基于RSS的OAI框架中元数据同步问题解决方法

王秀慧,陈立潮,谢斌红,袁 英
(太原科技大学 计算机科学与技术学院,山西 太原 030024)

摘 要:OAI协议的飞速发展促使了仓储数目的不断增大,导致服务提供者在短时间内无法及时收割数据提供者仓储更新的元数据记录。文中将RSS技术应用到OAI协议中,提出了一种OAI框架中元数据同步问题的解决方法。该方法以XSLT为基础,将数据提供方的元数据记录转换为对应的聚合文件(RSS1.0和RSS2.0),利用RSS技术的即时性、标准统一、易于整合等特点,有效地维护了数据提供方和服务提供方元数据的同步,从而实现了两方信息的同时更新并提高了信息的利用率。

关键词:OAI;RSS;元数据;同步
中图分类号:TP391 **文献标识码**:A **文章编号**:1673-629X(2009)08-0240-03

A Method to Solve Problem of Metadata Synchronization in OAI Framework Based on RSS

WANG Xiu-hui, CHEN Li-chao, XIE Bin-hong, YUAN Ying
(Institute of Computer Science and Technology, Taiyuan University of Science and Technology, Taiyuan 030024, China)

Abstract: The rapid development of OAI-PMH leads to the growing number of repository, which introduces the problem of synchronization. Stated simply, this problem arises when SP can not harvest DP's newest metadata records timely. In this paper, a method is proposed by using an XSLT-based transformation mechanism to transform DP's metadata records into syndication document formats in the OAI framework to solve the problem. Use the characteristics of real time, standard format and easily integration in the RSS to maintain metadata synchronization and keep resource consistent between DP and SP.

Key words: OAI; RSS; metadata; synchronization

0 引 言

随着信息技术的飞速发展,人类在认识世界、改造世界的科技活动中积累的科学数据不断地膨胀。对于这些分散在不同领域、不同地区的科学数据,将其有效地管理和组织,实现全社会的资源共享,能够使科学数据在全社会得到广泛的传播和应用,进而对科技创新、经济发展和社会进步起到不可估量的推动作用。因此,近年来开始投入大量资金建设科学数据共享工程。而科学数据共享工程的一个关键问题是要确保科学数据能够最及时地为领域专家等不同类型的用户提供服务,以充分发挥科学数据的价值。

收稿日期:2008-12-16;**修回日期**:2009-03-05
基金项目:太原市科技项目(ZB0701-04);太原科技大学青年基金(2007133)
作者简介:王秀慧(1981-),女,硕士研究生,CCF会员,研究方向为数据挖掘与人工智能;陈立潮,教授,研究方向为人工智能与模式识别、智能软件。

科学数据共享平台能够对分散在各地、各个领域的科学元数据收割,将收割回的数据进行加工处理后存储在科学数据中心,进而为用户提供搜索等增值服务。为了保证共享平台提供给用户的信息是各个领域的最新科学数据,需要解决数据提供者和服务提供者间元数据的同步问题。

针对基于OAI框架下的元数据同步问题,提出了一种基于RSS技术的OAI框架中的元数据同步问题解决办法。即利用RSS技术的及时性、标准统一、易于整合等特性来有效维护数据提供者和服务提供者元数据的同步更新,从而尽量保证用户在任何时候检索到的信息都是最新的。

1 OAI协议中同步问题

OAI-PMH(Open Archives Initiative Protocol for Metadata Harvesting,简称OAI协议)是近几年提出来的一种元数据互操作协议,为实现元数据的传播、共

享、分发与互操作提供了新的解决办法。由于 OAI 协议具有简单、低门槛、跨平台等优点,逐渐被应用于资源整合、跨库检索、文献传递、学科信息门户建立、个性化服务等各大领域[1,2]。

如图 1 所示,OAI 协议规定了两个角色,数据提供者(Data Provider,DP)和服务提供者(Service Provider,SP)[3,4]。DP 维护着一个或多个仓储,对来自服务提供者的请求做出响应,以 OAI 要求的格式(XML)向服务提供者提供元数据。SP 通过 OAI 协议收割来自不同 DP 的元数据记录并存储在数据库中,为用户提供查询等增值服务。从用户角度看,希望能够从 SP 中检索到各个仓储中的最新元数据,因此维护 DP 和 SP 间元数据的同步,进而确保 SP 所提供的信息的新颖性是 OAI 协议的一个关键问题。只有保证了 DP 和 SP 元数据的同步,才能保证 SP 服务用户的信息的新颖、及时、有效、准确。

