# 日志挖掘中的数据预处理

方 杰,朱京红
(合肥工业大学 计算机与信息学院,安徽 合肥 230021)

**摘 要**:日志信息的预处理是日志挖掘任务中的重要阶段,是当前研究的重点,同时也是整个日志挖掘过程的基础和实施有效挖掘算法的前提,在日志挖掘中起着重要的作用。目前主要的日志挖掘主要采用国外的几种软件,而日志挖掘中重要的数据预处理软件国内暂无。文中主要介绍了数据挖掘中的日志挖掘,分析了数据预处理的过程,以及如何实现日志挖掘中的数据预处理,并在Delphi开发工具中成功完成了IIS文本日志文件到Xls格式及XML格式文件的转换,实现了日志挖掘中的数据预处理。

**关键词**:XML;日志挖掘;数据预处理
**中图分类号**:TP311;TP39 **文献标识码**:A **文章编号**:1673-629X(2010)04-0017-04

# Data Pretreatment of Log Mining

FANG Jie, ZHU Jing-hong
(School of Computer & Information, Hefei University of Technology, Hefei 230021, China)

**Abstract**: Log information preprocessing is an important stage of the log mining task, which is the focus of current research. It is also the whole basis of log mining process and the implementation of the prerequisite of an effective mining algorithm. In the log mining, it plays an important role. Log mining is currently the main tool is the number of foreign software, and logs important data mining software internally no pretreatment. This paper introduces data mining in the log mining, a detailed analysis of the data pre-processing process, as well as how to log the data pre-processing mining, and Delphi development tools in the successful completion of the IIS log file to text format and XML format Xls document conversion, to achieve the log mining in data preprocessing.

**Key words**: XML; log mining; data pretreatment

## 0 引 言

随着Internet的发展应用,WWW上的信息量剧增,其中包含了大量的数据信息。如何从访问的Web日志数据中快速地抽取用户感兴趣的访问模式,通过对服务器日志的分析和挖掘获取用户访问路径及关注点,以便优化站点结构,为用户提供个性化Web服务,提高用户查找信息的质量和效率和进行个性化服务等,这就是目前Web日志挖掘的重点研究方向[1]。

当前Web日志挖掘一般包含以下三个阶段:数据预处理阶段、模式发现、模式分析。

而数据预处理是日志挖掘中最重要阶段,是后续数据挖掘和分析能否顺利进行的前提和关键。数据预处理是为了将日志文件转换成数据库文件而进行的工作,其目的是把Web日志数据转换为适合进行数据挖掘的精确数据[2]。

**收稿日期**:2009-07-09;**修回日期**:2009-10-18
**基金项目**:国家自然科学基金(60705015)
**作者简介**:方 杰(1980-),男,安徽霍邱人,硕士研究生,合肥师范学院现代教育技术中心讲师,从事人工智能、数据挖掘研究。

## 1 日志挖掘中的数据预处理

Web用户访问Web服务器时,Web服务器会自动创建访问日志信息,包括访问日志、引用日志、代理日志、错误日志等文件。文件里包含了大量的用户访问信息,如所访问用户的IP地址、访问日期和时间、访问方法(或)、访问结果、URL GET POST访问的信息大小等。以微软的IIS产生的访问日志文件为例,其日志文件包含数据形式为:"2009-3-2 08:26:25 127.0.0.1 GET /vv/10-01.xml 200",其中关键字段以空格分割,可以看出日志文件包含的信息只是普通的文本形式,并非符合关系型数据库的结构模型,而目前的数据挖掘一般是建立在关系型数据库基础上的,因此为了实现挖掘,首先就需要对日志文件的预处理,将日志文件转换为可以方便挖掘的数据库文件。日志源文件如图1所示。

数据预处理对于数据挖掘非常重要,在日志中存在许多对于数据挖掘无用的属性和数据,而对数据挖

掘算法而言，由于数据挖掘算法通常只能处理固定格式的数据，不正确的输入数据可能导致错误或者不准确的挖掘结果[3]。在当前的Web使用挖掘中对于数据源有两种处理方法：一种是将数据源数据预处理后转换为传统的关系型数据库；另外一种方法直接对Web记录日志数据预处理进行挖掘。

