织，这些组织既开发自己的网站，又需要从其他 Web 资源中获取信息，以了解市场和业界的布局和发展，获得高度的竞争智能。他们不仅从技术层面上，更要从系统和策略的层面上评测 Web 资源质量，以便以高质量的信息为基础制定出合理的商务政策。

显然，第一种需求注重 Web 资源的使用质量，第二种需求强调网站性能质量，第三种需求聚焦在整合 Web 资源，获取竞争信息，提高商务智能整个过程中的 Web 资源质量，需要从网站性能、信息内容和商务使用多个方面考量 Web 资源的质量。下面着重分析第三种需求。

3.2.3 关于质量评测方法

Web 质量评测方法目前有定性评价、定量评价、综合评测 3 类。

(1)质量定性评价。常采用用户问卷调查、投票评比、电话调查、网下抽样等方式来评价网站质量。例如，为了完成中国互联网发展状况统计报告，CNNIC 采用了电话抽样调查、网上问卷调查、网上自动搜索、相关单位上报统计数据等方式[8]。但问卷的设计、调查方法、样本数量、样本分布、各种误差、被调查者的主观判断等多种因素会影响 Web 质量定性评

价的可信度、全面性和准确性。

(2)质量定量评价。网站访问量统计、网站相应时间和超链分析技术是定量获得网站质量评价结果的常用方法,一些产品评价网站基于这类技术对 Web 资源进行评测,能获得较客观的结果,例如易比网采用这类技术对不同类型网站进行了评测排名,如图 1 所示。然而由于对网站访问量和独立用户的定义还不一致,网络存在单 IP 地址多用户、单用户多 IP 地址,以及使用代理等复杂因素,定量评价网站质量方法在评测的全面性、准确性等方面还存在问题。

(3)质量综合评测。质量综合评测方法结合了定性、定量方法的长处,可以较全面地考量 Web 资源质量。该类方法具有几个共同的特性。

· 多准则:准则是评测的度量参数,包含评测的目标和属性。目标是质量评测目的的量化表达,而属性是可测量的量值,该量值反映了达到质量评测目标的程度。需要多个属性评测待评对象,以达到一个或多个目标。

· 准则间的冲突:多个准则可能互相冲突,评测需要考虑准则间的平衡。

· 准则的权重:选择出的准则都在一个评测体系中有着不同的重要性,其量化值表示为权重值。权重可以直接或间接获得。如果没有权重,所有的准则是等权的。

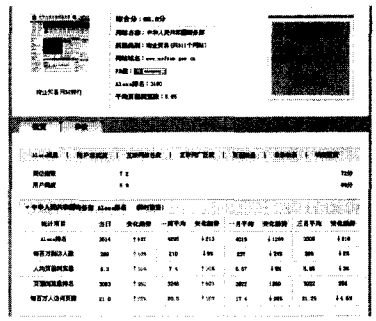

图 1 网站定量评测举例

综合上述特性,可以建立图 2 所示的质量综合评测框架。

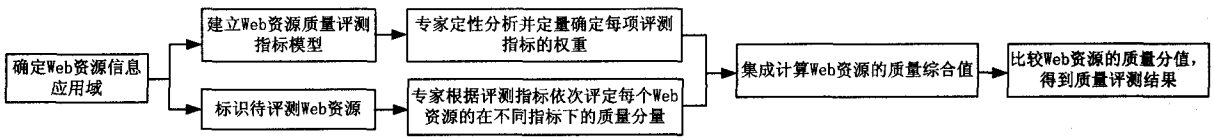

图 2 质量综合评测框架

3.2.4 关于多准则 Web 资源质量评测技术

多准则信息质量评测是综合评测的主要技术,其研究思路是建立一个多层次多维度质量模型。该模型的最上层是评测目标,以目标为指导,将信息源的综合质量向下投影到质量维和质量子维上,获得质量评测准则,形成树状层次结构,例如文献[25]提出的 Web 质量评测模型 WebQM。评测过程是建模的逆过程,首先根据最底层的质量准则评测信息质量,然后应用算法将得到的质量分量逐层向上集成为综合值,根据最后的综合质量值对信息源进行分析比较,确定并遴选最优信息源。