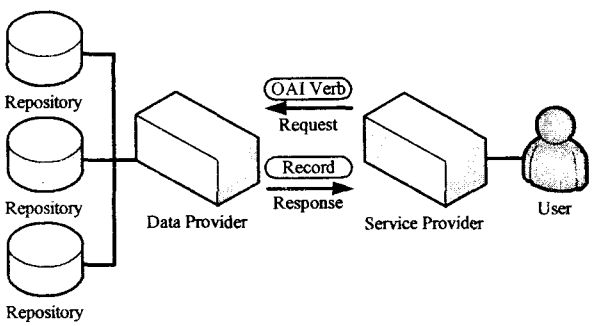

图 1　OAI 协议工作流程图

2　同步问题解决方法

目前有两种方法来实现 DP 和 SP 元数据的同步。

其一是 SP 方指定一个定时收割时间,每隔一定的时间向 DP 方发送命令来收割元数据记录,以确保及时更新 SP 中的元数据。然而定时收割时间的确定对维护系统同步至关重要,定时收割不足,则不能保证 DP 和 SP 元数据记录的一致,进而影响到用户所搜索信息的新颖度和准确度。定时收割过于频繁,则会加重系统负担,浪费系统资源,甚至造成网络堵塞。这要求 SP 应根据不同的仓储设定不同的收割时间。

其二是把注册服务器充当 DP 和 SP 信息反馈的中介[5],当数据提供者更新了它的元数据后,立即向注册服务器进行说明。当注册服务器获知更新信息后,向服务提供者进行信息推送,服务提供者被告知有元数据更新后,启动更新线程去收获元数据。但是由于该方法要求在数据有更新的时刻三方服务器都必须正常运行,而且,在通知过程中又借助了中介,影响了信息更新传播的速度,因此存在一定的弊端。

3　基于 RSS 技术的同步问题解决方法

针对目前元数据同步问题解决方法的不足,提出将 RSS 技术应用到 OAI 框架中,有效地维护 DP 和 SP 的同步,确保元数据的一致。

3.1　RSS 技术

RSS 作为一种基于 XML 的全新信息传播方式,已经在网络信息构建中得到越来越广泛的应用[6]。RSS 有三大优点:

(1)即时性:对于 RSS feed 的订阅者来说,能以最快的方式得到最新消息,而不用被动地到各网站搜索。

(2)标准统一:虽然由于 RSS 在发展过程产生了两大阵营:RSS0.9x/2.0 版本和 RSS1.0 版本,但每一版本都规定了统一的元数据规范,便于解读与管理。

(3)易于整合:RSS 文件以 item 元素表示各个不同的标题与内容,因此易于通过 RSS 阅读器进行快速整合,以选择最合适信息。

正是由于 RSS 具有即时性、标准统一、易于整合的特点,只要有新内容在服务器数据库中出现,该内容便第一时间被"推"给用户,极大地提高了信息的时效性和价值。因此将其应用到 OAI 协议中,能有效维护 DP 和 SP 之间元数据的同步。

3.2　同步更新机制

考虑到 RSS 与 OAI 协议中元数据的制定标准不一致,DP 需对仓储中的元数据进行格式转换。此外,RSS 在实现即时更新过程中需要搭载合适的时间参数,这要求 DP 需提供更新频率以便于 SP 针对不同的 DP 来调整自己收割时间。因此对同步问题的实现应该按如下步骤来进行。

(1)元数据格式转化。

OAI-PMH 把 Dublin Core (DC)作为互操作的标准元数据[7],但是由于 DC 的 15 个元素不能很好满足需求,通常情况下 OAI 协议中的元数据记录都对 DC 进行了扩展。典型的扩展方式是增加<datestamp>和<request>元素,<request>元素代表所请求的元数据的 URL,将其对应到 RSS1.0 和 RSS2.0 标准的 channel 的 link 子元素。datestamp 代表本条元数据的最近更新时间,对应到 RSS2.0 channel 的 lastBuildDate 元素,而 RSS1.0 本身不存在对应元素,需引入 syndication 标准 module,并将 datestamp 对应到 Sy:updateBase。

DC 与 RSS 各元素格式转换对照表如表 1 所示,从该表可以看到,DC 各元素与 RSS1.0 标准能够很好的匹配,RSS2.0 则不然,DC 元素集中的<contributor>、<relation>、<format>、<type>与<coverage>等元素,在 RSS 2.0 中没有存在适当的对应元素。然

而就 Service Provider 与 Data Provider 间元数据一致性问题,这几个元素并不重要,在转化过程中可以忽略掉。如果用户需要查看某条元数据的完整信息,可通过<dc:source>和< link>提供的链接来访问。

表 1 格式转化表

OAI Record	RSS 1.0	RSS 2.0
identifier	dc:identifier	guid
date	dc:date	pubDate
title	title	title
creator	dc:creator	author
publisher	dc:publisher	title
contributor	dc:contributor	–
descripton	description	description
subject	dc:subject	category
source	link	link
language	dc:language	language
rights	dc:rights	copyright
format	dc:format	–
type	dc:type	–
coverage	dc:coverage	–
relation	dc:relation	

