CN43-1258/ TP
ISSN 1007-130X

计算机工程与科学
COMPUTER ENGINEERING & SCIENCE

2007 年第 29 卷第 2 期
Vol. 29 ,No. 2 ,2007

文章编号:1007-130X(2007)002-0036-04

一种新型的 Web 挖掘数据采集模型*

A New Web Mining Data Integration Model Based on XML

胡迎松,宁海霞
HU Ying-song, NING Hai-xia
(华中科技大学计算机科学与技术学院,湖北 武汉 430074)
(School of Computer Science and Technology, Huazhong University of Science and Technology, Wuhan 430074, China)

摘 要:本文在简要论述了当前 Web 挖掘采用的数据源不足后,分析了 XML 文档结构与 Web 挖掘算法结构的相似性,提出了采用 XML 技术在应用服务层采集用户访问数据的数据源模型 X-DIM,并分析了它的优越性。该模型克服了以往基于 Web 访问日志在数据预处理中的一系列问题,具有数据完备、准确度高、便于为挖掘算法使用等优点,有较高的应用价值。

Abstract: The paper briefly describes the demerits of insufficient data sources adopted in the current Web mining, analyses the similarity between the XML document structure and the Web mining algorithm structure, proposes a data source model X-DIM of adopting the XML technology in the application service layer to sample users' access data, and analyes its advantages. The model overcomes a series of problems previously encountered in data preprocessing based on the Web access log, and features the merits of data completeness, high accuracy, ease of use in mining algorithms, and high application value.

关键词:XML;X-DIM;Web 挖掘;电子商务

Key words: XML; X-DIM; Web mining; E-commerce

中图分类号:TP311.13

文献标识码:A

1 引言

信息时代的网络技术日新月异,网络计算环境给商品流通领域和消费者购买行为带来了革命性的变化。电子商务的开展为企业的销售和消费者购买商品提供了极大便利,所有网上行为(用户行为)的可记录性和数据的迅速增长促使 Web 挖掘技术广泛运用到该领域中。利用这种技术可以对用户访问记录进行处理,以获取企业感兴趣的知识,发现潜在市场,更好地进行市场决策[1]。

在电子商务中应用 Web 挖掘的主要目的是发现用户访问站点的浏览模式,主要关注的是从访问信息中挖掘出关联规则、序列模式、分类规则和聚类分析等知识模型。一般而言,Web 挖掘过程主要分为资源发现和数据预处理、算法实施、模式分析以及分析验证等几个阶段[2]。目前,国内外各大供应商在算法的设计、分析和改进方面做了大量的研究。然而,对于数据预处理过程中如何选择一个数据源,又如何采集有效的数据这些方面,关注并不太多。高质量的数据源对整个挖掘过程至关重要,是 Web 挖掘过程的前提部分。本文以 Web 挖掘在电子商务中的应用为背景,利用 XML 技术,结合应用层日志,提出了一种新的基于会话的数据采集模型 X-DIM(Data Integration Model based on XML,简称 X-DIM),来提高数据源的综合性、易用性和实效性。

2 当前数据收集研究

当前电子商务 Web 挖掘的数据大部分来自于记录了用户在 WWW 上的网站访问浏览活动的服务器端的访问 log 文件,其它信息则来源于用户登记信息数据以及通过工具(如 CGI、Javascript)收集到的统计数据。服务器端 log 文件记录了来自用户的客户端 IP、访问的页面、代理服务

* 收稿日期:2005-09-02;修订日期:2005-10-27
作者简介:胡迎松(1966-),男,湖北武汉人,教授,研究方向为基于网络的计算机应用;宁海霞,硕士生,研究方向为基于网络的计算机应用。
通讯地址:430074 湖北省武汉市华中科技大学计算机科学与技术学院;Tel:(027)87547244;E-mail:ninghaixia@126.com
Address: School of Computer Science and Technology, Huazhong University of Science and Technology, Wuhan, Hubei 430074, P. R. China

