文章编号:1005－0523(2007)05－0117－05

面向网络营销的 IIS 日志分析系统

李明翠

(华东交通大学 信息工程学院,江西 南昌 330013)

摘要:web 日志分析对电子商务的网站推广和客户服务方面起到了非常重要的作用,文章主要阐述了一个面向网络营销的 IIS 日志分析系统的实现原理,提出了将文本文件形式的日志数据自动导入和批量导入数据库的方法,描述了日志分析系统采用的统计、聚类等分析方法的模型和实现.

关 键 词:网络营销;导入;统计;聚类

中图分类号:N945.23　　**文献标识码**:A

0 引言

随着 Web 网站的规模和范围的扩大,管理 Web 网站的工作也就变得更加复杂,它要求不只是控制通信量,而且还要观察对这些 Web 网站的外来访问,了解网站各页面的访问情况,根据各页面的点击频率来改善网页的内容和质量、提高内容的可读性,跟踪包含有商业交易的步骤以及管理 Web 网站"幕后"的数据等.为了更好地提供 WWW 服务,监控 WEB 服务器的运行情况、了解网站内容的详细访问状况就越来越显得重要和迫切了.而这些要求都可以通过对 web 服务器的日志文件的统计和分析来做到.

1 web 日志及日志分析

Web 日志是在服务器上有关 Web 访问的日志文件,这些文件里包含了大量的用户访问信息,如用户的 IP 地址、所访问的 URL、访问日期和时间、访问方法(GET 或 POST)、访问结果(成功、失败、错误)、访问的信息大小等.Web 服务器日志的格式取决于服务器系统的配置,比较常见有 NCSA 公用日志文件格式、W3C 扩展性日志文件格式以及 Microsoft IIS 日志文件格式等.下面是一个日志记录的结构表:

表 1 日志文件常用字段列表

属性域	描述	属性域	描述
date	用户请求页面的日期	cs－uri－query	请求参数
time	用户请求页面的详细时间	sc－status	协议状态
c－ip	客户端 ip 地址	sc－bytes	发送的字节数
cs－username	客户端用户名	cs－bytes	接收的字节数
s－computername	服务器的名称	time－taken	浏览耗费的时间
s－ip	服务器的 ip 地址	cs－version	客户端协议版本
cs－method	用户请求的方法	cs(user－agent)	客户端浏览器
cs－uri－stem	用户请求的页面	cs(referer)	用户从哪一页跳转过来

收稿日期:2007－06－26

基金项目:华东交通大学校立科研基金资助(课题编号:06ZKXX05)

作者简介:李明翠(1980－)女,瑶族,湖南人,华东交通大学助教,在读硕士研究生,主要研究方向:数据挖掘、web service.

web日志记录了web服务器接收、处理请求以及运行时错误等各种原始信息，较详细的记录了用户的行为信息.通过对这些信息进行归纳和分析，网站管理人员能够有效地掌握系统运行情况以及网站内容的受访问情况、加强对整个网站及其内容的维护和管理，发现用户的需求和最感兴趣的地方，找出用户行为的特征，对需求强烈的栏目提供优化以提高网站成功访问率.对于电子商务而言，通过对电子商务网站的日志分析，可以分析网站的流量模式，发现系统性能瓶颈，优化站点结构、提高站点效率，提高用户访问的有效性，发现用户的需要和兴趣[1]等，从而可以根据分析的数据制定有效的电子商务网站推广策略，发掘潜在的客户，以及针对不同的客户群开展特殊的营销等.

2 IIS日志分析系统的组成

本文设计和实现的日志分析系统如图1所示.

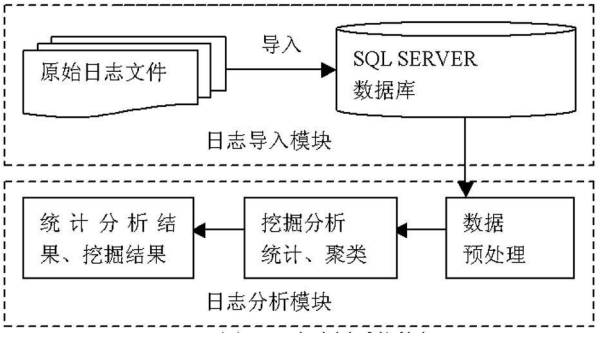

图1 日志分析系统构架

本系统主要由两个大模块组成：日志导入模块和日志分析模块.日志导入模块负责将系统自动生成的文本文件形式的日志数据导入到sql server数据库表中，以便于日志的分析.日志分析模块主要以提高网站营销为目标，通过统计分析方法对访问者数量、IP地址(地区分布)、访问时间分布、访问者操作系统、浏览器等信息进行分析，将分析结果采用直观图表显示出来；使用聚类分析[2]方法对访问站点的客户以及站点的页面进行分类；采用关联规则挖掘方法分析主要用户的访问路径；根据分析结果和不同的营销目标提出网站的调整建议，并提出相关的营销建议.

