基于 Web Services 的 Web 挖掘实现方案*

李艳霞，巩九洲，黎玉琴

(西北师范大学 数学与信息科学学院，甘肃 兰州 730070)

摘 要：随着信息技术的发展，Web 挖掘技术已成为数据挖掘技术的研究热点。本文针对 Web 挖掘中数据库环境的异构和信息半结构化等难题，提出了一个Web services框架下的Web挖掘实现方案，使用XML关键技术对异构信息进行包装，使之成为统一的数据模式，从而可以采用数据挖掘技术从海量异构信息中提取出更加有用的信息。

关键词：Web 挖掘；Web Services；XML；包装器；半结构化信息

中图分类号：TP311.12 文献标识码：B 文章编号：1003-7241(2008)05-0073-04

Implementation of the Web Mining Based on Web Services

LI Yan-xia, GONG Jiu-zhou, LI Yu-qing

(College of Mathematics & Information Science, Northwest Normal University, Lanzhou 730070 China)

Abstract: With the development of information technology , Web mining is becoming a hot point in data mining technology. In this paper, an implementation scheme is proposed based on the Web Services framework to solve some difficult problems, such as the different database environment and the half-structured information. In this scheme each of the resources is packaged with XML to form a unified data pattern, then the more significant information can be extracted from the large amount of heterogenous information.

Keyword: Web mining; Web Services; XML; wrapper; half-structured information

1 引言

随着 Internet 的飞速发展，网上的数据资源空前丰富。在这些海量异构的 Web 信息资源中，蕴含着具有巨大潜在价值的知识。人们迫切需要能够从 Web 上快速、有效地发现资源和知识的工具。然而信息检索工具和分析工具的相对落后，导致信息过载。目前人们从 Web 上获取信息的主要途径是通过搜索引擎，搜索引擎虽然部分的解决了资源发现问题，但其精度不够。因此，Web 挖掘技术应运而生，传统的数据挖掘是从大量的数据中发现隐含的规律性内容，解决数据的应用质量问题。充分利用有用数据，废弃无用数据，这是传统数据挖掘的主要应用。所谓 Web 挖掘[1][2]是指从大量的数据集合 C 中发现隐含的模式 p，如果将 C 看作输入，将 p 看作输出，那么 Web 挖掘的过程就是从输入到输出的一个映射 $\xi C \to p$。Web 上的数据最大的特点是半结构化。从数据库的角度看，Web 中的信息可以看作是一个更大的更复杂的数据库；每一个站点就是一个数据源，由于站点之间的信息和组织不同，因而形成一个巨大的异构数据库。如果对这些数据进行挖掘，首先要解决异构数据集成问题和半结构问题。而 Web Services 这些特点：跨平台性和高度集成性、普遍性、 完好的封装性、 松散耦合以及它的接口和封装是可以被 XML 定义描述和发现，并且支持与使用 XML 消息通过网络协议的其他应用软件进行直接交换。能够很好的解决 Web 挖掘上的一些难题。针对 Web 挖掘上的一些难题，鉴于 Web services的优势，我们提出了一种基于Web Services 的 Web 挖掘实现策略。

2 Web Services 框架概述

Web Service[3]是一种开放的分布式应用程序的模型，它能在所有支持 Internet 通讯的操作系统上实现。使用 Web Service 技术可以以独立于平台的方式，通过

* 基金项目：甘肃省科技攻关计划项目（2GS047-A52-002-04）
收稿日期：2007-11-27

标准的Web协议，建立可以由应用程序通过网络访问的应用程序逻辑单元。它的运行机制如下：

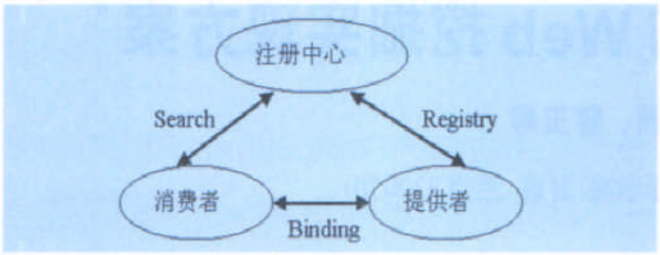

图1 Web Services框架

Web Service的运行机制基于三种角色（服务提供者、服务注册中心和服务请求者）之间的交互。交互涉及注册、查找和绑定操作。这些角色和操作一起作用于Web Service构件:Web Service软件模块及其描述。图1表示了这些操作、提供这些操作的组件及它们之间的交互。这些操作具体为：