器端 ID 等。P. Oirolli 等人利用信息检索技术,结合路径访问模式和站点拓扑结构来实现用户个性化访问[1],就是应用了基于此记录的 Web 日志挖掘算法。面对通过以上途径获得的看似海量、取之不尽的用户信息,在接下来的数据预处理工作中,包括数据清洗、数据规约等,通常会碰到以下问题[1]:

(1) 代理引发的数据真实性问题。在现实中,用户经常通过代理服务器访问网站,服务器日志记录了代理服务器端 Agent ID,而忽略了用户的真实 ID。这样,利用服务器端日志进行数据挖掘时,就存在单 IP-多用户、多 IP-单用户、多 IP-单会话(用户与网站会话中途亦可能更换代理)等情况。目前的 Web 分析工具仅能提供用户访问网站 Web 页的统计次数,对于大量的访问日志,正确识别用户和真实会话则无能为力,数据源的可信度降低[3]。

(2) 匿名信息缺失问题。目前,许多商业网站的销售活动仅仅限于已注册的消费群体;对于尚未注册、仅抱着浏览态度的匿名登录用户,有些网页或是不可见,或者网站干脆将其拒之门外。由于访问日志记录信息的局限性和缺少必要的用户登记信息,这些访问者的访问记录在进行数据过滤时往往显得多余。然而,实际上这些访问者的信息对于发现潜在用户群体是十分必要的。

(3) Cache 机制问题。为了减少网络流量,客户端采用了 Cache 机制暂存网页,导致服务器端 log 对用户访问的 Web 部分信息记录缺失。即使采用推理网站拓扑结构补全序列,但与用户实际访问情况仍有所差别。

(4) 保存信息不足问题。Web 服务器记录的只局限于用户访问网站浏览到的网页,而用户与网站交易的最终结果则记录在数据库文件中。用户对网页的操作,例如某用户对特定商品的查询、购买等操作并没有记录下来。而这些数据对于发现用户的购买心理和行为模式等十分关键。

分析以上问题不难得出结论:以往的 Web 挖掘数据来源虽然广泛,但作为用户访问挖掘的基础仍存在片面性和孤立性,而且仍不够充分。它们之间不能实现有效的结合,忽视了网站的需求,给数据分析人员带来了很多困扰。如果能够寻求到一种新的数据源模型,它可以克服以上数据源的不足,尽可能地采集到用户和网站会话期间来自用户各方面的访问数据,实现来自访问日志的用户浏览记录和用户对网页操作记录的有机融合,将极大地有助于分析人员对用户行为模式的研究[3,4]。

下面将根据在商务网站发布的网页上的各种数据类型进行 UML 建模,构造出用户和网站会话期间的数据结构类图,并结合当前因特网广泛应用的 XML 技术,提出一种新型的数据采集模型,以实现对用户访问网站涉及到的各种数据的采集与集成[3,5]。

3 基于 XML 的数据采集模型 X-DIM

3.1 Web 数据分类

Web 上的数据来源广泛,形态各异,结合用户与电子商务网站会话的实际情况,我们将 Web 挖掘过程涉及到的数据分为三类:页面结构数据、用户数据和访问数据。如何将这些数据有效地结合、收集起来,形成高效的、便于处理利用的数据源,是本文要关注的主要问题。

3.2 X-DIM 模型

在这里,以一个典型的电子商务活动为背景,来分析抽象出这个活动所要涉及到的各个实体。很容易想到,用户、商品以及用于发生交易的平台——商务网站,是不可缺少的部分。为了要将用户尽可能多的信息记录下来,需要进一步把用户的一些动作抽象出来,而在交易过程中发生的搜索某种特定商品、改变商品数目、查看购物篮等操作作为用户的某种特定的动作出现。另外,出于网站安全性以及性能上的考虑,交易都是在用户一次会话过程中完成的。因此,模型 X-DIM 将涉及到用户、会话、网页、交易、商品和用户动作这几个部分。对于各个部分的相关特性,涉及到的操作描述如下:

