Web 使用信息挖掘综述*)

郭 岩[1,2] 白 硕[1] 于满泉[1,2]

(中国科学院计算技术研究所软件室 北京 100080)[1] (中国科学院研究生院 北京 100039)[2]

摘 要 Web 使用信息挖掘可以帮助我们更好地理解 Web 和 Web 用户访问模式,这对于开发 Web 的最大经济潜力是非常关键的。一般来说,Web 使用信息挖掘包含三个阶段:数据预处理,模式发现和模式分析。文章以这三个阶段为框架,分别介绍了数据预处理的技术与困难,Web 使用信息挖掘中常用的方法和算法,以及主要应用。

关键词 数据挖掘,Web 挖掘,Web 使用信息挖掘

Survey of Web Usage Mining

GUO Yan[1,2] BAI Shuo[1] YU Man-Quan[1,2]

(Software Division, Institute of Computing Technology, Chinese Academy of Sciences, Beijing 100080)[1]

(Graduate School of the Chinese Academy of Sciences, Beijing 100039)[2]

Abstract Web usage mining can help us to understand Web and Web usage patterns deeply, which is crucial to develop full economic potential of Web. In general, Web usage mining consists of three phases, namely preprocessing, pattern discovery, and pattern analysis. This paper describes each of these phases in detail. Technologies and difficulties in preprocessing, common approaches and algorithms in pattern discovery, and main applications of Web usage mining are provided.

Keywords Data mining, Web mining, Web usage mining

1 引言

随着 Internet 在流量、规模和复杂度等方面的飞速增长,WWW 成为一个巨大的、分布广泛的、全球性的信息服务中心,它涉及新闻、广告、金融管理、教育、电子政务、电子商务等各方面的信息服务。Web 中包含了 Web 页面的内容信息、丰富的超链接信息,以及 Web 页面的访问和使用信息,为数据挖掘提供了丰富的资源。Web 挖掘旨在使用数据挖掘技术从 Web 资源中发掘出有用的规律和模式。Web 挖掘可以分为三类:

• Web 内容挖掘(Web Content Mining):对 Web 页面的内容进行挖掘,如传统的 Web 搜索引擎技术[4]。

• Web 结构挖掘(Web Structure Mining):对 Web 页面之间的结构进行挖掘,如 HITS 算法[5]。

• Web 使用信息挖掘(Web Usage Mining):对用户访问 Web 时留下的访问记录进行挖掘。Web 使用信息体现了用户使用 Web 资源的行为特点,以及隐藏在行为背后的更深层次的动因和规律。因此,Web 使用信息的挖掘作为 Web 挖掘的一个重要组成部分,有其独特的理论和实践意义。

本文的重点是对 Web 使用信息挖掘作综述,Web 内容挖掘和 Web 结构挖掘都不在本文的讨论范围内。文[1]对 2000 年以前 Web 使用信息挖掘领域的研究作了综述。文[2]从 Web 个性化角度对近年来的 Web 使用信息挖掘研究作了总结。文[3]着重介绍了 Web 使用信息挖掘中的数据预处理,并对 1999 年以前 Web 使用信息挖掘方面的研究作了小结。本文的部分内容将参考文[1~3],此外还将介绍近年来相关领域的主要研究成果。

文[6~8]第一次提出将数据挖掘技术用于 Web 使用信息这一概念。文[6]提出最大前向引用(maximal forward references)的概念,用于在 Web 日志预处理阶段辨识用户访问事务。文[7]从 Web 日志中发掘频繁访问路径。文[8]利用 Web 日志对 Web 访问者进行聚类。Web 使用信息的分析包括直接的统计,如页面访问频率,还有其他一些复杂的分析,如找出频繁访问路径等。通过 Web 使用信息挖掘,可以帮助我们更好地理解 Web 和 Web 用户访问模式,这对于开发 Web 的最大经济潜力是非常关键的。一般来说,Web 使用信息挖掘包含三个阶段:数据预处理,模式发现和模式分析。下面将详细介绍这三个阶段。

2 数据预处理

数据预处理阶段是要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。

2.1 数据的收集

2.1.1 *可使用的数据* 可用于 Web 使用信息挖掘的数据主要有以下四类:

• 使用数据(Usage):描述 Web 页面的使用模式的数据,如 IP 地址、页面访问时间等,主要存在于服务器日志中;

• 简档(User profile):描述 Web 用户的个人信息,包括注册信息以及其他一些显式或隐式的用户信息,如用户对产品、电影等对象的评价;

• 内容数据(Content):Web 页面的实际数据,如文本和图片;

• 结构数据(Structure):描述 Web 页面内容组织的数据,常用的结构信息是 Web 页面之间的超链接。

2.1.2 *数据源* 以上可用于 Web 使用信息挖掘的数据主要从 Web 服务器(Web server),客户端(client)和 Web 代理服务器(Web proxy)这三个级别收集。

Web 服务器的日志显式地记录了多个 Web 用户对单个站点的访问行为,是 Web 使用信息挖掘的重要数据源,但 Web 的多级缓存会使日志变得不那么可靠。此外,也可以利用捕包技术(Packet sniffing)直接从流入 Web 服务器的 TCP/IP 包中收集使用数据。Web 服务器还利用 cookies 和查询日志保存使用信息。除了提供使用数据,Web 服务器还提供内容数据、结构数据和 Web 页面的元信息(例如文件的大

*)本课题得到中国科学院计算技术研究所领域前沿青年基金资助(20026180-24)。

小和文件的更新时间)。

从客户端收集数据可以比较全面、准确地收集到用户数据。可以使用远程代理(remote agent),如 Javascript 或 Java applets;也可以通过修改浏览器的源代码达到收集数据的目的。但客户端的数据收集需要用户的合作,要么用户允许 Javascrip 和 Java applets 功能,要么用户自愿使用修改过的浏览器,这也是在客户端收集数据的困难所在。

Web 代理服务器作为 Web 服务器和客户端之间的中间级缓存,能够减少用户下载 Web 页面的时间,减少 Web 服务器和客户端之间的网络流量。Web 代理服务器的日志记录了多个 Web 用户向多个 Web 服务器的请求,可以用来辨识共享同一台 Web 代理服务器的用户组的访问行为。

服务器的日志格式一般都遵从 W3C 标准[10],如表 1 所示。

表 1 服务器日志示例

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"	200	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

