# Web使用挖掘的数据预处理*)

刘立军 周 军 梅红岩
(辽宁工学院计算机科学与工程学院 辽宁 锦州 121001)

**摘 要** Web使用挖掘的基本思想是将数据挖掘技术应用于Web使用数据源。在数据挖掘研究领域中,数据预处理起着至关重要的作用。Web使用挖掘的数据源最主要的是Web日志,介绍了Web日志的具体内容,针对Web日志的特点,介绍预处理过程中一些特殊情况的处理方法,并在事务的识别阶段给出了一种新的最大向前引用序列挖掘算法——剪枝算法。

**关键词** Web使用挖掘,数据预处理,剪枝算法,最大向前引用,事务识别

## The Pre-processing of Web Usage Mining

LIU Li-Jun ZHOU Jun MEI Hong-Yan
(Department of Computer Science, Liaoning Institute of Technology, Jinzhou121001)

**Abstract** The basic idea of Web usage mining is to apply the technology of data mining to the data source of Web usage. Data preprocessing plays an important role in the field of Web usage mining. The data source of Web usage mining is mainly composed of Web logs. This paper introduces detailed content of Web logs. In the light of features of Web logs, some special processing methods of preprocessing are introduced. At the same time, a newly algorithm of maximal forward references-pruning algorithm is proposed at the stage of transaction recognition.

**Keywords** Web usage mining, Preprocessing, Pruning algorithm, Maximal forward references, Transaction recognition

## 1 引言

随着以数据库、数据仓库等数据仓储技术为基础的信息系统在各行各业的应用,海量数据不断产生。如何从大量的数据中找到真正有用的信息成为人们关注的焦点,数据挖掘技术应运而生。数据挖掘(Data Mining)[1]就是从大量的、不完全的、有噪声的、模糊的、随机的实际应用数据中,提取隐含在其中的、人们事先不知道的、但又是潜在有用的信息和知识的非平凡过程。Web挖掘是数据挖掘的一种,是指使用数据挖掘技术在WWW数据中发现潜在的、有用的模式或信息。Web挖掘研究覆盖了多个研究领域,包括数据库技术、信息获取技术、统计学、人工智能中的机器学习和神经网络等。一般地,Web挖掘可分为3类:Web内容挖掘(Web content mining)、Web结构挖掘(Web structure mining)和Web使用记录的挖掘(Web usage mining)。Web挖掘的一般过程为:数据预处理、模式发现、模式分析。数据预处理是把从各种数据源得到的使用信息、内容信息和结构信息转换成模式发现阶段需要的数据抽象。1999年Pyle[2]提出在数据挖掘过程中增加数据预处理过程,Pyle强调了数据预处理的重要性:数据预处理过程在数据挖掘过程中占据了60%的时间。Cooley在文[3]提出Web日志挖掘的数据预处理的关键任务在于如何修复错误数据和处理缺失数据。Tan and Kumar[4]意识到Web日志挖掘中日志的冗余性,清理无关据:robots(蜘蛛)和其它软件代理的请求也应该成为日志预处理的任务。近几年,很多研究重点都放在Web日志的挖掘,但是对Web日志掘中的数据预处理方面却没有引起足够的重视。

文章针对Web使用挖掘数据源的特点,研究工作主要集中于在Web日志挖掘过程中如何对日志数据进行预处理,在不影响挖掘结果的基础上来提高日志挖掘过程的效率。在第2部分给出了一个完整的预处理过程,并在事务的识别阶段给出了一种新的最大向前引用序列挖掘算法--剪枝算法。

