基于客户聚类的商品推荐方法的研究

王宏超,陈未如,刘 俊
(沈阳化工学院 计算机科学与技术学院,辽宁 沈阳 110142)

摘 要:文中给出了一种新的数据源的获取方法,使用Web2.0技术直接从客户浏览行为中获取需要的数据,避免了传统Web使用数据挖掘时日志数据预处理时的大量繁杂工作,减少了噪声数据,提高了数据准确性。根据所获数据建立用户-商品矩阵,计算此矩阵的欧氏距离,在此基础上使用聚类算法将客户进行聚类,根据聚类结果对新来的客户进行有目的的商品推荐,并对聚类结果进行跟踪评价。目的是为了提高电子商务网站的个性化服务。

关键词:电子商务;Web使用挖掘;商品推荐;个性化;客户聚类
中图分类号:TP311 **文献标识码**:A **文章编号**:1673-629X(2008)07-0212-03

Research of Commodity Recommendation System Based on Customer Clustering

WANG Hong-chao, CHEN Wei-ru, LIU Jun
(Dept. of Computer Science and Technology, Shenyang Institute of Chemical Technology, Shenyang 110142, China)

Abstract: In this paper, proposed a new approach to achieve the data in order to cluster the customers; use the Web2.0 technology to acquire information from the behavior of customers who visited the Web site. Therefore, can avoid the lots of heavy work of the pretreatment of Web log data. Then, build the customer-commodity matrix and figure out the distance of Euclidean, and cluster the customers by using the arithmetic that is based on commodity recommendation of clustering. Based on the clustering result, recommend commodities to the new customers in order to see their reactions, and evaluate the result of customers clustering.

Key words: electronic commerce; web use mining; commodity recommendation; personalization; customers clustering

1 传统的Web日志数据挖掘

随着电子商务近年来的不断进步和完善,商品趋于多样化而竞争趋于激烈化。要想有效保留客户、防止客户流失、提高电子商务系统的销售能力,就要提高电子商务系统的个性化,让每一个客户都觉得这个网站是为自己量身定制的。商品推荐系统是电子商务中提高个性化的有效手段,向客户提供商品的信息和建议,根据大量客户的历史购物信息,帮助客户决定购买何种商品,预测客户的未来的购买情况,模拟销售人员向客户推荐商品并完成交易的过程,使客户感受到完全个性化的服务[1]。因此文中使用了基于聚类的商品推荐算法,进行客户聚类,并根据聚类结果向客户进行商品推荐,以跟踪聚类的效果。

收稿日期:2007-10-18
基金项目:辽宁省教育科学研究项目(05L338)
作者简介:王宏超(1980-),男,河南遂平人,硕士研究生,研究方向为数据挖掘;陈未如,教授,研究方向为构件平台技术与数据挖掘;刘 俊,副教授,研究方向为网络安全、网络集成、数据挖掘。

传统的Web挖掘使用的数据源都是通过对Web日志预处理而得到的[2],对Web数据预处理是一个费力费时的工作,主要经过以下步骤:数据清洗、用户识别、会话识别、路径补全、模式分析等过程。在这一系列的处理过程中,可能遗漏关键的信息,也可能产生一些噪声数据。这些都会对进一步数据挖掘的结果产生负面的影响。文中提出了一种全新的数据源获取方法,并且结合异步的AJAX(Asynchronous JavaScript and XML)技术使这种方法变为现实。

2 使用异步技术记录客户浏览信息

文中所采用的新的数据源获取方法是直接记录Web服务器上的客户浏览和购买信息,并将这些数据信息持久化到数据库中。由于在高峰时期电子商务网站的Web服务器和数据库服务器负荷较大,如果使用以往的方法,简单地将大量数据写入后台的数据库,势必会造成服务器的反应迟钝,并且也不利于客户的浏览体验,会延长客户等待时间。文中引入Web2.0技

术,使用 AJAX 和后台的服务器进行异步通信,因为 AJAX 技术是通过 JavaScript 和服务器进行异步通信,所以客户就不必等待服务器的返回结果而可以继续浏览网页或购买商品,AJAX 技术会在客户毫无察觉的情况下在后台为我们完成大量的数据收集工作。