早期的多准则信息质量评测研究成果有麻省理工的 TDQM 项目[21]、Naumann 等的质量驱动数据整合系统[28]和 Dagstuhl 研讨会报告[29]。这些研究工作考虑了信息源质量的多维性和信息源质量分量之间的影响,从而能获得综合的评测结果,但是这些研究仅针对传统的信息源,传统信息资源稳定,质量级别高,数据类型单一,数据有着优化的结构且通常不随时间变化而变化。

应用经典的多准则信息评测技术度量 Web 资源质量已证明是可行的[24-27]。然而 Web 资源质量的模糊性使得运用传统的精确评测技术增加了误差,降低了准确性。Web 资源质量的模糊性具体体现在两个方面。

(1)质量目标和质量属性重要程度的模糊性。主要反映在不同用户或不同应用领域对 Web 资源质量有不同程度的要求,这些要求投影到质量属性上,使得质量属性的重要性程度产生差异,而这样的差异是难以用精确数字度量的。

例如,Web 股票信息用户对信息在"来源可靠性"、"正确性"、"时新性"和"信息获取效率"几个方面的要求比在"表达格式"和"易操作性"方面更重要。而零售业的商务智能系统要获取网上的竞争情报,对信息质量的要求主要侧重在"来源可靠性"、"正确性"、信息源"易操作性"和"易于理解性"。因此在确定质量评测准则重要性权值时,采用模糊的语言词汇比精确的数值更能反映评测者的主观判断。

(2)质量度量的模糊性。某些 Web 资源质量很难用准确的数值度量。例如,Web 资源"运行状态"指出 Web 资源处于活跃状态,由主页面通过超链可达到其所有的链接页面,用户点击超链后,页面的读取时间在用户接受的范围之内。但用户点击超链后,等待某个页面出现的容忍度是因人而异的,因此无法用唯一确定的数值度量这个特点,而用语言词汇(如慢、中、快、极快等)定义 Web 资源的可运行性更合理。

为解决以上问题,将传统的评测方法与模糊逻辑相结合是较好的解决方案,Herrera-Viedma 等人的工作[30]和 Web 信息源质量模糊评测系统[31-33]阐述了这类技术的不同实现。文献[30]从普通用户的角度出发,采用模糊语言模型,应用 LOWA(Linguistic Ordered Weighted Averaging)和 LWA(Linguistic Weighted Averaging)评测方法直接对用户反馈的网站质量的语言评价值进行计算。文献[31-33]以 ISO/IEC 9126 为基础实现了 Web 信息源质量特性建模,在经典的质量多维评测技术 MCDM(Multi-Criteria Decision Making)基础上集成了模糊逻辑,能合理度量 Web 资源质量,特别对评测过程的鲁棒性、评测结果与质量度量参数的关系等方面进行了分析,实现了质量度量参数敏感性分析,并构建了一个 Web 信息源质量基础评测系统。

但上述工作中有两个重要问题尚未解决:

(1)质量多维评测综合技术常常涉及到两项重要指标:质量评测准则的权重和待评测对象在各质量准则下的质量评测分量。由于准则权重赋予和基于准则的质量评测在很大程度上受评测者主观因素影响,即使是采用模糊逻辑技术,评测值的集成计算仍然需要首先去模糊化,这些过程中产生的偏差可能影响到评测结果的正确性,从而降低了质量评测的可信度。如何更客观地评测 Web 质量,减少主观因素的影响是亟待解决的重要问题。

(2)Web 信息源质量评测因素很多,某些关键因素的微

小变化就可能产生完全不同的评测结果，这涉及到评测系统的敏感度或鲁棒性。模糊逻辑的引入大大增加了敏感度分析的难度，如何在质量模糊评测环境下确定敏感因素，优化评测过程，强化系统的鲁棒性是一个重要课题，但还没有对此开展深入的研究。