## 2 Web使用挖掘的预处理过程

Web usage mining即Web使用模式挖掘,主要是挖掘网站访问日志,从中挖掘用户访问模式[5]。目前市面上比较流行的Web服务器,例如IIS、Apache等,通常都保存了对Web页面的每一次访问的日志项,这些记录项又叫做Web log项;它忠实地记录了访问该Web服务器的数据流的信息。日志文件的格式并不复杂,例如IIS,它支持三种Web日志格式:Microsoft IIS日志文件格式,NCSA公用日志文件格式和W3C扩展日志文件格式,其中在IIS5.0中,W3C扩展日志文件格式是缺省的日志文件格式。日志文件记录些什么内容,还可以根据客户的不同需要,来调整记录些什么信息。例如IIS5.0中W3C扩展日志文件格式中,除了时间这些日志文件肯定有的元素外,还有多达19项可以选择记录的扩展属性,比较常用的属性是所请求的URI资源,客户端IP地址和时间戳。

Web日志一般符合W3C Working Draft推荐的CLF(Common Log File Format)和ECLF(Extended Log File Format)标准,其常用的字段和含义如下所示。

日期(date):用户请求页面的日期;时间(time):用户请求页面的具体时间;客户IP地址(ip):客户端主机的IP地址或DNS;客户名(username):客户端的用户名;用户代理(User-Agent):服务的提供者;服务器ip地址(ip):服务器的IP地

*)基金项目:辽宁省优秀青年骨干教师基金资助项目;辽宁省教育厅基金资助项目(20031066)。**刘立军** 硕士研究生,主要研究方向:数据挖掘;**周 军** 教授,硕士生导师,主要研究方向:数据挖掘、人工智能;**梅红岩** 硕士研究生,讲师,主要研究方向:数据挖掘、人工智能。

址；服务器端口(s-port)：服务器的端口号；方法(method)：用户的请求方法；URL 资源(Uri-stem)：用户的请求页面；URL 查询(Uri-query)：用户欲进行的查询；协议状态(status)：返回 HTTP 的状态标识；服务器名(computer name)：服务器名称；发送字节数(sc-bytes)：服务器发送的字节数；接收字节数(cs-bytes)：服务器收到的字节数；所花时间(time-taken)：完成浏览所花费的时间；协议版本(version)：传输用的协议版本；主机(host)：服务器的操作系统；Cookie(Cookie)：Cookie 标识符；参照(Referer)：用户浏览的上页。

在 W3C 扩展日志文件格式中，缺省的属性有：时间戳，客户端 IP 地址，访问方法，URI 资源，协议状态。典型的日志记录形式如图 1 所示。

| IP Address | User ID | Time | Method/URI/Protocol | Stauts | Size |
|---|---|---|---|---|---|
| 200.100.89.2 | -- | 10/Dec/2003:12:34:16-0600 | "GET/images/gaat.gif HTTP/1.1" | 200 | 44851 |
| 203.102.87.5 | -- | 10/Dec/2003:12:34:32-0600 | "GET/graduate.htm HTTP/1.1" | 200 | 7403 |
| 203.101.82.5 | -- | 10/Dec/2003:12:34:32-0600 | "GET/images/haha.jpg HTTP/1.1" | 200 | 18481 |
| 203.141.86.9 | -- | 10/Dec/2003:12:34:48-0600 | "GET/result.htm HTTP/1.0" | 20 | 12302 |
| 200.137.2.52 | -- | 10/Dec/2003:12:34:58-0600 | "GET/structure.htm HTTP/1.1" | 200 | 367 |
| 205.128.5.58 | -- | 10/Dec/2003:12:34:58-0600 | "GET/abcindex.htm HTTP/1.1" | 200 | 4370 |
| 208.153.99.78 | -- | 10/Dec/2003:12:34:58-0600 | "GET/abccontent.htm HTTP/1.1" | 200 | 12047 |
| 206.160.55.88 | -- | 10/Dec/2003:12:34:58-0600 | "GET/images/gty.jpg HTTP/1.1" | 200 | 22574 |

图 1　典型的日志记录形式

上面清单中前面部分是 CLF 的日志格式，后面斜体部分是 ECLF 增加的记录项。其中一些内容在实际应用中是用不到的，如 URI 查询。下面是一个实际的日志：

210.34.48.221- -[25/Sep/2005:22:10:10 -800]"Get bug.html HTTP/1.1"200 4219/Index.html Mozilla(IE6.0 win98).