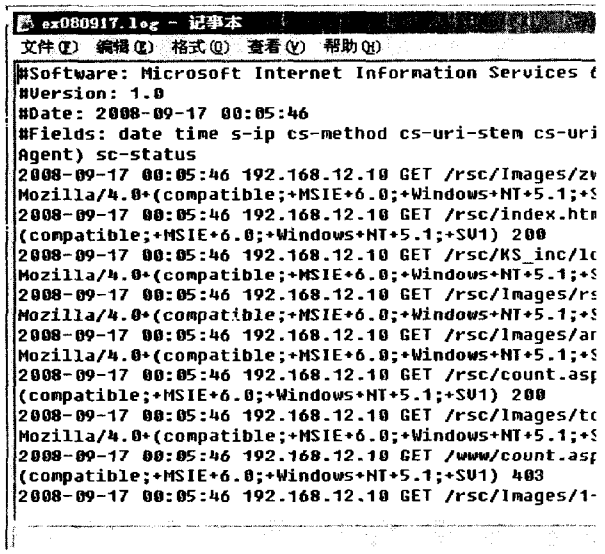



图1 日志源文件

目前，日志挖掘技术发展迅速，作为整个挖掘过程Web的基础和实施有效挖掘算法的前提，数据预处理的目的就是将原始日Web志记录经过处理形成用户的会话文件，为挖掘算法实施阶段作好数据准备。当前数据预处理一般包括以下四个阶段：数据收集及数据清洗、用户识别、会话识别和路径补充[4]。

### 1.1 数据收集及数据清洗

数据收集可以分为服务器端数据收集、客户端数据收集、代理服务器端数据收集。

(1)服务器端数据收集。

在服务器日志中记录了用户每次访问网站进行的每一次网页请求的信息，全面地记录用户登录页面的详细信息，比如：时间、日期、IP地址、访问的页面等等，并可通过记录Cookies和CGI的查询参数来描述各个不同用户的行为。使用Web服务器访问日志来实现数据采集是行之有效的，能方便地分析出用户的浏览行为，因而基于Web服务器日志的数据挖掘是目前日志挖掘的研究重点。

(2)客户端数据收集。

“客户端数据收集技术”(Client－side Data Collection)或“数据短标签”(data tagging for short)是实现客户端数据收集的新方法。它在网页中嵌入标签，并收集这些标签数据的脚本代码(如JavaScript)。当网页被客户端加载时，这些代码被客户端浏览器执行，代码将把标签信息和一些客户端信息发送到服务器，以做分析的基础。数据标签的采用可以解决由于分析Web服务器日志(Web Server Log)文件所产生的诸多问题。

(3)代理服务器端数据收集。

对于局域网来说大多用户都是通过代理服务器登录网站的，所以通过代理服务器不仅可以收集多个用户的行为，还可以收集对多个网站的行为。利用收集到的信息，系统管理员可以获取有价值的数据，从而有助于完善网络管理，有助于实现使用挖掘。

数据清洗是指根据需求对日志文件进行处理，包括删除无关紧要的数据、合并某些记录、对用户请求页面时发生错误的记录进行适当的处理等等。日志挖掘的目的是发现用户的浏览行为模式，而用户的行为是指用户点击超链接的动作。在页面中有时包含图像、声音、动画以及广告等文件，同时也需要将日志中文件后缀名为JPG、GIF、SWF、CSS、JS等请求项删除，通常数据清洗包含两个步骤：属性删减和记录删减。

### 1.2 用户识别

在对用户进行访问模式挖掘或用户聚类分析时，用户识别显得至关重要，因为群体是由个体组成的，只有对个体有了清楚的了解，才能识别群体的特征。然而由于本地缓存、代理服务器(网吧、局域网等环境)和防火墙的使用，使得用户识别这一步变得很复杂。

用户识别，是从日志文件中的每一条记录中识别出响应的用户。一般通过三条规则，结合用户提交的查询信息便可以给不同的用户赋予不同的用户ID号。

规则如下：

(1)如果用户的IP地址不同，则认为是不同的用户；

(2)如果IP地址相同，而代理agent日志中表明用户的浏览器或操作系统改变了，则可以假设为两个不同的用户；

(3)将访问日志、引用日志和站点拓扑结构相结合构造用户的浏览路径。如果当前请求的页面同用户已浏览的页面间没有链接关系，则认为存在IP地址相同的多个用户。

### 1.3 会话识别

会话识别就是将用户在一段时间内所有的请求页面分解得到能反映实际浏览习惯的用户会话，会话Session是指用户在一次访问网站期间从进入网站到离开网站所进行的一系列活动。

