基于用户访问记录的 Web 挖掘研究①

The Research of Web Mining Based on User Visiting Record

何　典　（湖南商学院计算机与电子工程系　湖南长沙　410205）
宋中山　（中南民族大学计算机科学学院　湖北武汉　430074）
刘济波　（湖南商学院计算机与电子工程系　湖南长沙　410205）

摘要：从 Web 日志挖掘存在的困难和不足出发，结合电子商务个性化服务的特点，引入用户访问记录进行 Web 挖掘，提出了一种 Web 挖掘中数据采集与预处理的新思路，指出了该思路的实现方法和特点。提出了引入用户访问记录后的 Web 挖掘体系结构。

关键词：Web 挖掘　用户访问记录　Web 日志　电子商务个性化服务

1　引言

Web 挖掘是指利用数据挖掘技术从 Web 文档和访问数据中发现和抽取知识。通常可以分为 Web 内容挖掘（Content Mining）、Web 结构挖掘（Structure Mining）和 Web 使用记录的挖掘（Usage Mining）[1]。

在电子商务中利用 Web 挖掘技术，可以在海量的 Web 访问数据中发现用户的兴趣爱好和购买习惯等，对用户进行在线推荐；可以发现用户的访问模式，用来调整网站结构，提供浏览建议，动态地为用户定制个性化的网站。通过电子商务个性化服务，使用户成为电子商务网站的中心，有利于将电子商务网站的浏览者转变为购买者、提高电子商务网站的交叉销售能力、提高用户对电子商务网站的忠诚度。

2　Web 日志挖掘技术

2.1　Web 日志挖掘的一般过程

电子商务个性化服务主要采用 Web 使用记录的挖掘，其数据主要来源是服务器端的 Web 日志文件。Web 日志挖掘的一般过程是：

（1）数据预处理。Web 日志挖掘首先对服务器端的 Web 日志文件中的数据进行预处理。数据预处理要去掉原始数据中的“噪声”，并且使处理后的数据便于实施挖掘算法。数据预处理的质量与 Web 挖掘的效率和结果密切相关。数据预处理的内容包括：数据净化、用户识别、会话识别、事务识别、页面过滤、路径补充等[2]。

（2）模式识别。然后要对预处理后的数据进行模式识别，即实施挖掘算法。模式识别的基本方法有：统计分析、聚类、关联规则、频繁访问序列、依赖建模、最近邻技术等。

（3）模式分析。模式分析的目的是根据实际应用，通过观察和选择、把发现的规则、模式和统计转换为知识。常用的手段有：信息过滤、可视化、联机分析处理等。

2.2　Web 日志挖掘的不足

Web 日志挖掘主要是提供面向用户的信息分析，所以首先要从 Web 日志中对用户和用户会话（一个用户在规定的时间内请求的所有 Web 页面）进行识别，以此作为信息分析的基础。由于本地缓存、代理服务器和防火墙的存在，使得 Web 日志中的数据并不精确[3]。例如，用户如果通过代理服务器或者防火墙访问网站，服务器端 Web 日志中记录下的 IP 地址将不准确，这增加了识别用户和用户会话的难度。在一些 Web 日志挖掘的实际应用中，利用客户端浏览器中的 Cookie，采用 Cookie 技术，通过服务器端得 Cookie 日志来识别用户和用户会话[4]，但是某些客户端的浏览

① 项目来源：06C460 湖南省教育厅科学研究项目（面向电子商务个性化服务的 Web 挖掘研究）

器由于安全原因禁用了 Cookie，使得识别难以进行。

而且，Web 日志非常庞大，热门的 Web 站点每天都有数百兆的日志记录[1]，但是其中大量的记录是无用的，提供的可用信息太少，有些甚至会影响挖掘结果，这使得数据预处理工作量比较大。动态页面的大量使用也使得分析日志更为困难。另外，一些重要数据例如查询关键字等没有记录，而这些数据可能与用户的兴趣密切相关。

针对 Web 日志挖掘存在的困难和不足，根据电子商务个性化服务的要求，我们提出在 Web 挖掘中使用用户访问记录作为主要数据来源。

3　用户访问记录在 Web 挖掘中的应用

3.1　用户访问记录的引入

由于网络速度和计算机软硬件处理能力的大大提高，使得服务器可以在用户访问网站的同时记录用户访问信息。具体做法是：在网页设计时，对各链接对象进行设置，当用户访问该链接对象时，系统自动将用户的信息和访问对象的信息实时传递到服务器端的用户访问记录文件中，然后结合历史数据和客户实时访问的信息进行数据挖掘和在线推荐，向用户提供个性化服务。这样，Web 挖掘将不依赖 Web 日志文件。

引入用户访问记录带来的网络访问时延并不足以影响用户的正常访问。在用户访问各链接对象的同时写用户访问记录，在网络中传递的不过是少量的文本，相比传递网页、图片、图像等数据来说，占用带宽和所需要的时间可以忽略。正因为数据量很小，所以写记录文件所需要的时间也是很少的。