(1) 用户(USER):它包含的属性有用户帐号(SUBSID)、会话参数(MID)、密码(PASSWORD)、登录信息(REGINFO)、状态(STATUS)、客户主机(DEVICEID)、请求页面的 URL;涉及的操作有登录(login)、退出(logout)。

(2) 会话(SESSION):包含的特性有用户帐号(SUBSID)、访问时间(ACCESSTIME)、会话生存期(SURVTIME);涉及的操作为退出会话(logout)。

会话是用户访问网站的核心,用户浏览网页均在会话期间完成。用户与网站的会话需要有最大会话时间限制,系统将对用户和网页会话时间进行判断:如会话时间过长,超出会话生存期,系统会认为用户离站,将自动结束会话(logout)。

(3) 网页(WEBPAGE):包含了用户浏览网页的 URL、该网页的商品列表、该网页的关键字和用户访问时间等特性。网页与交易动态关联,这是由用户消费行为的无目的性和随机性决定的。

(4) 交易(TRADE):交易类封装了用户在商务网站的交易,每一笔交易都涉及不同事件,包括用户帐号(SUBSID)、操作时间(ACCESSTIME)、所在页的 URL、商品名称(COMMODITYID)、数量(AMOUNT)等属性。另外,交易还涉及不同的事件,而事件又涉及用户帐号和商品 ID。交易需返回给系统布尔变量表示交易是否成功(ISSUCCEED)。交易要求在会话期间完成。

(5) 事件(EVENT):事件是交易的派生,即交易是由各种事件组成的。事件中有两个属性是固定的,即访问时间和商品数量。事件名称(ACTIONID)可以是搜索 SEARCH、购买 BUY、放弃购买 ABANDON、改变商品数目 CHANGEAMOUNT 等。

(6) 商品(COMMODITY):网页上涉及的商品与现实买卖的商品相比,是一种抽象概念,仅提供 ID 和所在的 Web 页的 URL。

根据上面的描述,我们建立了如图 1 所示的 X-DIM 模型的 UML 静态结构图,它体现了用户、会话、网页和商品这些在电子商务交易中的关系,将用户、网站和访问信息三种数据类型有机融合。

在 Web 挖掘预处理过程中,该数据采集模型具备以下一些优点,具体包括:

(1) 最大限度地保存与用户相关的数据。模型中对用户类设置了 SUBSID 属性,当用户为网站注册用户时,用户将保存在 SUBSID 里,用户对于商品的查询、购买(发生交

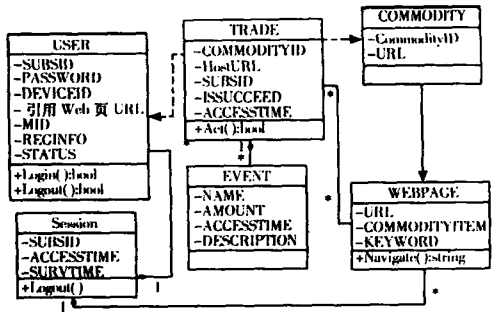

图 1　X-DIM 模型 UML 静态结构图

易)等数据也将保存在与这个 SUBSID 相关的数据里;对于非注册(匿名)用户,对于相关网页的查询、商品的浏览和查询等信息也将保存起来,把 SUBSID 设置成 NULL,表示并不与具体用户关联,这些数据也能作为后期数据挖掘的依据。

(2) 尽可能准确地保存用户数据。当是网站的注册用户时,以会话为基础的用户访问数据记录可以很好地将用户在浏览网站时进行的操作封装在基于同一 SESSION 的数据中,这样的数据能使后期数据识别更加准确,数据源可信度得到了提高。