3.3 Web资源质量元数据管理模式研究

3.3.1 Web资源质量元数据类型及特点

信息的生命周期由信息产生、管理和使用3个过程所组成，因此信息质量问题也出现在这3个过程中。在Web的基本体系中，信息产生和信息管理通常在Web信息源中（服务器端）实现，而信息使用主要是用户从信息源获取信息进行阅读（客户端），或将信息保存以便进一步处理。因此Web质量就映射到了Web信息源质量、Web信息内容质量和Web信息应用质量3个维上，定义为Web资源的质量元数据。这些质量元数据需捕捉Web信息资源内在的质量、信息内容动态与静态质量特点以及用户在信息使用过程中的质量性能，具有多维度多层次的特点。图3给出了质量元数据的多维多层结构。

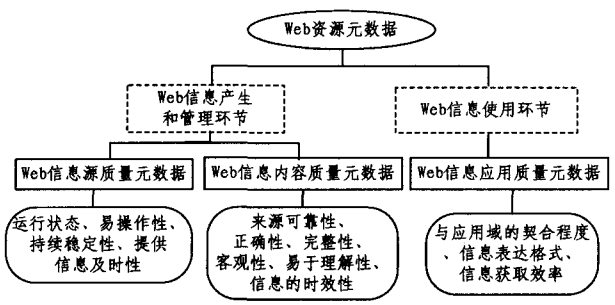

图3 Web资源质量元数据多层次多维结构

（1）Web信息源质量元数据

Web信息源既是信息的产生者又是信息的管理者，质量的重点是信息源的功能是否稳定和可用。Web信息源质量又可定义为4个子类：资源的运行状态、资源的易操作性、资源的持续稳定性和提供信息的及时性。

（2）Web信息内容质量元数据

Web内容质量元数据检查信息内容的各个质量层面。可以分解为以下6类元数据：来源可靠性、内容正确性、内容完整性、客观性、易于理解和信息的时效性。

（3）Web信息应用质量元数据

Web信息应用质量元数据定义Web信息在应用环境中的质量。由于不同的应用有不同的质量要求，选择大部分应用对质量的共性要求如下：信息与应用域具有较高的契合程度，信息表达格式满足应用的处理要求，以及信息获取具有较高的效率。

3.3.2 Web资源质量元数据管理方法

在数据库和数据仓库系统中，元数据在数据提取、数据清洗、数据整合转换等环节中发挥关键作用，在Web信息检索和提取方面元数据提供了信息特征支持[34,35]。已出现的基于元数据模型管理Web资源质量的研究方案是使用元数据机制捕捉质量信息，构建质量元数据（仓）库，并在此基础上利用查询语言查询质量数据，从质量语义的层次对质量进行评测，或挖掘质量模式。研究成果包括欧洲的DWQ项目[36]、Mihaila等的SCQD方法[37]以及Dagstuhl研讨会报告[29]。然而上述研究成果依赖于信息源在发布信息数据的同时也提供质量元数据。这样的假设并不适用于大部分的Web资源，因为在这些信息源中，质量元数据是作为信息（数据）本身的特征或附加数据，常常是模糊且非显式表现的，甚至是完全缺失的。此外，类SQL的查询式质量评测不能发现数据深层次的关系，缺乏探求隐含规律的力度。

文献[38,39]研究了万维网内容中元数据的提取和标注，但其关注的元数据是数据一般特性的描述，没有涉及到质量方面。

Yi和Sundaresan等利用元数据强化Web信息的收集[40]，但元数据内容仅局限于超链接和锚文本，并未涉及完整的Web质量。Hess等[41]实现了3种元数据的处理：（1）捕捉Web服务中的元数据，用各种服务域将它们分类标记，目的是进一步标注服务类型和解释服务操作的含义；（2）根据元数据将Web服务分组；（3）随着Web服务的扩展，指导性分类效率变低时，转而对Web服务聚类，从而创造适合的Web服务类别。虽然文献[41]进行了深入的元数据管理研究，给基于质量元数据的数据挖掘工作提供了很好的经验，但所处理的元数据只是Web服务信息中通用的上下文特性，而不是全面的质量数据。