(2)确定更新频率。

SP作为 RSS 技术中的订阅方,维护着一个 RSS feed 地址列表,列表内容即为仓储对应的 RSS 文件的链接。SP 通过收割器定时收割 RSS 地址内容,而定时收割时间取决于仓储元数据变化频率。为了得到 DP 方仓储的更新时间间隔,特定义了如下参数。假如 $\{r_1, r_2, \cdots, r_n\}$ 为所要研究的 M 个仓储,$\triangle t$ 表示固定的时间间隔,$\{t_1, t_2, \cdots, t_n\}$ 表示所观察的 n 个时间所要研究的 M 个仓储,$\triangle t$ 表示固定的时间间隔,$\{t_1, t_2, \cdots, t_n\}$ 表示所观察的 n 个时间点,且有 $t_{j+1} = t_j + \triangle t$。对于任意 t_j,t_c 表示处于 t_j 之前的仓储最近的修改时间[8]。

仓储更新状态(S):表示仓储 r_i 在时间 t_j 的更新状态,如果已更新则取值为 1。

$$S(r_i;t_j) = \begin{cases} 1 & \text{若仓储 } r_i \text{ 更新} \\ 0 & \text{其它情况} \end{cases} \quad (1)$$

记录更新速率(R):$R(r_i, t_j)$ 表示仓储 r_i 在时间 t_j 所更新的记录的条数。

仓储更新时间间隔(I):表示仓储 r_i 在时间 t_j 记录更新时间间隔。

$$I(r_i;t_j) = \begin{cases} 0 & \text{若 } S(r_i;t_j) = 0 \\ j - c & \text{若 } S(r_i;t_j) = 1 \end{cases} \quad (2)$$

平均仓储更新时间间隔(I):表示在观察期间 $\{t_1, t_2, \cdots, t_n\}$ 内,仓储 r_i 记录更新平均时间间隔。

$$U(r_i) = \frac{\sum_{j=1}^{n} I(r_i;t_j)}{\sum_{j=1}^{n} S(r_i;t_j)} \quad (3)$$

(3)同步实现。

定义一个 XML schema 用于描述数据提供方的基本情况。DP 在得到合适的时间参数后,根据该 schema 生成 XML 文件。文件首先包含仓储的基本信息,比如 URL,所采用的协议版本,E-mail 地址等,其次包含元素 UpdatePeriod, UpdateFrequency 和 UpdateBase,分别代表仓储的更新周期、更新频率和更新基准时间。SP 通过收割器向 DP 发送请求命令 Identify 以获取上述 XML 文件,从中解析出仓储的更新时间间隔和 RSS 链接地址,然后按照如下所示算法更新内容,将返回的 XML 文件解析后存入到数据库表中,进而为用户提供查询等增值服务。由于 SP 针对不同的仓储调整收割时间,使得更新频繁的仓储有密集的收割次数,较少更新的仓储能适时地对其修正,因此保证了在流量和系统负担都能得到控制的前提下,维护了 DP 和 SP 资源的同步,从而使得用户能够从 SP 中获得各个仓储的最新信息。伪代码如下:

```
Input:RepositaryList = {r1, r2, ···, rn}
  AverageUpdateInterval = {u1, u2, ···, un}
  LastHarvestTime = {t1, t2, ···, tn} = null;
  RssUrlList = {a1, a2, ···, an} = null;
  Procedure
  For(i = 1; i ≤ n; i ++){
    ai = getrssturl(ri);
  }
While(true){
    For(i = 1; i ≤ n; i ++){
      If(currenttime ≥ ui + ti){
        Getmetadatabyurl(ai, ti);
        ti = getresponsetime(ai);
      }
  }
Sleep(pre_defined_interval);
}
```

4 结束语

把 RSS 技术应用到 OAI 框架中,设计开发了一个原型系统来解决 DP 和 SP 元数据的同步问题。实验结果表明,该系统性能优良,在流量和系统负担都能得到控制的前提下将数据提供者信息的更新快速反映到相关服务提供者,有效确保了 DP 和 SP 元数据的同步,使得用户在任何时候都能够搜索到最新信息,从而

(下转第 246 页)

较。由表 2 可见,RBF 网络模型预测的相对误差均小于多元回归模型,具有更高的预测精度。

表 2 RBF 模型与回归模型的比较

Number	Springback (FEA)△θ(°)	RBF model		Regression model	
		△θ(°)	η(%)	△θ(°)	η(%)
1	8.59	8.62	0.41	9.61	11.83
2	0.41	0.50	21.09	0.28	31.71
3	1.44	1.49	3.57	1.05	27.08
4	-1.20	-1.16	3.33	-1.07	10.58
5	-0.76	-0.82	7.82	-0.83	9.69