定义1 用户会话S是一个二元组＜userid，RS＞，其中userid是用户标识，RS是用户在一段时间内请求的Web页面的集合。

RS包含了用户请求的页面的标识符Pid和请求

的时间 t,用户会话 S 可以表示为如下的元组:

$$S = < \text{userid}, \{(\text{Pid}_1, t_1), \cdots, (\text{Pid}_k, t_k)\} > \quad (1)$$

其中 $t_k$ 表示 $\text{Pid}_k$ 页面的访问时间。通过用户识别,可以得到关于用户的访问 Web 页面的一个序列,然后可以通过设定超时的方法识别用户会话。(1) 式中的用户会话要满足下面的条件(其中 $T$ 为设定的超时界限值):

$$t_i - t_{i-1} \leqslant T, 1 < i < K \quad (2)$$

具体算法略。

### 1.4 路径补充

路径补充就是将由于本地或代理服务器缓存而没有被日志记录的请求页增加到用户会话中。也可以根据引用日志和网络拓扑结构提供的信息把路径补充完整。解决的方法类似于用户识别中的方法[5]。

## 2 XML 概述及在 Web 日志挖掘中的优势

### 2.1 XML 概述

XML(eXtensible Markup Language, 可扩展标记语言)是由 W3C 于 1998 年 2 月发布的一种标准,具有简单、开放、高效可扩充、国际化等特点。随着计算机和 Web 技术的发展, XML 技术在异构系统及数据库间实现信息的高效传输、交换、共享,并在保证数据的完整性上发挥着越来越重要的作用。XML 数据由嵌套和标记元素组成,标记包含对文档存储形式和逻辑结构的描述, 这种结构使 XML 很适合描述 Web 上的半结构化数据, 甚至非结构化数据。这使得"以关系数据库为存贮手段,以 XML 为交换载体"的数据管理模式成为可能[6]。

### 2.2 XML 特点

首先 XML 解决了 HTML 不能解决的两个 Web 问题,即 Internet 发展速度快而接入速度慢的问题,以及可利用的信息多,但难以找到自己需要的那部分信息的问题。在 XML 中显示样式从数据文档中分离并且是纯文本文件, 具有跨平台及与软件、硬件无关的特性。因此, XML 能够作为一种广义的中介在异构应用间进行数据交换。使得我们对于 Web 这样一个复杂的异构数据库环境中使用 Web 日志挖掘成为可能。

对 Web 服务器日志数据进行预处理形成 Web 日志数据库。通过转换再将其转换成 XML 的网页,并对网页信息库进行存储。利用数据挖掘模型对 XML 数据进行挖掘,数据挖掘模型是系统功能实现的核心,提供各种数据挖掘算法和有效的解决方案以及结合用户的兴趣模型,最终为用户挖掘到所需要的信息,有效完成日志挖掘任务[7]。

## 3 Delphi7.0 环境下实现 WEB 日志到 XLS 以及 XML 文档等格式文件转换

Delphi[8]中实现日志文件的读取:

首先需要对日志文件预处理,将非有效的信息去除,通过 Topendailog 选定日志文件并将信息读取到 Tstringgrid 中,算法伪码如下:

```
var
    VLogfile,VProfile:TstringList; //文件读入此 StringList 中
    LinStr:Tstrings; i,k:integer; Row:array[1..40]of Tstrings;
  begin
    If opendialog1.Execute then //用于提取日志文本文件
    begin
    VLogfile:=TstringList.Create;LinStr:=TstringList.Create;
    VLogfile.LoadFromFile(opendialog1.Filename); //日志文件读取到字符串集中
    end;
    for i:=0 to VLogfile.Count-1 do
    begin
    LinStr.Clear;
    ExtractStrings([';',',','.','|',' '],['@',' '],PChar(VLogfile.Strings[i]),LinStr);
    if stringgrid1.RowCount<i+1 then //添加行
    stringgrid1.RowCount:=i+1; stringgrid1.Rows[i]:=LinStr; //整行赋值
    end;
    end;
```