上面的日志说明的是一个 IP 地址为 210.34.48.221 的客户端在 2005 年 9 月 25 号晚上 22:10:10 这一时刻使用 IE6.0 的浏览器在 index.html 页面上发出一个 HTTP 的 Get 请求，这个请求的目的是 bug.html。

数据预处理是一个十分关键的步骤，根据不同的业务，不同的情况，将海量的原始数据中抽取需要的数据，并且对于不完整的数据还需要做些处理等。Web 使用挖掘的数据预处理包括依赖域的数据净化、用户识别、会话识别和路径补充、事务识别等。对日志进行预处理的结果直接影响到挖掘算法产生的规则与模式。因此，预处理过程是保证 Web 使用挖掘质量的关键。

### 2.1　清理数据

当用户请求一个网页时，与这个网页有关的图片、音频等信息会自动下载，并记录在日志文件中；而如果我们挖掘的目的是用户访问模式，这些信息对我们来说显然用处不大(除非图片、音频等是用户显示请求的，即用户所需要的内容正是这些图片和音频等文件)，所以可以把日志中文件的后缀为 gif、jpg、jpeg、GIF、JPG、JPEG 等的记录删除。但是，当挖掘的目的是为了进行网络流量分析或为页面缓冲与预取提供依据时，这些信息又会显得格外重要，所以在删除这些记录的时候一定要把相关信息记录下来。选择将其中的"发送字节数"和"接收字节数"这两个域的内容记录下来。此外，后缀名为 cgi、js 和 JS的脚本文件因对后面的分析处理不造成任何影响，所以应该删除。可以定义一个缺省的规则库来帮助删除记录，而且这个规则库可以根据正在分析的网站类型进行修改。例如，对于主要包含图形的站点，日志中的图形文件可能代表了用户的显式请求，此时就不能将图形文件删除。可以预先将网站分为一般网站、图片网站、音视频网站等，分别建立对应的规则库；确定要分析的网站属于哪一类然后按照该类网站的规则库进行数据清理。当然，还可以根据需要对规则库进行删改。

### 2.2　用户的识别

由于本地缓存、代理服务器和防火墙的存在，使得有效识别用户的任务变得十分复杂。一般采用的方法是基于日志站点的方法，还可以使用一些启发性规则。例如：如果 IP 地址相同，但是代理信息变了(代理信息，在 IIS5.0 环境下的 W3C 扩展日志文件格式的 cs(User-Agent)字段)，表明用户可能是在某个防火墙后面的内网的不同用户，则可以标记为不同的用户；还可以将访问信息，引用信息 cs(Referer)字段和站点拓扑机构结合，构造出用户的浏览路径，如果当前请求的页面同用户已浏览的页面没有链接关系，则认为存在 IP 地址相同的多个用户。使用这些规则难以保证准确识别用户，因此用户识别是个难题。

### 2.3　会话识别

在跨越时间区段比较大的 Web 服务器日志中，用户可能多次访问该站点，会话识别的目的就是将用户的访问记录分为单个会话。最简单的方法是用超时的技术，如果两个页面之间请求的时间差值超过了一定界限就认为用户开始了一个新的会话。例如，可以设置 30min 等。

### 2.4　路径补充

在识别用户会话过程中的另外一个问题是确定访问日志中是否有重要的请求没有被记录。这就需要路径补充来完成这些记录了。如果当前请求的页面与用户上一次请求的页面之间没有超文本链接，那么用户很可能使用了浏览器上"BACK"的功能调用缓存在本机中的页面。检查引用信息确定当前请求来自哪一页，如果在用户的历史访问记录上有多个页面都包含与当前请求页面的链接，则将请求时间最接近的作为当前请求的来源，如果引用信息不完整，则可以利用站点的拓扑结构来代替。