3.2　用户访问记录的组成

一般的 Web 日志的内容包括：IP 地址、请求日期时间、方法（POST/GET）、被请求页面的 URL、HTTP 版本号、返回码、传输字节数、用户使用的操作系统以及浏览器的类型、页面请求的来源地等。

根据电子商务个性化服务的目标和 Web 挖掘的要求，用户访问记录可以设计为以下三个数据集：

（1）用户 URL 访问记录（见表 1）：

表 1　用户 URL 访问记录

数据项	说明
UserIP	用户 IP 地址
UserID	用户标识
TargetType	访问对象类型
TargetURL	对象 URL
Date	访问日期
Time	访问时间
Agent	代理
Referrer	页面请求的来源地

其中，用户标识是指该电子商务网站用户的用户名或者会员标识；访问对象类型包括导航页、内容页、资源、广告、图片、声音、视频等；代理是指用户使用的操作系统和浏览器的类型以及版本号。

（2）用户查询记录（见表 2）

表 2　用户查询记录

数据项	说明
UserIP	用户 IP 地址
UserID	用户标识；
SourceURL	查询所在网页的 URL
KeyWord	查询关键字
KeyValue	查询值
Date	访问日期
Time	访问时间
Agent	代理

（3）用户购物记录（见表 3）

表 3　用户购物记录

数据项	说明
UserIP	用户 IP 地址
UserID	用户标识
Date	访问日期
Time	访问时间
Action	用户动作
ObjectID	商品号
Agent	代理

其中 Action 表示的是用户将商品移入还是移出购物篮的动作。

3.3　用户访问记录在 Web 挖掘中的作用和特点

可以看出，用户访问记录既包含了一般的 Web 日志中与 Web 挖掘相关的数据项，又有其自身的特点。

(1) 引入了 UserID 这个参数。该参数是登录该电子商务网站的用户的用户名,或者会员标识等。在大部分电子商务网站中,用户交易是需要先输入用户名和密码进行登录后,才能进行交易。所以,针对电子商务个性化服务的 Web 挖掘,可以采用跟踪特定用户的访问的办法,记录他的访问时间、对象等信息。而 Web 日志中并没有记录该电子商务网站的会员标识。

有了 UserID 这个参数标识用户后,避免了由于代理服务器、防火墙的问题造成的 IP 相同造成的用户识别困难,极大地方便了用户识别工作,从而方便了后续数据预处理和模式识别工作。

当用户访问网站,输入其 UserID 登录后,个性化服务系统会立即根据 UserID 找到对应该用户的 Web 挖掘结果,很快生成对该用户的页面定制、浏览推荐等。

特别地,对于初次访问和未注册用户的访问行为,可以统一分配一个匿名标识进行记录。针对这些记录的数据挖掘,可以采用 IP 地址识别用户。如果 IP 地址相同时,可以采用 IP 地址结合代理进行用户识别。在代理这一数据项中记录的是用户所用的操作系统和浏览器的类型及版本号,当两条记录的 IP 地址相同而代理不同时,可以认为这两条记录来自不同的用户。

(2) 用户访问记录并不是将用户访问的所有对象都记录下来,而是有选择性的记录。在设计网站前要根据该网站的特点进行一些预测和分析,看哪些对象需要记录用户访问信息,而哪些对象不需要记录。一些与挖掘算法无关的对象,例如装饰用的图片等,则不需要记录用户访问信息。这样,用户访问记录的数据量要比 Web 日志小得多,也在一定程度上避免那些有可能延长挖掘时间、影响挖掘结果的数据,从而提高 Web 挖掘的效率和质量。

(3) 用户访问记录的数据采集的同时就已经对数据进行了分类,如区分用户访问的对象是导航页、内容页、资源,还是图片、声音、视频、广告等。这样的分类为后续数据挖掘打下了很好的基础。例如,当挖掘用户的频繁访问序列时,那么对于内容页、图片、资源等处于网站拓扑末端的访问对象就可以首先排除。这样,待挖掘的数据集将大大减小,从而可以降低挖掘错误,提高挖掘性能,并且,在挖掘算法的时间复杂度很高的情况下,可以减少挖掘时间。

(4) 用户访问记录在数据采集的同时已经完成了列维的确定。在 Web 日志挖掘中,由于 Web 日志中有些数据项与 Web 挖掘无关,所以需要在数据预处理时进行降维,将这些数据项去除。用户访问记录的每一个数据项均与 Web 挖掘相关,避免记录与 Web 挖掘无关的数据项,实际上在数据采集时已经完成了数据预处理中减少列维的工作,提高了数据预处理的效率。