为收集到的数据建立相应的二维表进行存储,如客户访问日志表(CustomerAccessLog)的结构如表 1 所示。

表 1　客户访问日志表

字段名	数据类型	空否	字段说明
CustomerIP	Varchar(50)	否	访问者 IP 地址
CustomerID	Varchar(50)	是	浏览者的标识
LoginTime	Datatime	否	登录时间
LogoutTime	Datatime	否	登出时间
Service	Char(20)	是	请求服务类型
Httpprotocol	Char(10)	否	HTTP 协议版本号
RequestWay	Varchar(50)	是	如何来到本网站
ServiceStatus	Integer	是	返回的状态码
Agent	Char(20)	是	浏览器类型

当然还有很多的数据信息需要收集,如:客户浏览网页顺序表(PageOrder),客户查看和订购商品表(CommodityOrder),客户搜索关键字表(RetrieveKey),客户注册信息表(CustomerInf)等。当在数据库中建好这些表后,就可以有针对性地进行数据收集了。

当客户登录服务器之后,如何获取这些信息并将其持久化到数据库中是非常关键的。可以通过读取 Request 对象来获得部分的数据。获得数据信息的部分程序代码如下:

```
//客户登录 id
String CustomerId = request.getParameter("CustomerId");
//输入的路径
String contextPath = request.getContextPath();
//客户的 IP 地址
String remoteAddr = request.getRemoteAddr();
//请求的 URL
String requestURI = request.getRequestURI();
//Session 编号
String sessionId = session.getId();
//客户登录时间
SimpleDateFormat df = new SimpleDateFormat ("MMddhh-
mmss");
df=new SimpleDateFormat("yyyy-MM-dd hh:mm:ss");
DataTime logTime = df.format(new Date());
```

然后创建 XMLHttpRequest 对象 xmlHttp,通过 xmlHttp.open("POST",url,true)方法将数据以异步方式提交给服务器,完成客户数据信息的收集工作。

3　使用基于聚类的商品推荐方法

3.1　购买商品的客户聚类

聚类分析是电子商务中很重要的一个方面,通过分组聚类出具有相似浏览或购买行为的客户,并分析客户的共同特征,更好地帮助电子商务的企业了解自己的客户,向客户提供更合适更全面的服务,便于开发和执行未来的市场战略。这种市场战略包括:自动地给一个特定的客户聚类群体发送特制销售邮件,为一个客户聚类群体动态地改变一个特定的站点等。

聚类分析分为对客户群体的聚类和 Web 页面的聚类。其中客户群体的聚类在电子商务和客户提供个性化服务的应用中起着重要的作用。根据现有客户的历史行为对客户进行聚类,得到聚类的结果模型,抽象出一个特殊客户代表这个客户聚类群体的特征。对于一个新来的客户运用这个模型将新客户归入相应的类别中,根据这个类别的特征有目的地为这个新来的客户进行商品推荐,并跟踪推荐效果。

3.2　聚类算法

聚类算法[3]是基于一定的距离尺度将相互距离很近的客户进行聚类。客户经过聚类后被归为若干个不同的类。聚类后的客户对象具有最大的类内相似性和最小的类间相似性的特征。在许多应用中,可以将一个簇中的数据对象作为一个整体来对待。通过聚类,可以识别密集和稀疏的区域,因而可以发现全局的分布模式,以及数据属性之间有趣的相互关系。按客户对站点的访问行为产生聚类,具有相似浏览或购买行为的客户将会聚成一类。

3.3　构建 Customer—Commodity 矩阵

个性化页面的推荐需要客户提供评价来确定客户的兴趣,在 Web 服务器的环境下,这种客户兴趣度的评价可以通过 Web 使用挖掘自动获得。有两个指标可以用来衡量客户的访问兴趣度:一个是客户访问某个页面的频率,即访问该页面的次数;通常认为,可能是网站的拓扑结构造成了客户频繁访问某个页面,所以这并不足以反映客户对该页面中商品的兴趣度;另一个是客户购买某一类商品的个数,如果客户经常购买某一商品或某一类商品,就说明客户对该类商品感兴趣,这一类商品购买的越多,那么他对该类商品的兴趣度就越高,这里采用第二个指标来衡量客户的兴趣度。