具体程序的其它算法和功能在文中不再说明。具体程序如图 2 所示,预处理导出的 Xls 文件如图 3 所示,预处理导出的 XML 文件如图 4 所示。

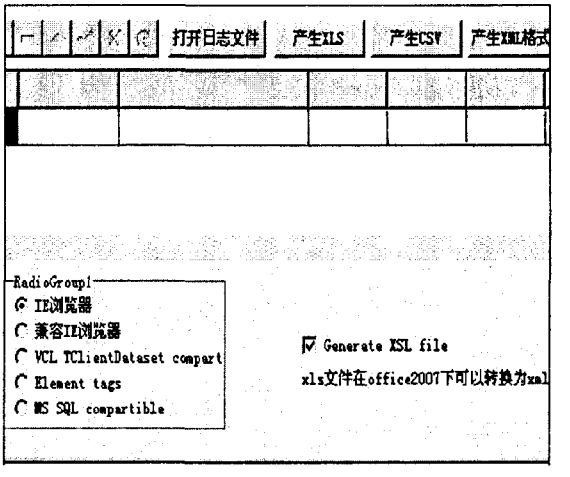
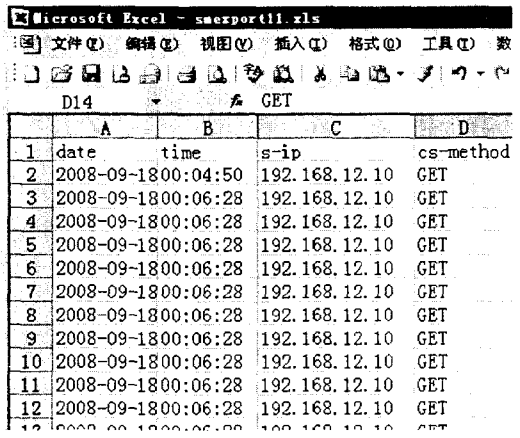



图 2　程序运行图

## 4 结束语

Web 日志数据准确性是 Web 日志挖掘中重要的前提和基础,只有准确的数据才能正确地反映用户的意图,才能保证分析沿着正确的方向进行。文中综述

了日志挖掘中数据预处理的基本概念和基本过程，并用开发工具制作了专门软件来完成日志文本信息到关系型数据格式的转换。但由于 Web 日志的多样性和不确定性，还有很多问题亟待解决，有待于进一步去研究和探索。


Microsoft Excel - smexport11.xls

文件(F) 编辑(E) 视图(V) 插入(I) 格式(O) 工具(T) 数

D14 GET

| | A | B | C | D |
|---|---|---|---|---|
| 1 | date | time | s-ip | cs-method |
| 2 | 2008-09-18 | 00:04:50 | 192.168.12.10 | GET |
| 3 | 2008-09-18 | 00:06:28 | 192.168.12.10 | GET |
| 4 | 2008-09-18 | 00:06:28 | 192.168.12.10 | GET |
| 5 | 2008-09-18 | 00:06:28 | 192.168.12.10 | GET |
| 6 | 2008-09-18 | 00:06:28 | 192.168.12.10 | GET |
| 7 | 2008-09-18 | 00:06:28 | 192.168.12.10 | GET |
| 8 | 2008-09-18 | 00:06:28 | 192.168.12.10 | GET |
| 9 | 2008-09-18 | 00:06:28 | 192.168.12.10 | GET |
| 10 | 2008-09-18 | 00:06:28 | 192.168.12.10 | GET |
| 11 | 2008-09-18 | 00:06:28 | 192.168.12.10 | GET |
| 12 | 2008-09-18 | 00:06:28 | 192.168.12.10 | GET |

图 3 预处理产生的 Xls 文件

```
<?xml version="1.0" encoding="UTF-8" ?>
- <dataroot xmlns:od="urn:schemas-micros
  xsi:noNamespaceSchemaLocation="smex
  - <smexport>
    <ID>1</ID>
    <data>2008-09-18</data>
    <time>00:04:50</time>
    <sip>192.168.12.10</sip>
    <method>GET</method>
    <cs-uri-query>/rsc/zsbbook/leavew
    <cs-username>8001</cs-username>
    <cip>74.6.22.183</cip>
    <ie>Mozilla/5.0+(compatible</ie>
  </smexport>
  - <smexport>
```

图 4 预处理产生的 XML 文件

注：本程序在 http://www.hftc.edu.cn/xjzx/datamining.rar 有下载。

**参考文献：**


[1] Cooley R, Srivastava J. Grouping Web page references into transactions for mining world wide Web browsing patterns [C]//Proceedings of KDEX'97. Newport Beach, CAUSA: [s.n.], 1997:2-7.