注册：为了使服务可访问，需要发布服务描述以使服务请求者可以查找它。发布服务描述的位置可以根据应用程序的要求而变化。

查找：在查找操作中，服务请求者直接检索服务描述或在服务注册中心中查询所要硕士学位论文基于Web Service技术的分布式异构数据库的集成求的服务类型。对于服务请求者，可能会在两个不同的生命周期阶段中牵涉到查找。

绑定：最后，需要解决的问题是如何实现对服务的调用。在绑定操作中，服务请求方通过分析从注册服务器中得到的服务绑定信息，可以知道调用该服务所需的详细要求，包括服务的访问路径、服务调用的参数、返回结果、传输协议、安全要求等，服务请求方根据这些信息对自己的系统进行相应配置，从而实现对服务的远程调用。从而在运行时调用或启动与服务的交互。

3 基于Web Services的异构分布信息的集成与挖掘方案

3.1 整体框架图

目前，Web挖掘所面临的主要问题是数据库环境的异构和信息的半结构化。在本文中，为了解决数据库环境的异构采用了基于Web Services框架进行了信息集成，而信息集成采用的是虚拟方法[4]。包装器将底层的数据对象用XML语言进行包装，将各种性息转换成统一的数据模型。挖掘程序读取了XML，通过XML解析器中的JDOM标准接口，使用它里面的SAXBuilder功能解析出符合JDOM模型的XML树。

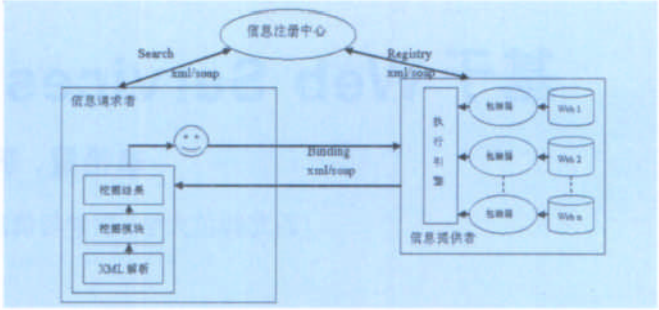

图2 基于Web Services的Web挖掘框架

3.2 Web数据的转化

信息源层处于最低层，是系统的数据提供者。在此包括各种类型，在数据管理上我们采用“虚拟集中”[4]的方式也就是Web数据库实现。每一个异构信息由一个数据源和一个“外套(wrapper)”构成。集成系统是面向各种信息源的，数据类型往往多种多样。由于XML[5][6]具有可扩展性和结构性等特点。因此，用XML模型作为集成系统的公共模型。从数据源中读取数据生成XML统一的文件格式。如下所示：

```
<?xml version="1.0" encoding="gb2312"?>
<root>
    <StudentItem>
        <Id>2006001</Id>
         <Name>Larry</Name>
         <Sex>女</Sex>
    </StudentItem>
     </root>
```

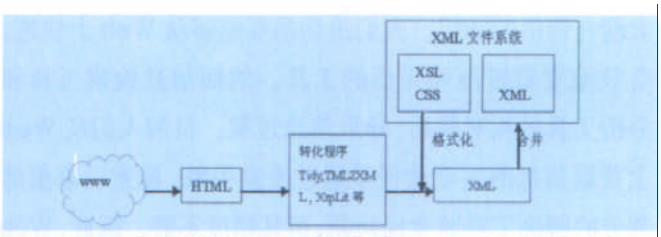

图3 各种格式网页转换为XML储存的过程

3.2.1 包装器的实现过程

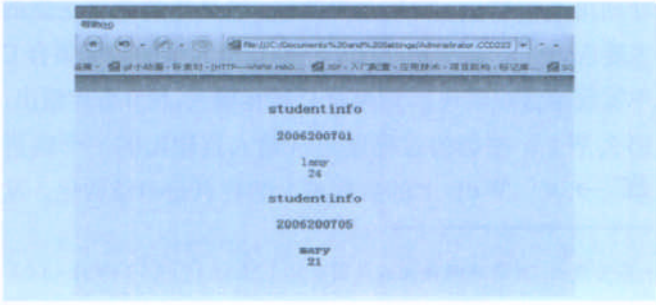

图4 Html的显示结果

目前很多网站都是用HTML语言构建的，所以获取源信息的关键是把现有的Web页面包装成 XML格式的数据，应用XML格式，不仅可以很好地兼容原有的Web应用和信息，而且可以更好地实现Web中的信息共享与交换。XML 可看作一种半结构化的数据模型，可以很容易地将XML的文档描述与关系数据库中的属性一一对应起来，实施精确地查询与模型抽取，以检索出适当的数据。Tidy工具是一个免费使用的产品，可用与改正HTML文档中的常见错误。如图5所示：