(3) 数据融合。将 Web 挖掘所需的部分来自服务器访问日志的网页浏览统计数据及用户信息和用户交易操作数据等进行了动态集成,使来自各方面的用户信息相互融合。

Web 数据多为半结构化和非结构化,用户对网页的操作是动态、多变的。从某种角度上讲,Web 上的数据处理模式和 XML 文件的树型结构具有相似性结构,可以被认为是一些节点的集合。目前,各种大型商用数据库系统对 XML 都提供了完善的支持,与用户密切相关的数据全部以 XML 的形式封装。这种封装形式将极大地有利于数据导入、交换等操作。而且,XML 这种树型结构的数据,为一些常用的数据挖掘算法提供了便利。根据 XML Schema 事先定义的 XML 文档结构对用户数据分类,可以很容易实现对用户的访问模式变化的获取。通过对某一网页的连续操作关联分析,可以得到用户购买行为和网页之间的潜在关联。

本文利用 ASP.NET 技术和 XML 技术在服务器端实时跟踪用户访问操作。当用户来访时,将用户访问数据按照预先设定的模式封装成 XML 格式的数据流,记录于应用层日志文件。这些数据可以日后保存在数据库中,以便于对用户访问历史纪录进行进一步的研究分析。

下面列出一个典型的以 XML 格式封装的数据流。假设某用户 *u*1 在电子网站访问期间进行了包括浏览网页、购买商品(COMMODITYID 为 001)以及修改等操作,应用层日志记录如下:

```
< ? xml version =“1.0 ”encoding =“utf-8 ”? >
< Session >
    < ACCESSTIME > 12/ Mar/ 2005; 13: 34: 53-0600 < / ACCESSTIME >
    < SURVTIME > 30 minites < / SURVTIME >
    < User SUBID =“u1 ”PASSWORD =“123456 ”DEVICEID =“202. 69. 194. 72 ”REGINFO =“ ”STATUS =“1 ”REQDEVICEID =“http: / / www. ppyt. com ”> < / User >
    < WebPage WebURL =“ http: / / www. ppty. com/ new/ Index. asp ”KeyWord =“new ”>
    < Commodity >
        < CommodityItem id =“001 ”/ >
        < CommodityItem id =“002 ”/ >
    < / Commodity >
    < / WebPage >
    < TradeList >
        < Trade id =“t1 ”CommodityID =“001 ”HostURL =“http: / / www. ppyt. com/ index. asp ”isSucceed =“true ”>
            < event name =“AddtoCart ”AMOUNT =“30 ”description =“NULL ”/ >
            < event name =“ChangeAMOUNT ”AMOUNT =“10 ”description =“NULL ”/ >
        < / Trade >
    < / TradeList >
< / Session >
```

正如上面所描述,用户访问网站时,服务器应用层日志生成了一个以 Session 为根节点的 XML 数据包,其中包含了用户访问电子商务网站期间的一切操作——浏览网页以及所作的交易。设置最大会话时间的意义在于排除了因多 IP-单会话等原因导致的会话模糊,能更加有效地进行会话识别。XML 语言的动态扩展性支持文档内部和文档之间的关联,很好地满足了 Web 挖掘的实时需要。建立在 XML 技术上的 X-DIM 利用了 XML 语言的平台无关性和 XML 文档内部的动态关联属性,为下一步的数据挖掘工作,特别是关联规则发现和序列模式预测以及数据文件在各个异构数据库之间的传输奠定了基础。

4　应用分析

X-DIM 模型参照用户在访问网站期间各种访问数据之间的属性关联,利用 XML 技术的树结构父子节点路径关系使看似毫无关联的操作很好地融合起来。根据日志已存的 XML 数据包,可以很容易生成关系数据库表:

(1) X-DIMSUBSCRIBER:用户信息表,包含字段 SUBSID、MID、PASSWORD、DEVICEID、REGINFO、STARTUS);

(2) X-DIMSUBSCRIPTION:用户访问表,包含字段 MID、ACCSSTIME、COMMODITYID、HOSTURL);