### 2.5　事务的识别

在 Web 日志挖掘领域中，用户会话是惟一具备自然事务特征的对象，但它对于关联规则挖掘任务而言粒度仍较大，需要特定的算法将用户会话分割为更小的事务。划分事务的主要方法是引用时长和最大前向引用。

(1)引用时长

Web 页面可以简单分为两类：内容页和导航页[6]。当页面中超链接达到一定数目时，可看成导航页，这是一种静态划

(下转第 204 页)

种动态文本具有较好的效果。该方法可用于对聊天过程中的聊天文本进行实时特征选取，也可广泛应用于对短文本进行文本挖掘时的特征选取。但是，随着聊天知识库的增大，如果某新文本为一新主题，遍历聊天知识库查询近似主题将可能非常耗时，并且每次文本特征选取完毕后，对聊天知识库进行修正或改进时同样面临这个问题。不过可将聊天知识库中的所有数据存储在数据库服务器中，通过改进查询算法、使用存储过程等数据库技术来缓解这个问题。

## 参 考 文 献


1 Han Jiawei, Kamber M. Data Mining Concepts and Techniques [M]. 北京：高等教育出版社，2001. 285～295

2 王继成，潘金贵，张福炎. Web 文本挖掘技术研究[J]. 计算机研究与发展，2000(5)：513～520

3 Li Fan, Lu Mingyu, Lu Yuchang. Research about newmethods of text feature extraction[J]. Journal of Tsinghua University (Science and Technology), 2001 (7)：98～101

4 Luo Jie, Chen Li, Xia Delin, et al. Research on Fast Text Classifier Based on New Keywords Extraction Method[J]. Application Research of Computers, 2006(4)：32～34

5 Kantrowitz M, Mohit B, Mittal W. Stemming and its Effects on TFIDF Ranking[J]. In: Proc. of the 23rd annual international ACM SIGIR Conf on Research and development in information retrieval. New York: ACM Press, 2000. 357～359

6 Colliez J, Dufrenois F, Hamad D. Optic Flow Estimation by Support Vector Regression[J]. Engineering Applications of Artificial Intelligence, 2006, 19(7)：761～768

7 http://bbs.ctbu.edu.cn


(上接第 201 页)

分法。内容页一般是用户关心的信息，浏览时间较长。导航页是用户快速找到所需信息而设置的路标，浏览时间较短。通过辅助页面在整个日志中所占的比例的估计，可以使用一个最大可能估计算法来划分辅助页面和内容页面的划分时间。通过对照划分时间，页面就可以划分为内容页或导航页，从而划分为不同的事务。

(2)最大前向引用

有时某些页面含有较多超链接，是用户关心的信息，却被作为内容页，此时可采用 Chen[7] 等人提出的最大向前引用路径(简称为 MFP) 来定义事务。对于每个用户会话，从开始页面为起点，每个最大前向引用路径即为一个事务。这里每一个事务定义为一组页面的访问，从第一次引用，直到在某处向后回溯为止。前向指引定义为一个从未在事务集合中出现的页面，后向指引指已经在前面的事务中出现的一个页面。当一个前向指引出现时，开始一个新的事务。获取最大向前引用序列的算法如下：

剪枝算法描述

Step1：初始化，定义存储最大向前引用序列的字符串变量和循环控制变量。

Step2：扫描数据库，执行连接和剪枝操作操作，输出最大向前引用序列。

Step3：重复执行 Step2，直到没有新的最大向前引用序列为止。

剪枝算法：

输入：大项集阶段转换后的序列数据库

输出：最大序列

Step1：初始化：令 $i=0$，字符串 $S=$ NULL//字符串 $T$ 用来存储当前的向前引用。

Step2：While($s_{i+1}$不空且 $s_{i+1}$比 $s_i$ 更深){连接 $s_i$ 到 $S$；$i=i+1$；}连接 $s_i$ 到 $S$；//$S$ 是一个最大向前引用。