如何从Web信息源中发现、标记并管理数据质量的元数据仍然是一个尚未深入探讨的课题。

3.4 链结构技术评价Web资源质量的研究

以高效检索Web信息为目的的链结构分析是Web信息源质量模式挖掘的另一着眼点，其评测准则是信息相关性和权威性。这类研究将Web信息源的拓扑结构模型化为一个图结构，结点代表网页，结点间的有向边表示超链。网页间的超链隐含着网页所有者对其他网页内容相关性和权威性的判断。越多的相关网页和权威网页链接某个网页，该网页的相关性和权威性就越高。某网页链接到相关网页和权威网页越多，对该网页的相关性和权威性评价的正效应就越大。

典型的算法是Google的PageRank[42]、Kleinberg的HITS算法[43]及各种改进、IBM Clever项目组的ARC算法[44]。近年来，超链分析或挖掘方法[45-47]受到广泛而深入的研究，覆盖了大量的应用领域，例如万维网图片标注、社会网挖掘、小世界现象、超链发现、组检测、推理与语义网、计算机犯罪与Web安全、基于超链的对象分类等。

然而，与超链有关的质量仅是Web资源的部分质量特征，Web资源质量不仅体现在各类静态指标上（如相关性、权威性、信息正确性等），还反映在其动态指标上（如网页的利用度、网页内容更换频率、超链的可达性等）。显然，早期的超链结构挖掘方法侧重于提高Web信息检索的相关性和检索速度。近年的研究在超链基础上，结合了锚文本、Web内容、语义理解等技术，更多地考虑了Web资源质量的其他指标，但还没有从Web信息产生、管理到Web信息应用三个维度上完整地解决质量异常模式的发现、标识、分类等难题。

结束语 挖掘Web资源质量模式，遴选高质量Web信息源作为各种高端应用中（如商务智能）不可或缺的外部信息源，不仅使商务智能可以在国家重点经济行业发挥积极的作用，支持全球性的业务分析，制定合理的发展决策，而且能将此项技术逐步推广到政务、交通等更多关键领域，在国家各项事业发展中有着极为广阔的应用前景。

Web资源的特性给这种新型信息源的质量管理工作带来了巨大困难，现有的技术和方法还有较多局限性，存在于学术理论、技术和方法中的大量关键问题还没有解决，亟待深入研究。虽然当前在Web资源质量模式挖掘及相关领域已有了一定的进展，但如何突破以往方法的限制，发展新的方法或拓展组合已有技术，管理日益重要的万维网信息源的质量；如何解决关键的理论问题，最大程度地提高Web信息源质量模式挖掘系统性能，提高遴选的准确性、客观性和自动化程度，均是极具研究价值的课题。

参 考 文 献

[1] Gartner Inc. Top 10 Strategic Technologies for 2009 [R]. Gartner Symposium ITxpo, 2008

[2] Srivastava J, Cooley R. Web Business Intelligence: Mining the Web for Actionable Knowledge [J]. INFORMS Journal on Computing, 2003, 15(2): 191-207

[3] 菲奈特. 光大银行商业智能应用案例[R/OL]. 2007. http://www.amteam.org/k/BusinessIntelligence/2007-8/587931.html

[4] 林德. BI系统的九大专题分析[N]. 人民邮电报, 2005

[5] O'Reilly T. Inventing the Future [J/OL]. O'Reilly Network. 2002. www.oreillynet.com/pub/a/network/2002/04/09/future.html

[6] Meta Group. Data Warehouse Scorecard [R]. Meta Group, 1999

[7] Gartner Inc. Using Business Intelligence to Gain a Competitive Edge[R]. Strategic Planning Report, ISBN 0-9741571-1-2. 2004