(3) X-DISUBSCRIPTIONHISTORY:用户访问明细表,包含字段 HISTORYID、MID、SUSID、EVENTNAME、DESCRIPTION、ACCESSTIME、COMMODITYID、AMOUNT、ISSUCCEED)。

服务器应用层日志负责对各种事件进行捕捉,生成的事件类型将适时保存到数据库的相应数据表中,在以后需要的时候可以根据这些表的记录生成用户的会话列表。X-DIM 在数据采集中既极大地融合了浏览、购买等各种事件,又兼容以往的 Web 挖掘算法,并支持多维数据挖掘。

电子商务中经常应用的 FP 算法是一种典型的应用树结构进行单维关联规则推导的算法,在用户访问数据构造频繁模式树 FP-tree 的过程中经常重复的操作是:

(1) 利用解释器将 Session 中的复合事件拆分为简单事件;

(2) 扫描 Sessions 并收集一项项目次数,按照支持度排序建立列表 *L*;

(3) 创建根节点 root *y*;

(4) 对于会话中的每一项,执行下述操作:

Insert(*x* (*name*, *value*), *y*)

该操作以 *y* 为根节点,将事件 *x* 节点插入,*name* 和 *value*

对应于事件 x 的名称和值;

(5) Insert(T_x, x):调用递归算法以节点 x 为根节点,将以节点 x 为根的子树插入。

利用 X-DIM 模型收集数据得到的数据库表,执行算法 GenSession:

输入:关系数据表 X-DIMSUBSCRIPTION、X-DISUBSCRIPTIONHISTORY;
输出:会话列表 SessionList。
方法:

```
executeQUERY(select b. SUBSID, a. MID, b. COMMODITYID, AMOUNT, HOSTURL, a. ACCESSTIME, ISSUCCEED
from X-DIMSUBSCRIPTION a LEFT OUTER JOIN X-DISUBSCRIPTIONHISTORY b
ON a. MID = b. MID orderby a. ACCESSTIME desc, b. SUBSID desc)
```

由算法生成的会话列表每一项都是简单事件,无需拆分,避免了对不同路径上节点的关联模式进行毫无意义的挖掘,不但简化了 FP-tree 关联规则挖掘算法,而且大大降低了搜索开销。

分析上面的例子,$u1$ 浏览网页 $p1$ 并访问商品 001,则添加子树 $x(001)$ 到 $p1$ 下,子树 x 的叶节点为($x1$(Addto-Cart,20))自动插入到节点 x 下,更改数量等操作只需要重复叶节点加入。由以上算法,我们可以分析产生例子模式集{(Browse, $p1$), (Deal,001)},发现浏览与对商品操作的多维关系。这在原来基于 Web 日志的挖掘算法是很难做到的。

特别地,用户对网页的操作在不断增加,用户交易不断变更,将经常变动的部分从大量用户访问数据抽取出来进行单独分析,可以反映用户访问模式的整体变化。利用用户访问网站留下的大量历史记录,特别是将行为或特征模式具有相似性的用户群会话记录中关联部分抽取出来,计算某时间点的商品购买和浏览网页顺序变化,可以预测到用户选择商品和服务的概率和倾向。

5 结束语

X-DIM 采用 XML 技术,在应用服务器层进行用户访问数据收集,还有一些技术实施的细则需要完善:以上讨论的 XML 数据以 Session 作为根节点,在日常的数据采集中,数据分析人员常常定期从服务器导出数据。如果以天为时间单位,则日期作为根节点更恰当。在这里没有使用 XML 解析器,但当日内所有用户访问结束之前,不能写入最后的根标记。一种可行的资源集中方法是使用 File 对象的 ReadToEnd 方法将一个现存的 XML 数据的所有内容都读入一个变量中,使用空字符串替换现存的结束根标记,添加一个新用户访问记录,然后将结束根标记输出到文档末尾。每当有用户记录要添加时,都要反复运行。当然,处理该问题还有更加便捷高效的方法,是下一步关注的焦点。