[2] Wong J S K, Nayar R. A framework for a world wide web based data mining system[J]. Journal at Network and Computer Applications, 2000, 21:163-185.

[3] Pei J, Han J. Mining access patterns efficiently from Web logs [C]//Sun Liping, Zhang Xiuzhen. PAKDD' 00, Kyoto, Japan2000. Efficient Frequent Pattern Mining on Web Logs. APWeb 2004. [s.l.]: [s.n.], 2004:533-542.

[4] Ezeife, Lu Yi. Mining Web Log Sequential Patterns with Position Coded Pre - Order Linked WAP - Tree[J]. Data Mining and Knowledge Discovery, 2005(10):5-38.

[5] 马瑞民，李向云. Web 日志挖掘中数据预处理技术的研究[J]. 计算机工程与设计，2007(10):2358-2359.

[6] 刘造新. 基于本体的 XML 关联规则挖掘方法[J]. 计算机应用，2008(9):2319-2320.

[7] 李雪竹. 一种基于 XML 的 Web 数据抽取的实现[J]. 科学技术与工程，2008(9):2473-2474.

[8] Delphi[EB/OL]. 2009-03-21. http://baike.baidu.com/view/3297.htm. baidu, Linking-2009-03-21.


(上接第 16 页)


Computer and System Sciences, 1993, 46(1):39-59.

[40] Quafatou M. a-RST: a generalization of rough set theory [J]. Information Sciences, 2000, 124(1-4):301-306.

[41] Beyond M. Reducts within the variable precision rough sets model :a further investigation[J]. European Journal of Operational Research, 2001, 134(3):592-605.

[42] 汪小燕，杨思春. 一种新的不一致决策表属性约简算法[J]. 计算机应用，2008，28(2):525-527.

[43] 陈鑫影，邱占芝. 不协调决策信息系统的约简[J]. 计算机工程与应用，2008，44(7):193-195.

[44] 贺 鹏，王庆林. 可重构制造系统故障诊断多 Agent 自学习模型[J]. 计算机工程与设计，2007，28(8):1741-1743.

[45] 朱永利，吴立增，李雪玉. 贝叶斯分类器与粗糙集相结合的变压器综合故障诊断[J]. 中国电机工程学报，2005，25(10):159-165.

[46] 张秋娜，董双勤. 粗糙集模型和概率粗糙集模型的若干研究[J]. 重庆文理学院学报：自然科学版，2007，26(4):13-14.

[47] 刘高峰，王 飞. 基于聚类分析的粗糙集模型及其应用[J]. 内江师范学院学报，2008，23(8):28-31.

[48] 印 勇，孙如英. 基于聚类有效性分析的模糊粗糙集归纳学习方法[J]. 计算机工程，2008，34(10):86-88.

[49] 宋云雪，张传超，史永胜. 基于模糊粗糙集的飞机远程故障诊断模型研究[J]. 中国民航大学学报，2007，25(6):15-19.

[50] 庄曰平，李 伟. 基于粗糙集和模糊理论的变电站电压无功控制策略[J]. 武汉大学学报：工学版，2007，40(5):112-115.

[51] 张 明，方 敏. 基于粗糙集和小波矩的车牌字符识别[J]. 安徽建筑工业学院学报：自然科学版，2007，15(3):95-98.

[52] 车志宇，夏明革，何 友. 基于粗糙集与小波分析的图像融合算法[J]. 电光与控制，2005，12(1):18-21.

[53] 郑小霞，钱 锋. 基于粗糙决策模型和蚁群算法的故障诊断[J]. 系统工程理论与实践，2007(3):140-144.

[54] 马 昕，林丽清. 蚁群算法在面向属性的数据约简中的应用[J]. 计算机仿真，2007，24(9):158-160.