[8] 中国互联网信息中心. 第二十四次中国互联网络发展状况统计报告[R/OL]. 2009. http://www.cnnic.net.cn/html/Dir/2009/07/15/5637.htm

[9] Kaplan D, Krishnan R, Padman R, et al. Assessing Data Quality in Accounting Information Systems [J]. Communications of the ACM, 1998, 41(2): 72-78

[10] Chengalur-Smith N, Ballou D P, et al. The Impact of Data Quality Information on Decision Making: An Exploratory Analysis [J]. IEEE Transactions on Knowledge and Data Engineering, 1999, 11(6): 853-864

[11] 郭志懋, 周敖英. 数据质量和数据清洗研究综述[J]. 软件学报, 2002, 13(11): 2076-2081

[12] Luebbers D, Grimmer U, Jarke M. Systematic Development of Data Mining-Based Data Quality Tools[C]// Proc. of the 29th VLDB Conference. Berlin: Morgan Kaufmann Publishers, 2003: 548-559

[13] Grimmer U, Hinrichs H. A Methodological Approach to Data Quality Management Supported by Data Mining[C]// Proc. of the 6th Conf. on Information Quality. Cambridge: MIT Press, 2001: 217-232

[14] Hipp J, Guntzer U, Grimmer U. Data Quality Mining: Making a virtue of necessity[C]// Workshop on Research Issues in Data Mining and Knowledge Discovery. Santa Barbara: ACM Press, 2001

[15] Chandola V, Banerjee A, Kumar V. Anomaly Detection-A Survey[R]. To Appear in ACM Computing Surveys, 2009

[16] Agyemang M, Barker K, Alhajj R. WCOND-Mine: Algorithm for Detecting Web Content Outliers from Web Documents[C]// Proc. of the 10th IEEE Symposium on Computers and Communications. IEEE Computer Society Press, 2005: 885-890

[17] Agichtein E, Castillo C, Donato D, et al. Finding High-Quality Content in Social Media[C]// Proc. of the Intl Conf. on Web Search and Web Data Mining. New York: ACM Press, 2008: 183-194

[18] Wilkinson G L, Bennett L T, Oliver K M. Evaluating criteria and indicators of quality for Internet resources [J]. Educational Technology, 1997, 37(3): 52-59

[19] Stoker D, Cooke A. Evaluation of Networked Information Sour-ces [C]// Ahmed H. Helal and Joachim W. Weiss, eds. Information Superhighway: the Role of Librarians, Information Scientists and Intermediaries: Proceedings of the 17th International Essen Symposium. Essen: Universitaetsbibliothek, 1994: 287-312

[20] 董小英, 张本波, 陶锦, 等. 中国学术界用户对互联网信息的利用及其评价[J]. 图书情报工作, 2002, 10: 29-40

[21] Wang R. A Product Perspective on Total Data Quality Management [J]. Communications of the ACM, 1998, 41(2): 58-65

[22] Calero C, Caro A, Piattini M. An Applicable Data Quality Model for Web Portal Data Consumers [J]. World Wide Web, 2008, 11(4): 465-484

[23] Kabassi K, Virvou M, Tsihrintzis G A. Web Services User Model Server Performing Decision Making [J]. International Journal of Pattern Recognition and Artificial Intelligence, 2007, 21(2): 245-264

[24] Olsina L, Rossi G. Measuring Web Application Quality with WebQEM [J]. IEEE MultiMedia, 2002, 9(4): 20-29

[25] 朱焱, 唐慧佳, 马永强. 基于ISO/IEC9126的Web信息源质量评测系统[J]. 西南交通大学学报, 2008, 43(2): 253-257

[26] Mich L, Franch M, Gaio L. Evaluating and designing Web site quality [J]. IEEE MultiMedia, 2003, 10(1): 34-43