2.2 数据的抽取

对数据源提供的数据需要进行抽象抽取。为了保证抽取结果的一致性,W3C Web characterization Activity(WCA)[9]发布了 Web 使用信息挖掘的一些概念,以下为其中的一部分:

• 用户(User):一个用户是通过浏览器访问一个或多个 Web 服务器的个体。这个定义很简单,但实际上辨识用户是一件很困难的事情(参见 2.3.1)。

• 页面文件(Page File):一个页面文件是 Web 服务器通过 HTTP 请求发给用户的文件。页面文件往往在 Web 服务器上静态存在,有时候 Web 服务器为了响应用户的请求,会动态生成一些页面文件。

• 页面视图(Page View):一个页面视图由一组页面文件组成,如 frame、图片和 script 等,它们在用户浏览器上同时显示。在分析用户行为时,不是页面视图中的所有文件都有用(参见 2.3.1)。页面视图通常与一个用户的行为相关,如一次鼠标点击(本文以下内容中的页面如无特别说明,均指页面视图)。

• 点击流(Click Stream):也称连续 HTTP 请求序列,是由用户从客户端浏览器上连续发出的 HTTP 请求序列。

• 一次访问用户(One User at a Time):是指一位通过一个客户端浏览器向一个 Web 服务器发出连续 HTTP 请求序列的访问者。这个访问者是针对 Web 服务器而言的,是 Web 服务器所能辨识的用户,通常和一个真实用户的一次访问相对应。如果一个真实的用户每隔一段较长的时间对一个 Web 服务器发出一次连续 HTTP 请求序列,那么对该 Web 服务器而言,就有多个一次访问用户进行了访问。如果一个真实的用户通过不同的客户端浏览器对一个 Web 服务器发出连续 HTTP 请求序列,那么对该 Web 服务器而言,就有不同的一次访问用户进行了访问。这个概念的提出将一个真实的用户和该用户的一次访问作了区分。

• 用户访问会话(User Session):是指由一个用户发出的对 Web 的一次连续 HTTP 请求序列。

• 服务器用户访问会话(Server Session):简称用户访问事务(User Transaction)。是指一个用户对一个 Web 服务器的一次访问,由这次访问中的请求页面序列组成。

• 访问片断(Episode):任何有意义的用户访问会话或用户访问事务的子集。

W3C Web characterization Activity(WCA)发布的这些概念是整个 Web 使用信息挖掘的基础。有些文献还在这些概念上提出了一些扩展的概念,比如文[6]提出了最大前向引用(maximal forward references)的概念,是指用户在一次访问中,点击浏览器中的回退(BACK)键之前访问的最后一个页面。例如,一个用户访问会话中包含这样的访问请求序列:$A \rightarrow B \rightarrow A \rightarrow C \rightarrow D \rightarrow C$,那么这个会话的最大前向引用就是 B 和 D。最大前向引用的意义在于用户回退后访问的页面一定是已经访问过的页面。这个概念的提出有助于辨识用户访问事务。

2.3 数据的预处理

2.3.1 *使用数据的预处理* 使用信息的预处理主要是服务器日志的预处理,一般包括以下五个步骤:

(1)数据清洗(data cleaning):用户的一次请求可能会让浏览器自动下载多个附属物,如一些图片等,下载的所有文件构成一个页面视图,造成一次请求对应多个日志项的情况。数据清洗就是要除去这些附属物对应的日志项。一般采用的方法是除去 URL 中包含后缀为 gif,GIF,jpeg,JPEG,jpg,JPG,map 等的文件的日志项。

(2)用户辨识(user identification):辨识用户的困难主要是由本地缓存和代理服务器造成的。为了提高网络的性能,减少网络流量,绝大多数 Web 浏览器缓存已请求得到页面,这样,当一个用户点击回退键时,缓存的页面被显示出来,而 Web 服务器并不知道页面被再次访问了。Web 代理服务器提供了一个中间层的缓存,给用户的辨识带来了更多的麻烦,例如,所有通过一个 Web 代理服务器的请求都具有相同的 IP 地址,造成多个用户的请求被误认为单个用户请求的情况。那么,如何较好地辨识用户呢? cookies 是 Web 设计者用来标记和跟踪访问用户的,可以使用 cookies 来辨识用户,但这需要用户允许浏览器使用 cookies。还可以通过用户注册来标记用户,但注册往往被看成是一种对隐私的侵犯,用户往往不愿意登录需要注册的 Web 站点。这些依赖用户的合作,辨识用户的方法虽然简单,但因为涉及隐私问题,所以不容易实现。一些启发式信息可以用来帮助辨识用户。例如,对于 IP 地址相同的日志记录,可以观察日志记录中主机代理(agent)这个字段,这个字段记录了访问者使用的浏览器或操作系统的版本,如 Win95、IRIX6.2 等。如果该字段显示浏览器或操作系统有所变化,那么可以假设是不同的用户使用了同样的 IP 地址。另一种辨识用户的启发式信息是通过访问日志和站点拓扑结构为每个用户构造的浏览路径。如果发现一个用户的一次请求的页面不可能通过该用户这次请求之前已访问的任何一个页面的链接直接到达,那么可以假设这个请求是具有同一 IP 地址的另一个用户发出的。但这些仅仅是启发式信息,不能完全依赖它们辨识用户。比如两个用户使用同样的 IP 地址,同样机器上的浏览器,而且他们访问同样的 Web 页面集,那么根据启发信息,他们很可能被看成是同一个用户。如果一个用

户在同一台机器上运行了两个不同的浏览器，或者他在浏览器中直接敲入URL，没有使用站点链接结构，那么根据启发信息，很可能被误认为多个用户。

(3) 用户会话辨识(session identification)：如果一个用户的日志记录跨过很长的时间，那么可以猜测，该用户多次访问了Web。用户会话的辨识就是把用户的访问日志分割成一个个的会话。一般地，以一段固定时间作为时限，如30分钟，一个用户每30分钟以内的访问序列被看作是该用户的一个会话。时限的选择可以通过日志的统计分析来确定。