(5) 用户访问记录还了记录用户的其他信息,如记录用户查询关键字,记录用户的购物过程(用户将什么商品移入/移出购物篮),记录购物结果等。这些信息丰富了数据挖掘所需要的可用信息。这些信息结合用户的注册信息、交易信息、访问信息等,也是商品在线推荐的主要数据来源。

可见,在面向电子商务个性化服务的 Web 挖掘应用中,基于用户访问记录比基于 Web 日志更具有优势。

4 Web 挖掘的体系结构

引入用户访问记录后,Web 挖掘的体系结构可以通过图 1 来表示。

在图 1 中,用户访问数据记录是 Web 挖掘的主要数据来源。交易数据库存放的电子商务网站用户交易的情况,包括用户交易情况,用户所获得的产品,另外还包括对产品的评价,问卷调查结果等。

一般来说,Web 挖掘和在线推荐的特征的获取和规则生成是离线处理的,而当用户再次登录该网站时及时返回结果,通过个性化智能 Agent 进行个性化服务。

挖掘算法和推荐策略可以根据不同类型站点的要求来具体选择,挖掘结果和推荐集通过个性化智能 Agent 反馈给用户。

对于电子商务网站的会员用户,通过会员标识登录网站以后,其访问信息将会被记录到服务器端。这些数据将在经过预处理后,在专用的数据挖掘模块中,通过具体的挖掘算法和推荐策略来进行模式识别和模式分析。用户访问信息也会传到个性化智能 Agent,Agent 根据用户的标识,向挖掘模块抽取对应用户的挖掘结果和推荐集,将其可视化地反馈给用户,达到个性化服务的目的。

采用 Agent 代理访问技术,将在线的挖掘和推荐

工作交给个性化智能 Agent 去做，减轻了 Web 服务器的压力，而且可以提高系统的智能化，使个性化服务具有自主性、自学习、应激性和合作性等特点[5,6]。

对于首次访问或未注册用户，可以采用 Web GIS 技术进行电子商务个性化服务[7]。来自同一地点，如同一城市、同一社区的用户可能有相同或者相似的访问模式、交易习惯、消费兴趣等，而 IP 地址在一定程度上与地理位置相关。所以，可以根据用户 IP 地址，利用以前对应该 IP 地址地理范围的用户的 Web 挖掘结果，来预测和估计该用户的访问习惯、消费行为等，对其进行个性化服务，从而提高客户获得的成功率。

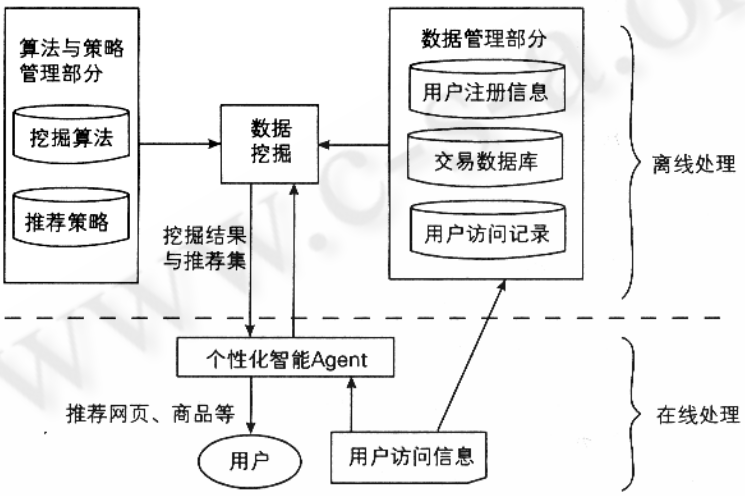

图 1　引入用户访问记录后的 Web 挖掘体系结构

5　结束语

Web 挖掘的研究在电子商务中个性化服务中的实际应用并不多见。本文提出了一种新的思路来进行数据采集和数据预处理，为后续的挖掘算法的实现打下了良好的基础。

参考文献

1 Jiawei Han, Micheline Kamber. Data Mining Concept and Techniques[M]，范明、孟小峰等译，机械工业出版社，2001.
2 陆丽娜，Web 日志挖掘中的数据预处理的研究[J]，计算机工程，2000.26(4)：66～67.
3 杨怡玲，一个简单的日志挖掘系统[J]，上海交通大学学报，2000.34(7)：932～935.
4 肖立英，Web 日志挖掘技术的研究与应用[J]，计算机工程，2002.28(7)：276～277.
5 苏安洋，电子商务中的 Agent 虚拟协商与智能决策[J]，中国工程科学，2003.5(10)：56～62.
6 李煊，基于关联规则挖掘的个性化智能推荐服务[J]，计算机工程与应用，2002.38(11)：200～204.
7 毛克彪，基于 Web GIS 的电子商务数据挖掘研究[J]，测绘学院学报，2003.20(3)：180～182.
8 李颖基，Web 日志中有趣关联规则的发现[J]，计算机研究与发展，2003.40(3)：435～439.