[27] Davoli P, Mazzoni F, Corradini E. Quality assessment of cultural Web sites with fuzzy operators[J]. Journal of Computer Information Systems, Fall, 2005: 44-57

[28] Naumann F. From Databases to Information Systems-Information Quality Makes the Difference[C]// Proc. of the 6th Conf. on Information Quality. Cambridge: MIT Press, 2001: 244-260

[29] Gertz M, Oezsu M T, et al. Report on the Dagstuhl Seminar[J]. ACM SIGMOD Record, 2004, 33(1)

[30] Herrera-Viedma E, Peis E, Moralesdel Castillo J M, et al. A Fuzzy linguistic model to evaluate the quality of Web sites that store XML documents[J]. International Journal of Approximate Reasoning, 2007, 46: 226-253

[31] Zhu Y, Buchmann A P. Evaluating and Selecting Web Sources as External Information Resources of a Data Warehouse[C]// Proc. of the 3rd Intl Conf. on Web Information Systems Engineering. Los Alamitos: IEEE Press, 2002: 149-160

[32] Zhu Y. Group Assessment of Web Source/Information Quality Based on WebQM and Fuzzy Logic[C]// Proc. of the 3rd Intl. Conf. on Rough Sets and Knowledge Technology. Heidelberg: Springer-Verlag, 2008: 660-667

[33] 朱焱. 应用模糊分析层次法可靠评测Web资源质量[J]. 计算机科学, 2009, 36(4): 221-223, 267

[34] 王继成, 杨晓江, 潘金贵, 等. 基于元数据与Z39.50的分布协作式Web信息检索[J]. 软件学报, 2001, 12(4): 620-627

[35] 张铭, 银平, 邓志鸿, 等. SVM+BiHMM: 基于统计方法的元数

据自动抽取混合模型[J]. 软件学报,2008,19(2):358-368

[36] Vassiliadis P, Bouzeghoub M, Quix C. Towards Quality-Oriented Data Warehouse Usage and Evolution[C]// Proc. of the 11th International Conference on Advanced Information Systems Engineering. Heidelberg: Springer-Verlag, 1999: 164-179

[37] Mihaila G A, Raschid L, Vidal M E. Using Quality of Data Metadata for Source Selection and Ranking[C]// Proc. of the 3rd Intl. Workshop on the Web and Databases. Heidelberg: Springer-Verlag, 2000: 93-98

[38] Mayer M, Karkaletsis V, Archer P, et al. Quality Labeling of Medical Web Content [J]. Health Informatics Journal, 2006, 12 (1): 81-87

[39] Vadrevu S, Nagarajan S, Gelgi F, et al. Automated metadata and instance extraction from news Web sites[C]// Proc. of the IEEE/WIC/ACM Intl. Conf. on Web Intelligence. Los Alamitos: IEEE Press, 2005: 38-41

[40] Yi J, Sundaresan N, Huang A W. Using Metadata to Enhance a Web Information Gathering System[C]// Proc. of the 3rd Intl. Workshop on the Web and Databases. Heidelberg: Springer-Verlag, 2000: 38-57

[41] Hess A, Kushmerick N. Learning to attach semantic metadata to Web services[C]// Proc. of the 2nd Semantic Web Conference. Heidelberg: Springer-Verlag, 2003: 258-273

[42] Brin S, Page L. The Anatomy of a Large - Scale Hypertextual Web Search Engine[J]. Computer Networks, 1998, 30(1-7): 107-117

[43] Kleinberg J M. Authoritative Sources in a Hyperlinked Environment [J]. Journal of the ACM, 1999, 46(5): 604-632

[44] Chakrabarti S, Dom B, Raghavan P, et al. Automatic resource compilation by analyzing hyperlink structure and associated text [J]. Computer Networks, 1998, 30(1-7): 65-74

[45] Adibi J, Chalupsky H, Grobelnik M, et al. KDD-2004 Workshop Report[J]. SIGKDD Explorations, 2004, 6(2): 136-139