(4) 补全路径(path completion)：由于缓存等原因使得访问日志中并没有完全记录用户的访问行为，补全路径就是要将用户会话中的访问路径补全，从而更好地反映用户的访问过程。用于辨识用户的方法都可用来补全路径。例如，如果一个用户的一次请求的页面不是从上一次请求的页面中链接而来，而且这次请求的页面是用户最近曾经请求过的页面，那么可以假设用户使用了浏览器的回退键，重新使用了缓存的页面副本，和辨识用户会话类似，也可以使用站点拓扑结构来帮助补全路径。

(5) 事务辨识(transaction identification)：从用户访问会话中找出有意义的页面访问序列。有不少用于辨识事务的算法，如基于最大前向引用的事务辨识[6]和基于访问长度的事务辨识[11]。基于访问长度的事务辨识方法基于这样的假设：一个用户驻留在一个页面上的时间和这个页面对于该用户来说是否重要成正比。

2.3.2 *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使用信息挖掘的格式。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[12,13]；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。

2.3.3 *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$P=\{p_1,p_2,\cdots,p_n\}$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。从概念上讲，我们可以把每一个事务t看成是一个具有l长度的序列对$t=\langle(p_1^i,w(p_1^i)),(p_2^i,w(p_2^i)),\cdots,(p_l^i,w(p_l^i))\rangle$，其中$p_i^i(i=1,2\cdots,l)$是$P$中的页面，$w(p_i^i)$是页面$p_i^i$在事务$t$中的权重。

页面的权重可以根据不同的需要采用不同的策略赋值。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，页面的权值是基于用户的评价而计算的。

用户事务可以被看成集合(不考虑页面间的顺序信息)，也可以被看成序列(考虑页面间的顺序信息)，这需要根据具体的分析和应用的目标决定。对于序列分析和频繁浏览模式的发现，必须保留事务中的顺序信息。对于聚类、分类和关联规则发掘，可以把用户事务看成集合，表示成n维页面向量，分量是页面在事务中的权重，这样得到$(m\times n)$的(用户事务-页面)矩阵。文[14]描述了一种用于聚类的用户访问频率矩阵，以Web页面的URL为行，以用户的UserID为列建立URL-UserID的关联矩阵，矩阵中的元素为用户对Web页面的访问次数。

3 模式发现

模式发现阶段旨在使用各种数据挖掘技术发掘隐藏在数据背后的规律和模式。可以使用统计、数据挖掘、机器学习和模式识别等各领域中已开发的方法和算法，但把这些方法和算法应用于Web使用信息挖掘时，要考虑Web数据的特性。常用的技术有统计分析，关联规则发掘，生成序列模式，聚类和分类，以及依赖关系的建模等。

3.1 统计分析

统计分析技术是最常用的从Web用户行为中抽取知识的方法。通过分析服务器日志文件，可以得到各种统计分析描述，如用户驻留在某页面上的时间，用户浏览路径长度的中值和平均值等。许多Web跟踪分析工具[15,16]可以定期报告一些统计分析结果，如最频繁访问页面、页面的平均浏览时间、浏览某站点的路径平均长度等。这种分析虽然看起来缺乏深度，但分析结果往往对提高系统性能，加强系统安全性，辅助网站设计，提供市场决策等方面有着不可替代的作用。文[17]把服务器日志载入数据立方体结构中，执行传统的OLAP(On-Line Analytical Processing)分析操作。

3.2 关联规则

关联规则挖掘技术用来在事务中发掘页面与页面之间的非序列关系。关联规则的生成基于页面在事务中的共现模式，即关联规则中的页面经常在同一个会话中被访问，这种共现模式不考虑页面之间被访问的顺序。

绝大多数发掘关联规则的方法都是基于Apriori算法[18,19]。Apriori算法能够找出在许多事务中频繁同时出现的对象(在Web使用挖掘中指页面)，称为频繁集。是否"频繁"取决于是否满足用户指定的最小支持度阈值。频繁集中的页面之间可能没有超链接直接连接。

IBM从Official 1996 Olympics Web site的服务器日志中发掘出如下的关联规则[20]：

——浏览了Indoor Volleyball的访问者中，45%的人还浏览了Handball的页面。

——浏览了Badminton and Diving的访问者中，59.7%的人还浏览了Table Tennis的页面。

发掘出的关联规则可以用来优化站点结构。例如一个站点并没有提供页面A到页面B的直接链接，发掘出的关联规则$A\Rightarrow B$则会提示网站设计者，应该让页面A直接链接页面B，从而方便用户浏览。关联规则还可以作为启发式信息用于缓存中的页面预取，减少用户的下载延迟。

关联规则发掘的一个较大问题就是使用全局的最小化支持度阈值。由于阈值的限制，频繁集中可能会丢掉那些较少出现但非常重要的页面。例如Web站点中的内容页面或产品导向页面往往位于整个超链接结构的较深层次，它们出现的频率一般会比在第一层的浏览导向页面少得多，所以包含内容页面或产品导向页面的规则通常会得到较小的支持度，然后被丢弃。但实际上，在Web个性化应用中，找出包含内容页面或产品导向页面的规则，并依此向用户作推荐是非常重要的。文[21]提出使用多种最小支持度阈值的策略，用户可以对不同的页面指定不同的支持度，从而保证频繁集中能够包含那些较少出现但很重要的页面。文[22]证明了在Web个性化中，使用多种支持度阈值的关联规则能够很大幅度地提高推荐质量。

3.3 序列模式

序列模式的发现是在时间戳有序的事务集中找出这样的内部事务模式：一些页面被访问后紧接着另一些页面也被访问了。从Official 1996 Olympics Web site的服务器日志中发掘出如下的序列模式[20]：

——9.81%的访问者在浏览了Atlanta主页后紧接着浏览了Sneakpeek主页。

——0.42%的访问者在浏览了Sprots主页后紧接着浏览了Schedules主页。

序列模式可以分为非邻接序列模式(Sequential patterns)和邻近序列模式(contiguous sequential pattern)两种。邻近序列模式要求模式中的页面访问是连续发生的，也就是说访问之间是邻近的；而非邻接序列模式只要求模式中的

页面访问是顺序发生的，不考虑访问之间是否邻近。邻近序列模式可以用来描述用户的频繁浏览路径[23,24]，非邻接序列模式则描述了整个站点中更通用的浏览模式。由于关联规则中的频繁集只是关注页面在会话中的出现，而不考虑它们出现的顺序，因此频繁集描述了受到最少约束的浏览模式。