训练得到的 RBF 网络用作函数发生器,来计算每组工艺参数条件下的回弹值。进化策略的种群大小为 50,经过 428 步的迭代,优化搜索达到收敛。最优的工艺参数如表 3 所示。与有限元分析、一般的进化策略方法的优化结果相比,改进的进化策略方法可以得到更好的优化结果。利用 RBF 网络与 ES 方法相结合得到了回弹最小的优化工艺参数。

表 3 优化结果的比较

	t(mm)	r/t	c/t	h/t	σ_s/E	Δθ(°)
FEA	0.5	8	1.1	10	0.78	0.22
NORMAL ES	0.73	7.93	1.06	12.91	1.27	0.22
IMPROVED ES	0.92	7.33	1.16	16.91	1.76	0.20

5 结束语

提出了一种将数值模拟、RBF 网络和进化策略相结合的板材成形优化设计模型。利用神经网络的逼近功能,结合有限元数值模拟,以板材弯曲回弹量最小为目标,完成了板材成形工艺的优化设计。从前面的分析可得出以下结论:

1)RBF 神经网络对回弹这类高度非线性问题具有良好的逼近能力[9]。与 BP 神经网络相比,RBF 网络通过较少的迭代次数,就能得到更高精度的预测结果。与基于数值统计的回归方法相比,具有更高的预测精度。

2)由于 RBF - ES 方法不需要目标函数的梯度信息,以及全局搜索特性,对于存在不可微的目标函数的非线性优化问题,能以较快的速度和较大概率收敛于全局最优解。

3)对进化策略方法进行了改进,实例分析证明,该算法全局搜索能力强,不易陷入局部最优,能得到较高精度的优化解。

4)RBF 神经网络与进化策略具有并行处理特性,将两者结合可达到快速准确优化设计,适用于金属塑性成形等领域的优化设计。

参考文献:

[1] Mehrotra K, Mohan C K, Ranka S. Elements of Artificial Neural Networks[M]. [s.l.]: MIT Press, 1997.

[2] 易荣贵,罗大庸. 基于遗传算法的物流配送路径优化问题研究[J]. 计算机技术与发展,2008,18(6):13 - 15.

[3] 谢红薇. 基于双基因变异方式的混合进化策略[J]. 微计算机信息,2008,24(2 - 3):230 - 232.

[4] 余永权. 神经网络模糊逻辑控制[M]. 北京:电子工业出版社,1999.

[5] 戴 葵. 神经网络实现技术[M]. 长沙:国防科技大学出版社,1998.

[6] 党建武. 神经网络网络技术及应用[M]. 北京:中国铁道出版社,2000.

[7] Bose N K, Liang P. Neural Network Fundamentals With Graphs, Algorithms and Applications[M]. [s.l.]: Tata McGraw - Hill, 1998:407 - 440.

[8] 邵鹏飞,王秀喜,车 玫. 板料成形中的回弹计算和模具修正[J]. 机械强度,2001,23(2):187 - 189.

[9] 胡 平. 一种板材成形压机速度敏感性的描述方法[J]. 中国机械工程,2003,14(12):106 - 107.

(上接第 242 页)

提高了信息的时效性和价值。

参考文献:

[1] Lagoze C. The Open Archives Initiative for Metadata Harvesting[EB/OL]. 2004 - 12 - 25. http://www.openarchives.org/OAI/openarchivesprotocol.html.

[2] Liu X. Federating Heterogeneous Digital Libraries by Metadata Harvesting[D]. [s.l.]: Department of Computer science, Old Dominion University, 2002.

[3] 张佩毅,丰 艳,陈一民. 基于 OAI 协议的数字资源互操作平台[J]. 计算机应用与软件,2007,24(10):46 - 47.

[4] 牛振东,丘俗标,朱先忠,等. 基于 OAI—PMH 的服务提供者的设计与实现[J]. 北京理工大学学报,2004,24(5):411 - 415.

[5] 朱遵学,徐汝兴,郑巧英. OAI 注册服务器功能的探讨[J]. 图书馆杂志,2004,23(8):61 - 62.

[6] 张会娥. RSS 的应用研究[J]. 图书馆杂志,2005,24(2):53 - 58.

[7] Liu Xiaoming, Maly K, Zubair M, et al. DP9: An OAI Gateway Service for Web Crawlers[C]//Proceedings of the 2nd ACM/IEEE - CS joint conference on Digital libraries. Portland, Oregon, USA: [s.n.], 2002.

[8] Liu Xiaoming, Maly K, Zubair M. et al. Repository synchronization in the OAI framework [C]//Proceedings of the ACM/IEEE Joint Conference on Digital Libraries. Houston: [s.n.], 2003.