3 日志数据导入模块

日志读取模块用于实现IIS日志数据的采集功能，具体包括将文本形式的日志读取到sql server数据库表中，实现日志数据的自动定时转储到数据库的功能.

在默认情况下，日志数据存储在文本文件中.保存在文本文件中的日志数据属于半格式化数据，直接分析较为麻烦，如果能将日志数据保存在数据库表中，这将非常有利于我们的数据分析.一些操作系统也提供了ODBC日志存储形式，即将网站日志直接写入SQL数据库表中，但这种方式会耗费额外的计算机资源，速度慢，不适合访问量大的网站.另外ODBC日志格式的字段是固定的，无法保存“referrer”和“cookie”这样一些扩展字段.因此，较为理想的办法是先将日志数据写入文本文件，在必要时再将日志数据导入到数据库中.可以用SQL SERVER的数据转换服务(Data Transformation Services，DTS)实现文本数据导入数据库的功能.

具体的实现方法是：先根据文本日志文件中记录的字段在SQL SERVER中创建相关的数据库和表.现在就可以用数据库表的“导入数据”任务选项通过“DTS导入/导出向导”来手动的导入日志数据了.

3.1 自动导入

由于多数站点都是对日志文件进行按日分割(即每天产生一个文件)，日志文件是离散的，因此通过上面的方式需要进行很多次重复的手工操作才能完成某段时间日志文件的导入.本系统采取的解决方法是创建一个DTS包，使DTS包每天在服务器较空闲时自动完成前一天日志数据的导入.DTS包如图2所示：

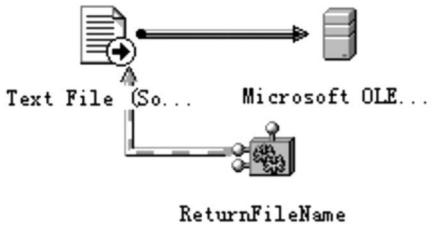

图2 自动获取前一天日志文件的DTS包

该DTS包完成从文本文件数据到SQL SERVER数据库表数据的转换.在执行转换任务前先完成动态属性任务ReturnFileName，该动态属性任务通过调用一个存储过程根据日志文件名的规律性来获取前一天的IIS日志文本文件名.创建完DTS包后用调度包任务设置该包每天运行的时间.采用该方法的特点是根据网站日志的按日分割特性，由系统自动完成文本文件形式的日志数据导入到SQL SERVER数据库表中的工作.

3.2 批量导入

日志分析系统还提供了批量导入的功能.自动导入虽然能在无人工操作的情况下实现文件的自动导入，但只能在DTS包正确运行后，每天导入一个日志文件中的数据，这就无法满足用户需要一次性导入较长一段时间内的大量日志的需求.针对这种需求，系统提供了批量导入日志数据的功能，具体的实现方法是：先将要导入的多个日志文件合并成一个文件，然后再用SQL SERVER的数据导入功能一次性将数据导入到数据库表中.合并日志文件的主要代码(c#)如下：

```
string[] dirs = Directory.GetFiles(winDir1);//获得要合并的文件
StreamWriter writer = new StreamWriter(winDir2);//建立或打开目标文件
    i=0;
    foreach(string dir in dirs)
```

```
{
    StreamReader reader=new StreamReader(dir);
    try
    {
      for(int j=0;j<4;j++)
      {reader.ReadLine();}//跳过前 4 行,日志文件的前四行是说明信息
      do
      {
        temp1=reader.ReadLine();
        temp1=temp1.Trim();
        writer.WriteLine(temp1);
      }
      while(reader.Peek()!  =-1);
      i=i+1;
    }
```

程序中主要用到了 C# 的 GetFiles 方法将文件列表读到一个标准数组中,再用 StreamReader 对象进行文件的读写.程序最后得到一个合并后的文件,然后用 SQL SERVER 强大的数据导入功能将合并后的文件中的数据一次性导入到数据库表中,这种实现方法能实现大量日志数据的导入,速度较快.