Markov 模型常用来发掘序列模式。通常，一个 Markov 模型由一个状态集合$\{s_1, s_2, \cdots, s_n\}$和一个状态转移概率矩阵$\begin{bmatrix} p_{11} & p_{12} & \cdots & p_{1n} \\ p_{21} & p_{22} & \cdots & p_{2n} \\ \vdots & \vdots & \vdots & \vdots \\ p_{n1} & p_{n2} & \cdots & p_{nn} \end{bmatrix}$组成，其中 $p_{ij}(i=1,2,\cdots,n; j=1,2,\cdots,n)$表示从状态 s_i 转移到 s_j 状态的概率。可以用 Markov 模型对页面访问序列进行建模，把从一个页面的访问到另一个页面的访问看成是状态的转移，用 Markov 链描述页面访问之间的概率转移。文[25,26]使用 Markov 模型生成序列模式，用于 Web 预取和系统优化。序列模式还可用于商业和市场的决策，以及站点的优化。

3.4 聚类和分类

聚类是将具有相似特征的对象聚成一个 cluster。在 Web 使用信息挖掘中，可以进行两种有趣的聚类：用户聚类(包括用户访问会话聚类和用户访问事务聚类)和页面聚类。用户聚类是要建立具有相似浏览模式的用户 cluster。这样的规则对于电子商务中的市场决策和向用户提供个性化服务是非常有帮助的。页面聚类是要发掘具有相关内容的页面 cluster，这对于 Internet 搜索引擎和 Web 提供商都是非常有用的。

PageGather 算法[12,13]基于页面在用户访问会话中的共现对 Web 站点的页面作聚类。对聚类结果中的每一个 cluster，系统自动生成一个包含该 cluster 中所有页面链接的 Web 页面，称为索引页面。每一个索引页面反映了一组用户可能具有的共同兴趣。利用索引页面，可以提高用户的浏览效率。由于索引页面是系统自动生成的，因此 PagaGather 算法可以使 Web 站点具有自适应性。文[62,63]提出了基于密度的递归聚类算法 RDBC (recursive density based clustering algorithm,)，该算法基于 Web 服务器日志对 Web 页面进行聚类。RDBC 是基于 DBSCAN 的一种改进算法，能够智能地、动态地修改其密度参数。文[60,61]对 Web 服务器日志作用户会话聚类，得到简档。文[28]对用户的评价记录作聚类，作为协同过滤的先前步骤，试图弥补 k-nearest-neighbor 算法的规模问题。文[59] 提出了 K-paths 路径聚类方法。他们认为，用户对 Web 站点的访问代表了用户对 Web 站点上页面的访问兴趣，这种兴趣程度可以通过用户对 Web 站点上页面的浏览顺序表现出来。K-paths 路径聚类方法根据群体用户对 Web 站点的访问顺序进行路径聚类。文[29]提出了基于 Web 使用信息挖掘进行 Web 个性化服务的框架，其中把用户访问事务聚类和页面聚类作为框架中的一个组成部分。

分类是将一个对象分到事先定义好的类中。在 Web 使用信息挖掘中，分类可用于为一组特定用户建立简档，这需要抽取并选择最能描述这组特定用户的特征。分类可以使用监督学习算法，如决策树、naive Bayesian 分类器、k-nearest neighbor 分类器和支持向量机等。例如对服务器日志作分类，可能会发现类似这样的有趣规则：在 /product/Music/SunYanZi 在线定购的用户中 30%年龄在 18～25 之间，并且住在海淀区。文[46]提出了基于决策树技术的 Web 缓存替换策略。

3.5 依赖关系的建模

依赖关系的建模在 Web 使用信息挖掘中也是很有用的模式发现技术。目标是要建立能够描述 Web 领域中各变量之间有意义的依赖关系的模型。例如，建立一个描述一个用户在一个 Web 在线商店中购物的不同阶段的模型。一些概率学习技术可以用来对用户浏览行为建模，如 Hidden Markov Models[27]和 Bayesian Belief Networks。Web 使用模式的建模不仅为分析 Web 用户行为提供了理论框架，而且对提高用户的访问效率，提高网上产品销量，预测未来 Web 的资源消耗大有用处。

4 模式分析及应用

模式分析是 Web 使用信息挖掘过程中最后一个阶段。模式分析旨在根据具体的实际应用，过滤掉在模式发掘阶段得到的那些没有用的规则或模式，把有用的规则和模式转换为知识，应用到具体领域中，因此模式分析和应用是分不开的。

4.1 应用的分类

将 Web 使用信息挖掘的应用进行分类的方法有很多种，主要可根据以下五维来分类：

- 收集数据的来源：Web 服务器，Web 代理服务器，客户端。大多数应用使用 Web 服务器级数据；
- 分析数据的类型：所有应用都要分析使用数据，有些应用还需要分析内容数据、结构数据或用户信息；
- 在分析数据集中涉及的用户个数：大多数应用都涉及多用户，涉及单用户的应用往往用于提供个性化服务；
- 在分析数据集中涉及的 Web 站点数：大多数应用都只涉及单站点，涉及多站点的应用往往需要分析来自 Web 代理服务器或客户端的数据；
- 应用针对的领域：Web 个性化服务，系统优化，Web 站点辅助设计，商业智能，Web 使用特性的研究等。

下面基于应用针对的领域介绍主要应用。

4.2 应用

已经有很多基于 Web 使用信息挖掘的应用。有一些应用并不针对某个特定的领域，只是关注通用的 Web 使用信息挖掘技术。例如文[6]提出最大前向引用(maximal forward references)的概念，用于在网络日志预处理阶段辨识用户访问事务。IBM Watson 的 SpeedTracer[25]致力于数据预处理方面的研究。文[68]研究了利用 Web 服务器日志重构会话的启发信息，并对它们的性能作了评价。文[69]研究了重构会话的启发信息在预测应用(如 Web 个性化)中的影响。Web Utilization Miner(WUM) 系统[31]提供一种功能强大的挖掘语言，帮助分析者辨识感兴趣的频繁路径。文[17]把 Web 服务器日志载入数据立方体结构中，执行 OLAP (On-Line Anlytical Processing)操作，例如数据的上钻(roll-up)和下钻(drill-down)。他们的 WebLogMiner 系统能够发掘关联规则，进行分类和时序分析(如事件序列分析和倾向性分析等)。文[32,33]研制了少有的依赖客户端数据的 Web 使用信息挖掘系统。