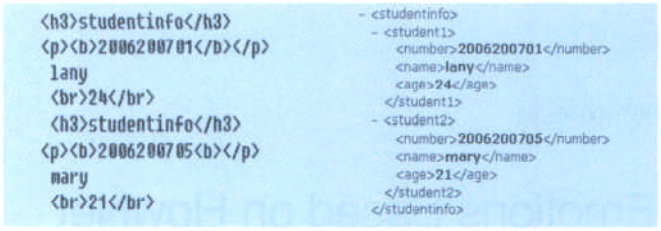

图5 HTML转换为XML的结果

通过构造名为XMLHelper的Java类来完成这一任务以及其它与XML相关任务。通过使用Tidy库提供的函数在XMLHelper TidyHTML ()方法中执行转换。这个方法接受URL作为一个参数并返回一个XML文档作为结果。该过程为抽取页面，转换成XML。

3.3 数据挖掘模快

数据挖掘模块对XML解析器中提供的数据进行挖掘。通过使用挖掘算法库和知识库对JDOM模型的XML树进行挖掘，挖掘出的模式集合经过模式评估和解释得出挖掘结果。如图6所示：

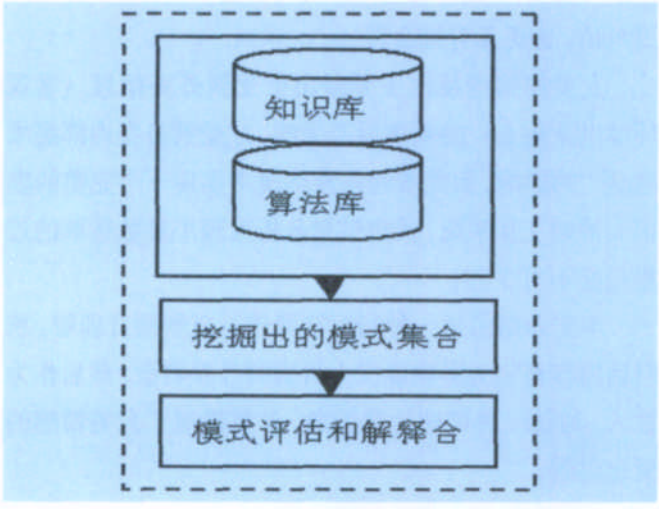

图6 数据挖掘模块

3.4 XML解析器

XML解析器将得到的XML格式的信息文档通过JDOM解析器生成相应的符合JDOM模型的XML树，并根据树形结构进行相应的操作和处理。使用JDOM解析器需将jdom.jar和xerces.jar文件复制到Tomcat网络应用程序的WEB-INF\lib目录中。解析过程中比较重要的几个方法如下：

(1)实例化一个合适的解析器对象：SAXBuilder sb=new SAXBuilder()。

(2)以包含XML数据的文件为参数构建一个文档对象：myDocument Document myDocumet=sb.build(xmlpath);// xmlpath是xml文档的路径。

(3)获得根元素及其子元素：Element rootElement=myDocument.getRootElement();List list=element.getChildern()。

4 基于Web Services挖掘方案的特点

(1)开放性于先进性：Web服务基于开放的的标准(UDDI 、SOAP、HTTP、XML)，将这些技术有机的结合起来，减少投资。

(2)松散耦合：在应用过程中易于修改并且对程序及流程的运行没有任何影响，可以方便地连接异构的平台和系统。

(3)动态集成：接口改变后，应用程序能够方便的重新获取服务描述文档，重新生成调用接口，并于代码进行动态绑定。

(4)简单：基于Web Services的Web挖掘开发和部署都比较简单。

5 结束语

与目前的Web挖掘相比，web services架构提供了很多优势，及开放性、方便、经济、高效。本文就如何应用Web Services 技术构建Web挖掘过程提出了一种实现策略。这种基于Web Services的方式能够满足各种Web挖掘的要求，体现了松散耦合、位置透明、协议独立的特点，能够支持随需应变的业务需求，具有广阔的应用前景。但是挖掘过程中将涉及到海量的数据，如何提高挖掘算法的效率和实时性还需进一步的研究。

参考文献：

[1] 张蓉.Web挖掘技术研究[J].计算机工程,2006-8,32(15):4-5.