[46] Borodin A, Roberts G O, Rosenthal J S, et al. Link Analysis Ranking: Algorithms, Theory, and Experiments [J]. ACM Transactions on Internet Technology, 2005, 5(1): 231-297

[47] Getoor L, Diehl C P. Link Mining: A Survey[J]. SIGKDD Explorations, 2005, 7(2): 3-12

(上接第 188 页)

维网格算法在消息长度较小进程数较少时表现尚可,但随着消息长度和进程数的增多用时增长剧烈;2)消息长度较小时,深腾 7000 上表现较好的是简单算法和 Bruck 算法,曙光 5000A 上表现较好的是原始算法和 Bruck 算法。值得注意的是成对交换和环算法在两个系统的短消息通信性能都不错;3)对于中长消息,深腾 7000 上表现较好的依次是成对交换、环算法和简单算法,随着消息长度进一步增加成对交换优于其他两种算法;曙光 5000A 上表现较好的依次是环算法、成对交换和原始算法,但是成对交换算法的进程扩展性优于环算法,同时中等数据长度的原始算法进程扩展性较差;4)简单算法在深腾 7000 上表现较好,但在曙光 5000A 上性能一般,但是较稳定。简单算法的用时一般情况下较环算法、成对交换算法、原始算法的多是因为存在节点冲突。

结束语 本文测试和分析了曙光 5000A 及深腾 7000 上主流 Alltoall 算法不同消息长度和不同进程数的性能。由递归倍增以及二维三维网格算法的测试结果可知,以较多的消息传递减少启动时间在曙光 5000A 以及深腾 7000 上并不能提高 Alltoll 的性能,为此集合通信操作应尽量避免额外的通信。消息长度较大时,消息的传输时间占主导地位,减少网络和节点等冲突是提高性能的关键,使两个系统的成对交换和环算法的性能都较好。

Alltoall 的优化是未来的主要工作,包括稀疏 Alltoall 和数据长度不等的 Alltoallv。调用非阻塞点对点操作的原始算法在测试中表现较好,因此非阻塞集合通信[10]也是重要的探索方向。

参 考 文 献

[1] Thakur R, Rabenseifner R, Gropp W. Optimization of Collective Communication Operations in MPICH[J]. International Journal of High Performance Computing Applications, 2005, 1(19): 49-66

[2] Faraj A, Yuan Xin. An Empirical Approach for Efficient All-to-All Personalized Communication on Ethernet Switched Clusters [Z]. ICPP, 2005: 321-328

[3] 陈靖,张云泉,张林波,等. 一种新的 MPI Allgather 算法及其在万亿次机群系统上的实现与性能分析[J]. 计算机学报,2006,29 (5):808-814

[4] MPICH-A portable implementation of MPI[OL]. http://www.mcs.anl.gov/mpi/mpich

[5] LAM/MPI Parallel Computing. http://www.lam-mpi.org/

[6] Kale L V, Kumar S, Vardarajan K. A framework for collective personalized communication[C]// Proceedings of the 17th International Parallel and Distributed Processing Symposium(IPDPS '03). 2003

[7] Bruck J, Ho C-T, Kipnis S, et al. Efficient algorithms for all-to-all communications in multiport messagepassing systems[J]. IEEE Transactions on Parallel and Distributed Systems, 1997, 8 (11): 1143-1156

[8] Sun Jiachang. Multivariate Fourier Series over a Class of non Tensor-product Partition Domains[J]. Journal of Computational Mathematics, 2003, 12(1): 53-62

[9] 姚继锋,孙家昶. 平行十二面体区域上的快速离散傅立叶变换及其并行实现[J]. 数值计算与计算机应用,2004,25(4):303-314

[10] Hoefler T, Lumsdaine A, Rehm W. Implementation and performance analysis of non-blocking collective operations for MPI [C]// Proceedings of the 2007 ACM/IEEE conference on Supercomputing. Reno, Nevada, November 2007