更多的应用和具体领域有关，这里主要介绍关注个性化服务，系统优化，Web 站点辅助设计，商业智能和 Web 访问特性研究的应用。

4.2.1 Web 个性化 Web 的信息资源和信息服务的数量和复杂度都在以惊人的速度增长着，一个站点能否吸引访问者，能否成功地引导访问者获得有用的恰到好处的信息，成为这个站点是否能够成功的关键。这使得 Web 个性化成为 Web 组织者和终端用户必需的工具，也使得提供 Web 个性化服务成为 Web 使用信息挖掘的重要应用之一。

传统的个性化方法(如协同过滤等)具有一些缺陷，例如用户主观评价的可靠性差，规模的限制，面对高维和稀疏数据时推荐质量差等。把 Web 使用信息挖掘用于 Web 个性化有助于克服这些缺陷。基于 Web 使用信息挖掘的 Web 个性化的目标是：根据用户偏好向用户动态地提供特定内容。简单流程就是：把当前用户的会话(或已经保存的该用户的简档)和通过 Web 使用信息挖掘得到的使用模式相匹配，得到当前用户的兴趣偏好，然后根据当前用户的兴趣偏好向其推荐一组

对象，这些对象可能包括链接、广告、产品或服务等。

文[34]使用 Web 服务器日志发掘具有相似访问模式的用户。他们的系统包括一个离线模块用于执行用户聚类分析，一个在线模块提供 Web 页面的动态链接生成。对每个站点访问者，系统根据其当前的浏览模式将其指定到一个用户 cluster 中，基于这个用户 cluster 中其他访问者已经访问过的页面，为该访问者提供动态链接。SiteHelper[35]通过观察每一个用户的访问行为来学习用户的偏好。如果一个用户在页面中的一组关键词上逗留了相当长的时间，这组关键词就会被总结出来提供给用户。系统根据用户对关键词列表的反馈为用户推荐页面。WebWatcher[36]跟踪用户的浏览行为，辨识出用户可能感兴趣的链接并推荐给用户。对每个用户，WebWatcher 先对其兴趣做简单的描述，然后基于该用户的浏览行为和具有相似兴趣的其他用户的浏览行为学习该用户的兴趣。Letizia[39]是一个客户端主机代理(agent)，用于在 Web 上自动查找和这个用户已经看过的或标记过的页面相关的页面。页面推荐系统[38]对 Web 服务器日志作页面聚类，根据当前会话找出最接近的页面 cluster，从中选择用户没有看过的页面，以及和用户当前访问的页面没有直接链接的页面推荐给用户。

文[39～41]把关联规则挖掘的结果用于推荐系统或个性化系统。文[41]提出的 top-N 推荐系统首先从订购信息中挖掘出关联规则，然后把消费者的历史订购信息和规则的左半部分相匹配，找出这个消费者支持的所有规则，把这些规则的右半部分(物品)根据可信度从高到低排序，最后向消费者推荐前 N 件物品。基于关联规则挖掘的推荐系统面临的问题之一是，如果数据集是稀疏的，将无法生成推荐，文[39]提出了两种解决方法。文[40]提出了一个基于关联规则挖掘的推荐系统的可扩展框架。推荐算法使用有效的数据结构存储频繁集，从而能够实时地生成推荐内容，不需要从频繁集中事先生成所有的关联规则。文[67]提出这样的结论：基于关联规则和序列模式的个性化推荐模型的效果受 Web 站点的结构特性(如拓扑结构、连通度等)的影响很大。他们提出一种混合 Web 个性化系统的框架，能够根据站点的连通度和访问者在网站中的位置，在不同的推荐模型之间智能地切换。实验结果表明，当访问者浏览高连通度站点时，混合系统会选择具有较少约束的个性化推荐模型，如频繁集；当用户在站点的较深层次浏览，且当前站点的连通度较低时，混合系统会选择基于序列模式的推荐模型。文[70]描述了 WUM(Web Usage Mining)的框架，该框架尤其适用于 Web 个性化的的应用。文[71,72]利用遗传算法建立用户的简档，用于个性化系统中。

文[73]把 Web 使用信息挖掘应用于信息检索中，提出了基于用户查询日志的查询扩展统计模型。模型将用户查询中使用的词或短语与页面中出现的相应词或短语以条件概率的形式连接，利用 Bayesian 公式挑选出页面中与该查询关联最紧密的词加入原查询，以达到扩展优化的目的，提高查询精度。

文[66]提出把语义知识集成到基于 Web 使用信息挖掘的个性化过程中。他们认为目前的基于 Web 使用信息挖掘的个性化系统没有考虑当前领域的语义知识。没有这样的语义知识，个性化系统就无法根据不同类型的复杂对象的当前属性产生推荐，而且系统没有能力自动解释或推理用户模式或用户推荐。因此将语义知识集成到基于 Web 使用信息挖掘的个性化过程中是下一代个性化推荐系统的主要挑战。

4.2.2 *系统优化* 用户对 Web 是否满意取决于 Web 的性能和服务质量。Web 使用信息挖掘提供了理解 Web 流量行为的途径，有助于更好地研究 Web 缓存、网络传输、负载平衡或数据分布等策略。例如，利用 Web 代理服务器日志可以模拟各种 Web 缓存替换策略，从而为缓存替换策略的研究提供了实验环境。从 Web 代理服务器的访问信息中可以分析用户的访问模式，从而预测用户对 Web 页面的访问，提高 Web 缓存的性能。另外，安全性逐渐成为 Web 服务关注的问题，尤其是当电子商务正在以指数速度增长的今天，Web 使用信息挖掘为检测入侵、欺诈和非法进入等安全技术提供了有用的模式。

文[42]提出了用于预测时间局部性和空间局部性的模型。局部性度量可以帮助确定 Web 代理服务器的预取策略和缓存策略。文[43]从服务器日志中生成路径的简档，基于当前用户的简档和路径的简档提前生成动态的 HTML 页面，从而减少页面生成的延迟。文[44,45]使用 Web 代理服务器日志研究了页面的预取。文[46]利用 Web 代理服务器日志提出基于 Web 数据挖掘的 Web 缓存替换策略模型 DMM(Data Mining weight Model)，使缓存具有自适应性，提高了缓存的智能特性。