把 $S$ 加到数据库中；$i=i+1$；如果 $s_i=$NULL，结束。否则，转到步骤(Step3)。

Step3：while($S_i$ 比 $S_{i+1}$大时)//判断是否是向后引用

{$i=i+1$；}$j=i-1$；

while($s[j]$比 $s[i]$大)

{删除 $s[j]$；$j--$；}转到第 Step2。

Step4：End

Answer=Maximal Sequence $YS$。

### 2.6 内容和结构数据的预处理

内容和结构数据的预处理是根据具体的应用把 Web 页面中的文本、图像、script 以及 Web 页面间的超链接等数据转化成用于 Web 使用挖掘的格式。例如根据一个 Web 页面的文本内容，描述该页面涉及的概念主题，用于 Web 页面的聚类[8,9]；根据 Web 页面之间的超链接信息构造 Web 站点的拓扑结构图，用于辨识用户。

### 2.7 数据预处理的结果

经过以上的预处理后，可以得到一个页面集合 $P=\{p_1, p_2, p_k, \cdots, p_n\}$和一个用户事务集合 $T=\{t_1, t_2, \cdots, t_m\}$其中 $t_i \in T$ 是 $P$ 的子集。我们可以把一个事务 $t$ 看成是一个具有 $l$ 长度的序列对 $t=\langle(p_1^t, W(p_1^t)), (p_2^t, W(p_2^t)), \cdots, (p_l^t, W(p_l^t))\rangle$其中 $p_i^t(i=1,2,\cdots,l)$是 $P$ 中的页面，$W(p_i^t)$是页面 $p_i^t$ 在事务中的权重。页面的权重可以根据不同的需要采取不同的策略赋值有两种常用的赋值策略，一种是权值为二进制，表示在一个事务中这个页面是否存在，另一种是使用用户在这个页面上的驻留时间的函数，在协同过滤中，页面的权值是通过用户的评价而计算的。用户的事务可以看成是集合(不考虑用户之间的顺序信息)，也可以看成序列(考虑用户之间的顺序信息)，这需要具体的分析和应用的目标决定。

**总结** 介绍了 Web 使用挖掘的前期工作——数据预处理的过程，其目的就是尽量使得预处理后的数据比较真实和完整，为后面的数据挖掘打好基础。此外，还可以根据 Web 使用挖掘的目的在此基础上进行改进，为 Web 使用挖掘的预处理提供更好的算法。

## 参 考 文 献


1 李雄飞，李军. 数据挖掘与知识发现[M]. 北京：高等教育出版社，2003

2 Pyle D. Data Preparation for Data Mining. Morgan Kaufmann Publishers Inc, San Francisco, CA, 1999. 540

3 Cooley R, Mobasher B, Srivastava J. Data preparation for mining World Wide Web browsing patterns. Journal of Knowledge and Information Systems, 1999, 1(1)：5～32

4 Tan P, Kumar V. Discovery of Web robot sessions based on their navigational patterns. Data Mining and Knowledge Discovery, 2002, 6：9～35

5 Jetal S. Web Usage Mining：Discovery and application of usage patterns from Web data[J]. SIGKDD Explorations, 2000, 1(2)：12～23

6 Cooley R, Mobasher B, Srivastava J. Data Preparation for Mining World Wide Web Browsing Patterns[J]. Journal of Knowledge and Information Systems, 1999, 1 (1)：5～32

7 Chen MS, Park J S, Yu PS. Data Mining for Path Traversal Patterns[A]. In：Proc. of the 16th Int'l Confon Distributed Computing System[C]. Hong Kong, 1996

8 Perkowitz M, Etzioni O. Towards adaptive Web sites：Conceptual framework and case study[J]. Artificial Intelligence, 2000, 118：245～275

9 Etzioni P M. Adaptive Web Sites：Automatically Synthesizing Web Pages[A]. In：Proceedings of Fifteenth National Conference on Artificial Intelligence[C]. Madison, 1997