(下转第79页)

情绪组。

我们判断这句话为对愤怒情绪的刺激因素，于是我们选择快愤怒情绪的状态变化转移矩阵

$$p_s=\begin{pmatrix}0.06&0.78&0.04&0.07&0.03&0.02\\0.07&0.80&0.03&0.05&0.02&0.03\\0.03&0.82&0.04&0.03&0.05&0.03\\0.04&0.83&0.03&0.02&0.03&0.05\\0.04&0.79&0.05&0.03&0.03&0.06\\0.05&0.84&0.02&0.04&0.02&0.03\end{pmatrix}$$

则情绪变化量 $\Delta e_s=e_s^n p_s=$

$[0.046,0.492,0.022,0.022,0.023,0.020]$

由公式3得 $e_s^{n+1}=e_s^n+\Delta e_s=e_s^n+e_s^n p_s=$

$[0.107,0.553,0.319,0.0805,0.085,0.0815]$ 其中愤怒情绪维度最大，为0.553，那么我们判断此时产生愤怒的情绪。然后再通过中文分词，判断句中有程度副词，取出“有些”，通过查询，判断为低量相对程度副词。

新的情绪状态变为愤怒情绪，快乐的程度为第一等级。这与实际情况是相符的。

继续进行对话，输入“台风‘韦帕’要来了，令人很害怕！”取出句中“害怕”，通过搜索情绪组发现为对恐惧情绪的刺激因素，于是我们选择快恐惧情绪的状态变化转移矩阵

$$p_s=\begin{pmatrix}0.03&0.03&0.05&0.04&0.06&0.79\\0.05&0.02&0.04&0.06&0.02&0.81\\0.03&0.03&0.02&0.01&0.06&0.85\\0.02&0.03&0.01&0.02&0.02&0.90\\0.06&0.02&0.03&0.02&0.04&0.83\\0.06&0.07&0.03&0.04&0.02&0.78\end{pmatrix}$$

则情绪变化量 $\Delta e_s=e_s^n p_s=$

$[0.052,0.055,0.040,0.047,0.116,1.01]$

由公式3得 $e_s^{n+1}=e_s^n+\Delta e_s=e_s^n+e_s^n p_s=$

$[0.159,0.608,0.719,0.1275,0.201,1.09]$

其中恐惧情绪的维度最大，为1.09，因为1.09>1，取值为1，判断此时产生恐惧的情绪。然后再通过中文分词，判断句中有程度副词，取出“很”，通过查询，判断为高量程度程度副词。新的情绪状态变为恐惧情绪，快乐的程度为第三等级。这与实际情况是相符的。

4 结束语

情感建模是一项涉及心理学、生理学、数学、计算机科学以及信息科学等领域的复杂工作。情感计算作为自然和谐人机交互的关键技术已经在应用方面取得了许多进展。但是，由于情绪心理学理论方法的多样性与争议性，导致情感建模的理论与方法都不成熟，很多理论与方法问题亟待解决[8]。

本文尝试建立了一种新的情感模型，首先利用知网进行自然语言处理，然后通过情感模型来模拟人的情感变化。经过验证，在一定程度上符合人类情绪的变化过程，具有良好的人机情感交互能力。

参考文献:

[1] 赵积春,王志良,王超.情绪建模与情感虚拟人研究[J].计算机工程.2007,33,(1):212-215

[2] 张冬蕾.情绪研究在计算机科学领域的发展与应用[J].信息技术,2005,3,(5):33-40

[3] MINSKY M.The Society of Mind[M].New York:Simon and Schuster,1986.

[4] 董振东,董强.知网http://www.keenage.com

[5] 刘群,李素建.基于《知网》的词汇语义相似度计算[C].第三次汉语词汇语义研讨会,台北.2002

[6] 王志良,乔向杰,王超等.基于自定义空间和OCC模型的情绪建模研究[J].计算机工程.2007,33,(4):189-192.

[7] 王力.中国现代语法[M].北京:商务印书馆,1985.131-132

[8] 杨国亮,王志良.情感建模研究进展[J].自动化技术与应用,2004,23(11):1-4

作者简介:王琦(1983-),女,硕士研究生,研究方向:人工智能。

(上接第75页)

[2] YUEFENG Li,NING ZHONG.Web mining model and its applications for information gathering[J].Knowledge-Based Systems.2004,(4):207-208.

[3] J.Mark Pullen,Ryan Brunton,etc.Using Web services to integrate heterogeneous simulations in a grid environment[J].Future Generation Computer Systems.2004.(9):98-99.

[4] 杨先娣,彭智勇等.信息集成研究综述[J].计算机科学,2006,33(7):55-56

[5] CHANGTAO QU,WOLFGANG NEJDL.Integrating XQuery-enabled SCORM XML Metadata Repositories into an RDF-based E-Learning P2P Network[J].Educational Technology & Society,2004,7(2):51-60

[6] 杨建物,陈晓鸥.XML相关标准综述[J].计算机科学,2002,29(2):25-27.

作者简介:李艳霞（1981-),女,硕士研究生,研究方向:网格技术,Web计算。