4.2.3 *站点辅助设计* 一个站点能否在内容和结构方面设计得足够吸引用户，对于很多 Web 商家是很关键的。Web 使用信息挖掘为网站设计者提供了详细的用户反馈，帮助他们根据实际用户的浏览情况，调整网站的网页链接结构和内容，对网站进行优化，从而更好地为用户服务。

许多应用的研究结果能够用于重新设计或改造一个站点的结构和内容。例如自适应的网站[12,13]提出 PageGather 算法，基于页面在用户访问会话中的共现对 Web 页面作聚类。对聚类结果中的每一个 cluster，系统自动生成一个包含该 cluster 中所有页面链接的 Web 页面，称为索引页面。每一个索引页面反映了一组用户可能具有的共同兴趣。利用索引页面可以提高用户的浏览效率。由于索引页面是系统自动生成的，因此 PagaGather 算法可以使 Web 站点的结构自动改变，从而具有自适应性。

4.2.4 *商业智能* 消费者是如何使用 Web 站点的，这对于 Web 电子零售商来说是非常重要的信息。Web 零售系统可以捕捉到大量的网上交易的细节，为进一步的商业智能挖掘提供数据。

文[47]提出了一个知识发掘过程，用于从 Web 数据中发掘市场智能。他们定义了一个 Web 日志数据的超立方体，用于把 Web 使用数据和用于电子商务的市场数据结合起来。此外还定义了消费者关系生命周期中的四个重要的阶段：吸引消费者，消费者的逗留，消费者购物和消费者离去。他们的知识发掘技术支持这四个阶段。

有一些商业产品提供 Web 流量分析，用于收集商业智能，如文[48～51,16]。Accrue[49]、NetGenesis[50]和 Aria[51]被设计用来分析电子商务事件，例如产品的买卖和广告点击率等。Accrue[49]提供了一个用于分析浏览路径的可视化工具。IBM 的 SurfAid[48]除了提供对页面浏览的统计分析之外，还提供了对数据立方体的 OLAP 操作和用户聚类。文[52]使用 Web 服务器日志为指定站点生成 Web 页面访问模式的可信度量(beliefs)，基于规则的未预料程度(unexpectedness)发掘有趣的规则。

4.2.5 *Web 使用特性的研究* 很多针对 Web 使用特性、内容特性和结构特性的研究都没有考虑和数据挖掘的关系。实际上，在 Web 使用特性研究和 Web 使用信息挖掘之间有很多重复的地方。

文[53]讨论了 Georgia 技术学院的研究结果。他们用于实验的 Web 浏览器 Xmosaic 是经过改造的，能够记录客户端的行为。研究结果中提供了用户和浏览器交互的细节信息，以及浏览特定站点的策略。此外他们还对各种客户端事件的出现作了详细统计，例如回退/前进按钮的点击，保存一个文件，添加标签等。文[54]提出一个模型，用于预测指定站点中的用户对各种页面访问的概率分布。模型根据页面的各种属性为

站点中的所有页面赋值。模型中使用的公式和阈值都是根据各种浏览群体的浏览模式依经验推导而来的。文[55]讨论了Web服务器的各种性能评价,以及针对不同负载,这些评价之间的关系。文[56]根据指定站点当前的负载生成定制的标准评测。这个标准评测被称为是自设置的标准评测,可利用它来研究Web服务器的规模和负载平衡。文[57]开发了可视化工具WEEV(Web Ecology and Evolution Visualization),用于研究随着时间的变化,Web的使用信息、内容信息和站点拓扑结构的相应变化。文[64]基于服务器日志,研究了中国WWW的业务特性,重点研究了Web页面请求的概率分布,Web静态页面大小的概率分布和Web静态页面的访问距离的概率分布。

5 隐私问题

随着电子商务的飞速发展,隐私问题越来越吸引人们的注意,成为Web使用信息挖掘不可回避的问题。绝大多数的Web用户希望在Web上保证严格的匿名,他们非常厌恶那些监视他们访问了哪些网站,浏览了多长时间的人。而另一方面,网站管理者想方设法地对网站各方面使用情况进行统计分析,试图优化网站的设计,从而最大限度地满足访问者。网站管理者还希望能够辨识每一位访问者,进而提供个性化服务。

为了折衷这样的矛盾,Web用户和网站管理者需要遵从一些原则,比如当网站管理者对Web使用信息作各种分析时,应尽可能地不去涉及任何一个具体用户的个人信息;网站管理者还要保证不把Web使用数据作为商品进行交换或出售;Web用户则需要预先知道要访问的站点的隐私协议,从而提供自己认为合适的个人信息。当然所有这样的原则都需要在法律的支持范围内。W3C制定了P3P(Platform for Privacy Preferences)[58],提供了用于解决隐私问题的一些准则。

小结 在Web迅猛发展的今天,几乎所有的公司、企业及政府部门都创建了网站,提供Web服务,如网上购物、产品介绍、信息发布等等。随着Web资源越来越丰富,如何利用这宝贵的资源成为大家关注的热点。Web使用信息挖掘旨在对Web使用资源进行各种定量或定性分析,揭示隐藏在数据背后的各种关系,如关联关系、时序关系、页面类属关系、客户类属关系等,找出频繁访问路径和频繁访问页面,从而向用户提供个性化服务,提高Web服务质量,为Web站点的设计者提供优化站点的参考,为企业制定更有效的市场营销策略提供依据,帮助企业确认目标市场,改进决策,获得更大的竞争优势。

一般来说,Web使用信息挖掘包含三个阶段:数据预处理,模式发现和模式分析。本文以这三个阶段为框架,分别介绍了数据预处理的技术与困难,Web使用信息挖掘中常用的方法和算法,以及主要应用。

随着语义Web概念的逐渐升温,把语义知识和领域的本体知识集成到Web使用信息挖掘过程中,将成为Web使用信息挖掘的未来发展方向。

参考文献

1 Srivastava J, Cooley R, Deshpande M, Tan P-N. Web Usage Mining: Discovery and Applications of Usage Patterns from Web Data. SIGKDD Explorations, ACM SIGKDD, Jan. 2000

2 Mobasher B. Web Usage Mining and Personalization Draft Chapter in Practical Handbook of Internet Computing. In: Munindar P. Singh, ed. CRC Press. To appear in 2004. http://maya.cs.depaul.edu/~mobasher/pubs-subject.html#usage_mining