总体而言,该模型封装了用户在一次会话中发生的多个操作,解决了以往数据源预处理的一系列问题,数据更精确、更合理;XML 语言的通用性和平台无关性使该模型得到的数据源可以在现有的各种平台下分析、传输和存储,克服了 Web 上数据的不兼容问题;XML 文档的结构模型与 Web 挖掘算法构造的树模型具有相似性,可以及时应用到电子商务关联分析、序列模式、聚类分析等常见的挖掘算法中。X-DIM 可以帮助企业发现潜在市场,改进商品营销策略,为电子网站商业智能化和推出个性化推荐服务提供可靠保障。

参考文献:

[1] Kohavi R, Mason L, Parekh R, et al. Lessons and Challenges from Mining Retail. E-Commerce Data [J]. Machine Learning Journal, 2004, 57(1/2): 83-113.

[2] 涂承胜,陆玉昌. Web 使用挖掘技术研究[J]. 小型微型计算机系统,2004,25(7):1177-1184.

[3] 卢正鼎,张素智. 集成 Web 数据的系统框架与实现方法[J]. 小型微型计算机系统,2003,24(10):1759-1762.

[4] Esti éenart F, Francois A, Henrard J, et al. A Tool-Supported Method to Extract Data and Schema from Web Sites [A]. Proc of the 5th IEEE Int'l Workshop on Web Site Evolution[C]. 2003.

[5] 李颖基,彭宏,郑启伦. 统一事件 Web 挖掘模型[J]. 计算机应用研究,2004,21(3):47-49.

(上接第 11 页)

处理。具体做法是,支付网关对收到的授权请求 $Request\_Auth$ 进行解密可以得到顾客的数字证书,首先确认顾客 C 的身份,其中,$Request\_Auth = SIGN\text{-}SK_M(CERT_M \| CERT_C \| EN\text{-}K_{CM}(EN\text{-}SK_C(PI)) \| EN\text{-}PK_P(PIN \| K_{CM}) \| EN\text{-}K_{CM}(H(EN\text{-}K_{MC}(m))))$;其次,支付网关还会对收到的验证信息 $Info\_Val$ 加以解密,得到交易商品信息 m_2,其中,$Info\_Val = EN\text{-}PK_P(m_2 \| K_{MC} \| SIGN\text{-}SK_C(OI) \| SIGN\text{-}SK_M(H(EN\text{-}K_{MC}(m_2))))$。在顾客 C 抵赖进行商品 m_2 交易的情况下,支付网关选择介入,让商家对商品 m_2 进行报价,由支付网关通过金融专用网完成交易双方资金的划拨,保证了交易的不可否认性。

4 结束语

A-SET 协议不仅保持了 SET 协议原有的安全特性,同时能够有效克服 SET 协议在商品原子性以及交易不可否认性方面的不足,增强了 SET 协议的安全性。基于 A-SET 协议的安全支付系统已在武汉精伦电子开发的金融终端上得到了成功的应用,该终端可以提供金融增值服务,如小额电子支付、银行卡缴费以及公交卡充值等。因此,从实践来看,A-SET 协议具备一定的应用价值和前景。

参考文献:

[1] 阙喜戎,孙锐,龚向阳,等. 信息安全原理及应用[M]. 北京:清华大学出版社,2003.

[2] 关振胜. 公钥基础设施 PKI 与认证机构 CA[M]. 北京:电子工业出版社,2002.

[3] 陈豫,陈向阳. SET 协议的分析与改进[J]. 微型机与应用,2004,6(1):34-35.

[4] 尹存燕,谢俊元. 一个公平、有效的安全电子交易协议[J]. 计算机应用研究,2002,19(1):58-63.

[5] Schneier B. 应用密码学:协议、算法与 C 源程序[M]. 吴世忠译. 北京:机械工业出版社,2000.