设网站共有 n 类商品构成集合:Commodity = $\{C_1,C_2,\cdots,C_n\}$,由客户查看和订购商品表(CommodityOrder)可以得到客户的购买信息,设有 m 个客户,则构成客户集合 Customer = $\{U_1,U_2,\cdots,U_m\}$,客户购买行为被映射成为二维的向量,所以客户

访问集合 T 可以用一个 $m \times n$ 的矩阵表示，其中每行表示客户购买商品的集合，每列表示购买该商品的客户集合，每个矩阵元素项 t_{ij} 表示 客户 U_i 对商品 C_j 的购买次数(单位可以自己定义，这里为个数)，即客户对该类商品的购买兴趣度大小，当 $t_{ij} = 0$ 时，表示客户 U_i 没有购买商品 C_j。

文中所使用的数据来源于大集体电子商务平台(http://www.dajiti.com)，大集体电子商务平台是国内著名的专业 B2C 电子商务平台，其日平均点击率在 2 万次左右。抽取部分数据作为基于聚类的推荐算法的实验所用，在这次实验中抽取 2007 年 4 月份的 6 个客户购买商品的记录，由于个人隐私文中隐去了这些客户的姓名和所购商品的名称，用 U_1 到 U_6 和 C_1 到 C_6 来进行代替。表 2 是这些客户的购买商品的记录。

表 2 客户购买商品记录

客户	C_1	C_2	C_3	C_4	C_5	C_6
U_1	1	1	0	0	0	0
U_2	1	1	0	0	1	0
U_3	0	0	1	1	0	0
U_4	0	0	1	1	1	0
U_5	0	0	0	0	1	1

由上面的客户购买记录可以得到如下的 Customer - Commodity 矩阵 $T_{5\times6}$：

$$T_{5\times6} = \begin{bmatrix} 1 & 1 & 0 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & 1 & 0 \\ 0 & 0 & 1 & 1 & 0 & 0 \\ 0 & 0 & 1 & 1 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 & 1 \end{bmatrix} \tag{1}$$

3.4 客户相似度计算

文中使用欧氏(Euclidean)距离公式来进行计算，其计算公式见式(2)：

$$S(i,j) = \left(\sum_{k=1}^{m} | T_{ik} - T_{jk} |^2\right)^{\frac{1}{2}} \tag{2}$$

当然计算距离的函数还有很多，如明考夫斯基(Minkowski)距离、曼哈顿(Manhattan)距离等，但这类算法大同小异，结果也差不多。从直观上看，属于同一类的客户事务对象在空间中应该互相靠近，而不同类的客户事务对象之间的距离会大得多，所以，距离越小，客户间的相似性越大[4]。距离定义满足以下的四条规则：

若 $S(i,j) = 0$ 则 $U_i = U_j$

对于任意 U_i、U_j，$S(i,j) \geqslant 0$

对于任意 U_i、U_j，$S(i,j) = S(j,i)$

对于任意 U_i、U_j、U_k，有 $S(i,j) \leqslant S(i,k) + S(k,j)$

使用欧氏(Euclidean)距离计算方法的计算相似度矩阵的结果见式(3)：

$$S_{5\times5} = \begin{bmatrix} 0 & 1 & 2 & \sqrt{5} & 2 \\ 1 & 0 & \sqrt{5} & 2 & \sqrt{3} \\ 2 & \sqrt{5} & 0 & 1 & 2 \\ \sqrt{5} & 2 & 1 & 0 & \sqrt{3} \\ 2 & \sqrt{3} & 2 & \sqrt{3} & 0 \end{bmatrix} \tag{3}$$