3 Cooley R, Mobasher B, Srivastava J. Data preparation for mining world wide web browsing patterns. The Journal of Knowledge and Information Systems, 1999, 1(1). http://maya.cs.depaul.edu/~mobasher/papers/webminer-kais.ps

4 李国辉,汤大权,武德峰. 信息组织与检索. 科学出版社,2003

5 Kleinberg J M. Authoritative sources in a hyperlinked environment. In: Proc. 9th ACM-SIAM Symposium on Discrete Algorithms, 1998

6 Chen M S, Park J S, Yu P S. Data mining for path traversal patterns in a Web environment. In: Proc. of the 16th Intl. Conf. on Distributed Computing Systems, 1996. 385~392

7 Mannila H, Toivonen H. Discovering generalized episodes using minimal occurrences. In: Proc. of the Second Int'l Conf. on Knowledge Discovery and Data Mining, Portland, Oregon, 1996. 146~151

8 Yan T, Jacobsen M, Garcia-Molina H, Dayal U. From user access patterns to dynamic hypertext linking. In: Fifth Intl. World Wide Web Conf. Paris, France, 1996

9 World wide web committee web usage characterization activity. http://www.w3.org/WCA.

10 http://www.w3.org/pub/WWW/

11 Global Reach Internet Productions. GRIP. 1997. http://www.global-reach.com

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

13 Perkowitz M, Etzioni O. Adaptive Web sites: Automatically synthesizing Web Pages. In: Proc. of AAAI98

14 宋擒豹,沈钧毅. Web日志的高效多能挖掘算法. 计算机研究与发展,2001,38(3):328~333

15 access watch. http://www.accesswatch.com/

16 Webtrends log analyzer. http://www.netiq.com/webtrends/default.asp

17 Zaiane O R, Xin M, Han J. Discovering Web access patterns and trends by applying OLAP and data mining technology on Web logs. 1998. http://citeseer.nj.nec.com/zaiane98discovering.html

18 Agrawal R, Srikant R. Fast Algorithms for Mining Association Rules. In: Proc. of the 20th Intl. Conf. on Very Large Data Bases (VLDB'94), Santiago, Chile, Sept. 1994

19 Agrawal R, Srikant R. Mining Sequential Patterns. In: Proc. of the Intl. Conf. on Data Engineering (ICDE'95), Taipei, Taiwan, March 1995

20 Elo-Dean S, Viveros M. Data mining the IBM o. cial 1996 Olympics Web site: [Technical report]. I BM T. J. Watson Research Center, 1997

21 Liu B, Hsu W, Ma Y. Association Rules with Multiple Minimum Supports. In: Proc. of the ACM SIGKDD Intl. Conf. on Knowledge Discovery and Data Mining (KDD'99, poster), San Diego, CA, Aug. 1999

22 Mobasher B, Dai H, Luo T, Nakagawa M. Effective Personalization Based on Association Rule Discovery from Web Usage Data. In: Proc. of the 3rd ACM Workshop on Web Information and Data Management (WIDM01), Atlanta, Georgia, Nov. 2001

23 Spiliopoulou M, Faulstich L. WUM: A Tool for Web Utilization Analysis. In: Proc. of EDBT Workshop at WebDB'98, LNCS 1590, Springer Verlag, 1999. 184~203

24 Schechter S, Krishnan M, Smith M D. Using Path Profiles to Predict HTTP Requests. In: Proc. of the 7th Intl. WorldWideWeb Conf. Brisbane, Australia, April 1998

25 Deshpande M, Karypis G. Selective Markov Models for Predicting Web-Page Accesses. In: Proc. of the First Intl. SIAM Conf. on Data Mining, Chicago, April 2001

26 Sarukkai R R. Link Prediction and Path Analysis Using Markov Chains. In: Proc. of the 9th Intl. World Wide Web Conf. Amsterdam, May 2000

27 Wang Shi. Mining Interest Navigation Patterns with Hidden Markov Model. SCI2000, Orlando, Florida, USA, July 2000. 105~111