4　日志分析模块

日志分析模块的目标是:通过对 IIS 日志的分析,实现对网站客户的活动信息的分析.分析模块用统计方法对访问者数量、IP 地址(地区分布)、访问时间分布、访问者操作系统、浏览器等信息进行分析,对分析结果采用直观图表显示;用聚类挖掘方法对网站的客户和网页进行聚类,以便发现用户的偏好,给用户提供合适的页面推荐,为同一类别中的用户提供相似的服务;用关联规则挖掘方法挖掘主要用户的访问路径[3];根据分析结果和不同的营销目标提出网站的调整建议,为网站的营销策略提供依据.日志分析模块主要包括日志数据预处理、算法分析和应用分析几个子模块.

4.1　日志数据预处理

数据预处理是在统计分析或聚类分析前根据要求,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,对用户请求的页面时发生错误的记录进行适当的处理等.

在一个用户请求页面的过程中,不仅仅将该页的请求信息写入了日志,在该页中出现的图像、音频和视频的文件会随着这个网页的下载而下载,也使用了同样的请求方法写入了日志,大大增加了日志的记录数目[4].在一般情况下,我们需要进行分析处理的信息仅仅是对应的网页,而不是网页中的图像文件、音频文件和视频文件.所以在这种情况下,那种随着网页一块下载的图像、视频和音频文件对于分析而言是干扰数据,需要过滤掉.过滤算法较容易实现.只要将日志数据表中 cs－uri－stem 字段值为图像(.bmp、.jpg 等)、音频、视频等对应格式的记录过滤掉就能实现.当然要过滤的数据需要根据分析目的,视具体的情况而定.

4.2　统计分析

统计分析模块主要采用 SQL 语言的聚合函数(count()、average()等)对日志数据表的某个字段进行统计,并用 OWC 组件将统计结果用图形的方式展现给系统用户.如系统根据[c－ip]、date/time 字段对一段时间内每天或某个时间的访问客户数进行了统计,根据[cs－uri－stem]字段对页面的点击率进行了统计,根据[sc－status]字段对页面的访问状态进行了统计等等.图 3 是用户访问页面出现的错误的部分统计结果.网站管理人员可以根据统计结果对相应的页面进行改进.

页面	状态码	状态码出现次数
/KeyuMarket/firstPage/newsCenter.aspx	302	1
/KeyuMarket/firstpage/search.aspx	302	3
/KeyuMarket/firstPage/shopping.aspx	302	1
/KeyuMarket/Keyu/register.aspx	302	2
/KeyuMarket/login.aspx	302	22
/explain/jinengjianding/eclab.doc	304	1
/explain/jinengjianding/train.doc	304	1

图 3　用户访问页面错误统计

4.3　聚类算法分析

在聚类分析中,我们进行两种类型的聚类,即用户聚类和页面聚类.用户聚类主要是把主要用户划分成许多组,具有相似浏览模式的用户分为一组,将访问模式不同的用户区分开.页面聚类,则可以找出具有相关内容的网页组.可以根据这两类聚类结果向用户推荐相关的网页超链接,为用户提供上网帮助和个性化服务.

在基于日志的用户聚类分析中,通过定义一个具体的用户模型来表示用户,这个模型中包含了日志中的得到的信息,可以是对网页的访问次数、对网页的浏览时间、对网页的浏览次序以及对上面的三种的参数的一种函数值.将站点主要用户应用到该模型,然后用向量来表征该模型,这样用户的集合就转化成了向量的集合,把这些向量的集合作为聚类算法的输入,就可以得到对应的用户的聚类结果.网页聚类分析,是将网页作为主体,以从日志中得到用户对网页的访问情况做聚类的标准,为网页建立相应的数据模型,而数据模型用向量的方法来表示,然后将向量作为聚类算法的输入,从而把用户访问的类似的网页聚为一类,访问情况的不同的分在不同聚类中,用来对网页有所区分.可以根据用户对一个网页的访问,来推荐给他属于同一个聚类的另一个网页.具体实现模型如下:

$$\begin{bmatrix} x_{11} & \cdots x_{1j} \cdots x_{1n} \\ x_{i1} & \cdots x_{ij} \cdots x_{in} \\ x_{m1} & \cdots x_{mj} \cdots x_{mn} \end{bmatrix}$$

将统计分析后得到的站点主要客户进行编号,同时对站点中访问量最多前 N 个页面也进行编号,建立数据矩阵 URL-UID.矩阵的元素值 Xij 表示用户 j 在特定时间段内访问页面 i 的次数.根据列向量的划分得到用户聚类,行向量的划分得到网页聚类.由于大型网站的页面和访问用户数量都非常大,而一个用户在某个时间段内一般不会访问网站的每一

个网页，往往只访问少量网页，因此数据矩阵 URL-UID 会存在许多的零变量，因此系统还需提供处理稀疏矩阵的能力.