28 O'Conner M, Herlocker J. Clustering Items for Collaborative Filtering. In: Proc. of the ACM SIGIR Workshop on Recommender Systems, Berkeley, CA, Aug. 1999
29 Mobasher B, Dai H, Nakagawa M, Luo T. Discovery and Evaluation of Aggregate Usage Profiles for Web Personalization. Data Mining and Knowledge Discovery, 2002, 6: 61～82
30 Wu K-L, Yu P S, Ballman A. Speedtracer: A web usage mining and analysis tool. IBM Systems Journal, 1998, 37(1)
31 Spiliopoulou M, Faulstich L C. Wum: A web utilization miner. In: EDBT Workshop WebDB98, Valencia, Spain, Springer Verlag, 1998
32 Shahabi C, Zarkesh A M, Adibi J, Shah V. Knowledge discovery from users web-page navigation. In: Workshop on Research Issues in Data Engineering, Birmingham, England, 1997
33 Zarkesh A, Adibi J, Shahabi C, Sadri R, Shah V. Analysis and design of server informative www-sites. In: Sixth Intl. Conf. on Information and Knowledge Management, Las Vegas, Nevada, 1997
34 Yan T, Jacobsen M, Garcia-Molina H, Dayal U. From user access patterns to dynamic hypertext linking. In: Fifth Intl. World Wide Web Conf. Paris, France, 1996
35 Ngu D S W, Wu X. Sitehelper: A localized agent that helps incremental exploration of the world wide web. In: 6th Intl. World Wide Web Conf. Santa Clara, CA, 1997
36 Joachims T, Freitag D, Mitchell T. Webwatcher: A tour guide for the world wide web. In: The 15th Intl. Conf. on Articial Intelligence, Nagoya, Japan, 1997
37 Lieberman H. Letizia: An agent that assists web browsing. In: Proc. of the 1995 Intl. Joint Conf. on Articial Intelligence, Montreal, Canada, 1995
38 Mobasher B, Cooley R, Srivastava J. Creating adaptive web sites through usage-based clustering of urls. In: Knowledge and Data Engineering Workshop, 1999
39 Fu X, Budzik J, Hammond K J. Mining Navigation History for Recommendation. In: Proc. of the 2000 Intl. Conf. on Intelligent User Interfaces, New Orleans, LA, ACM Press, Jan. 2000
40 Mobasher B, Dai H, Luo T, Nakagawa M E. ective Personalization Based on Association Rule Discovery fromWeb Usage Data. In: Proc. of the 3rd ACMWorkshop on Web Information and Data Management (WIDM01), Atlanta, Georgia, Nov. 2001
41 Sarwar B M, Karypis G, Konstan J, Riedl J. Analysis of Recommender Algorithms for E-Commerce. In: Proc. of the 2nd ACM E-Commerce Conf. (EC'00), Minneapolis, MN, Oct. 2000
42 Almeida V, Bestavros A, Crovella M, de Oliveira A. Characterizing reference locality in the www: [Technical Report TR-96-11]. Boston University, 1996
43 Schechter S, Krishnan M, Smith M D. Using path profiles to predict http requests. In: 7th Intl. World Wide Web Conf. Brisbane, Australia, 1998
44 Cohen E, Krishnamurthy B, Rexford J. Improving end-to-end performance of the web using server volumes and proxy filters. In: Proc. ACM SIGCOMM, 1998. 241～253
45 Aggarwal C C, Yu P S. On disk caching of web objects in proxy servers. In CIKM 97, Las Vegas, Nevada, 1997. 238～245
46 Bonchi F, et al. Web Log Data Warehousing and Mining for Intelligent Web Caching. Elsevier Science, April 2001
47 Buchner A, Mulvenna M D. Discovering internet marketing intelligence through online analytical web usage mining. SIGMOD Record, 1998, 27(4): 54～61
48 Surfaid analytics. http://surfaid. dfw. ibm. com/web/home/index. html
49 Accrue. http://www. accrue. com
50 Netgenesis. http://www. netgenesis. com
51 Andromedia aria. http://www. andromedia. com
52 Padmanabhan B, Tuzhilin A. A belief driven method for discovering unexpected patterns. In: Fourth Intl. Conf. on Knowledge Discovery and Data Mining, New York, 1998. 94～100
53 Catledge L, Pitkow J. Characterizing browsing behaviors on the world wide web. Computer Networks and ISDN Systems, 1995, 27(6)
54 Huberman B, Pirolli P, Pitkow J, Kukose R. Strong regularities in world wide web surfing. Technical report, Xerox PARC, 1998
55 Arlitt M F, Williamson C L. Internet web servers: Workload characterization and performance implications. IEEE/ACM Transactions on Networking, 1997, 5(5): 631～645
56 Manley S L. An Analysis of Issues Facing World Wide Web Servers. Undergraduate, Harvard, 1997
57 Chi E H, et al. Visualizing the evolution of web ecologies. In CHI '98, Los Angeles, California, 1998
58 Platform for privacy project. http://www. w3. org/P3P/
59 王实, 高文, 李锦涛, 谢辉. 路径聚类: 在 Web 站点中的知识发现. 计算机研究与发展, 2001, 38(4): 482～486
60 Nasraoui O, Frigui H, Joshi A, et al. Mining Web Access Logs Using Relational Competitive Fuzzy Clustering. In: Proc of the 8th Int'l Fuzzy Systems Association Congress. Taiwan, 1999
61 黄松, 刘晓明, 宋自林. 基于归纳化会话的网络用户的聚类. 计算机研究与发展, 2001, 38(10): 1224～ 1228
62 Su Zhong, Yang Qiang, Zhang Hongjiang, Xu Xiaowei, Hu Yuhen. Correlation-based Document Clustering using Web Logs. http://citeseer. nj. nec. com/su01correlationbased. html
63 苏中, 马少平, 杨强, 张宏江. 基于 Web-Log Mining 的 Web 文档聚类. 软件学报, 2002, 13(1): 99～104
64 郝沁汾, 祝明发, 郝继升. WWW 业务访问特性分布研究. 计算机研究与发展, 2001, 38(10): 1172～1180
65 Wexelblat A. History-Based Tools for Navigation. In: Proc. of the Hawai'I Intl. Conf. On System Sciences. IEEE Press, Jan. 1999
66 Dai H, Mobasher B. Integrating Semantic Knowledge with Web Usage Mining for Personalization. Draft Chapter in Web Mining: Applications and Techniques, Anthony Scime (ed.), IRM Press, Idea Group Publishing. To appear in 2004. http://maya. cs. depaul. edu/～mobasher/pubs-subject. html #usage_mining
67 Nakagawa M, Mobasher B. A Hybrid Web Personalization Model Based on Site Connectivity. In: Proc. of the WebKDD Workshop at the ACM SIGKKDD Intl. Conf. on Knowledge Discovery and Data Mining, Washington, DC, Aug. 2003
68 Spiliopoulou M, Mobasher B, Berendt B, Nakagawa M. A Framework for the Evaluation of Session Reconstruction Heuristics in Web Usage Analysis. INFORMS Journal of Computing, Special Issue on Mining Web-Based Data for E-Business Applications, 2003, 15(2)
69 Berendt B, Mobasher B, Nakagawa M, Spiliopoulou M. The Impact of Site Structure and User Environment on Session Reconstruction in Web Usage Analysis. In: Proc. of the 4th WebKDD 2002 Workshop, at the ACM-SIGKDD Conf. on Knowledge Discovery in Databases (KDD'2002), Edmonton, Alberta, Canada, July 2002
70 Shahabi C, Banaei-Kashani F. A Framework for Efficient and Anonymous Web Usage Mining Based on Client-Side Tracking, WEBKDD 2001 - Mining Web Log Data Across All Customers Touch Points, Springer-Verlag, New York, 2002, ISBN 3-5404-3969-2
71 Chen Y-S, Shahabi C. Improving User Profiles for E-Commerce b Genetic Algorithms, E-Commerce and Intelligent Methods Studies in Fuzziness and Soft Computing. Vol. 105, VIII. In: J. Segovia, P. S. Szczepaniak, M. Niedzwiedzinski eds. 2002, ISBN 3-7908-1499-7
72 Shahabi C, Chen Y-S. An Adaptive Recommendation System without Explicit Acquisition of User Relevance Feedback. Distributed and Parallel Databases Journal, Kluwer Academic Publishers, 2003, 14(3): 173～192
73 崔航, 文继荣, 李敏强. 基于用户日志的查询扩展统计模型. 软件学报, 2003, 14(9): 1593～1599