由于 web 日志的数量非常大，选择的聚类算法要能够较好的处理大规模数据，一个高度可伸缩性的聚类算法是日志处理的关键.本系统采用了 K—均值聚类算法.K—均值聚类算法的算法思想比较简单，实现容易，运行速度比较快[2]，是一种简单实用的无监督学习算法.该算法对海量数据的处理非常有效.K—均值聚类算法，必须在算法的开始输入一个 K 值，作为该算法在数据集上分割并计算后输出聚类的数量.本次实验数据来自本校电子商务实验中心服务器的 web 日志，部分聚类结果如下.

表 2 部分用户聚类结果

用户所属聚类编号	用户编号及对应 ip
0	0,192.168.0.6 1,192.168.0.3 15,66.249.65.70 19,172.16.15.245 22,218.64.37.115 20,202.101.27.40
1	14,124.118.191.191 29,222.80.144.136 22,202.108.1.12
2	16,162.105.146.15 17,162.105.146.11 25,218.12.194.315 26,220.181.19.76

表 3 部分页面聚类结果

页面所属聚类编号	页面编号及对应 URL
0	0,/KeyuMarket/firstPage/shopping.aspx 3,/CRM/Opportunity/OppList.asp 6,/cai/studentmenu.asp 11,/KeyuMarket/firstPage/cartList.aspx
1	10,/ecp/netsafe/index.html 24,/law/home.asp 27,/ecexam/
2	7,/teachplatform/ 17,/tiaoma/tm4/promote-engineer/index.htm 19,/java/text1—0—0.htm

聚类的结果具有较明显的分类特征，第 0 类用户大部分为校内用户且经常访问网上零售的页面，第 1 类用户较关注计算机技术方面的内容，第 2 类用户常访问与安全相关的页面.根据聚类的结果，可以对不同的客户群分别采用相应的营销策略，为其推荐相关的网页.如果推荐的网页同用户未来访问的网页同属于一个聚类，就认为推荐成功.

5 结束语

通过对 web 日志进行统计、分析，能为网站管理人员提供科学全面的决策分析图表和数据，发现和排除错误原因、了解客户访问分布等，更好的加强系统的维护和管理.

系统对象定位于 Web 的使用日志记录，从中分析出具体的用户访问模式，通过挖掘算法得出潜在的用户的使用规律，并用这些规律来进一步的指导网站的建设，辅助于商业的决策，了解潜在的商业用户的心理，挖掘潜在的商业用户，为用户提供更好的、更加人性化的 Web 服务.

参考文献：

[1] Laura Thomson. Mining User Preferences from Web Access Logs and User Public Information [R]. The Eleventh Australasian World Wide Web Conference, 2005. http://ausweb.scu.edu.au/aw05/papers/refereed/thomson/paper.html.

[2] 毛国军，段立娟. 数据挖掘原理与算法[M]. 北京：清华大学出版社，2005.

[3] 王书舟，高中文. Web 使用挖掘在电子商务中的应用[J]. 微机发展，2003,(2):41—43.

[4] 肖立英，李建华，谭立球. WEB 日志挖掘技术的研究与应用[J]. 计算机工程，2002,(7):276—277.

[5] 潘 登，董小社，杨麦顺. 从 Web 数据中挖掘频繁访问模式[J]. 西安交通大学学报，2002,(6):631—644.

Analysis System of IIS Log for E—Marketing

LI Ming-cui

(School of Information Engineering, East China Jiaotong University Nanchang 330013, China)

Abstract: Web log analysis plays an important role in E—Commerce website popularizing and customer service. This Paper presents the realization principle of an IIS log analysis system, puts forward the automatic and batch importing method for LOG data from text files to database. The statistic and clustering models and realization used in the system are also depicted in the paper.

Key words: E—Marketing; import; statistic; clustering