3.5 聚类客户并进行商品推荐

在得出客户相似度矩阵(距离矩阵)后，就要开始对客户进行聚类。具体的聚类方法是：确定一个距离阈值 θ，如果 $S(i,j)$ 小于这个距离 θ，那么就将第 i 个客户和所有满足这个条件的第 j 个客户划分为一类客户，得到客户的聚类[5]。距离 θ 的计算公式如下：

$$\theta = \sum_{i=1}^{n}\sum_{j=1}^{m} d_{i,j} / [n \times (n-1)] \tag{4}$$

对于 $\forall d_{i,j} \in S_{n\times n(1\leq i\leq n, 1\leq j\leq n)}$，所以计算得到 $\theta = 1.79$，$\theta > \sqrt{3}$。最后可以得到客户聚类的结果为以下两种情况：

CustomerCluster = $\{(U_1、U_2、U_5),(U_3、U_4)\}$

CustomerCluster = $\{(U_1、U_2),(U_3、U_4、U_5)\}$

上面的客户聚类结果出现了交叉的情况，当然这跟 θ 的计算公式是有关的，而且 θ 的值也是可以根据经验和反复的实验来确定的，θ 的选择是一个需要权衡的问题。如果 θ 的选择过小，则聚类过细，网站进行个性化困难；如果 θ 的选择过大，则聚类过于笼统不利于商品的准确推荐。阈值 θ 的选择需要经过多次的系统模型试验，寻找最合适的值。在这里可以为客户 U_5 推荐商品以观察其反应，从而确定客户 U_5 是属于哪一个客户聚类的。如给客户 U_5 推荐客户 U_1、U_2 都购买的商品 C_1、C_2 和客户 U_3、U_4 都购买的商品 C_3、C_4，如果客户 U_5 购买了商品 C_1 或 C_2，则认为第一种聚类情况正确；如果客户 U_5 购买了商品 C_3 或 C_4，则认为第二种聚类情况正确。

4 结束语

当然仅凭客户购买商品的习惯去聚类客户群体，有其不足之处。通过算法聚类在一起的客户不一定就是真正的一类，而且客户也是在时时的改变，一个客户不可能永远属于一个聚类。这就需要增加聚类的实时性和维度来提高聚类结果的准确性，如在进行客户聚类时还可以加入客户的浏览习惯、客户的注册信息、客户搜索时所使用的关键字等信息，这样就提出了基于多维的客户聚类情况。

(下转第 221 页)

网络通信协议支持,处理异构系统中电子政务数据报文传送、管理以及报文转换等过程中各种需求不同的报文。

(2)路由支持:数据交换平台将接受到的文件解包后提取目的地信息,查询路由表,将文件按要求格式重新打包路由到下一级交换平台或与其连接的系统。

3.6 监控中心

监控中心由数据监控、安全支撑、错误处理、日志管理四个部分组成。

(1)数据监控:实现对数据交换流量和数据文件交换状态(创建、激活、挂起、终止等)的控制、管理、查询、统计与审计等。

(2)安全支撑:在利用防火墙、入侵检测、漏洞扫描等网络安全技术的基础上,保障交换平台的网络通讯与传输安全;在提供统一的CA身份认证前提下,控制用户的数据请求、访问权限;在提供加密、认证、数字签名等技术的前提下,实现数据的保密性、完整性和不可否认性。

(3)错误处理:使数据交换过程成为一个事务性过程,保证数据交换中数据的准确性和传输过程的完整性,提高通信的效率、健壮性和可靠性。

(4)日志管理:完成交换数据备份、记录系统异常和网络异常信息等。

4 结束语

数据交换不仅仅局限于电子政务系统方面,在其它系统的集成过程中也是经常要面临的问题。文中提出的数据交换平台方案通过数据交换拓扑网络,将各个物理位置分散的异构系统连接为一个统一的整体,不仅可以有效地解决电子政务系统中各应用中数据交换与信息共享问题,同时也适用于其他领域。通过这种方案构建的体系结构,无论是新子系统,还是旧的业务系统都有其相对独立性、灵活性,为系统进一步扩展提供良好的扩展性。

参考文献:

[1] 王　琰,徐　玲.电子政务理论与实务[M].北京:清华大学出版社,2004:2-17.

[2] 向　真,吴秋云,陈　荦.电子政务三网模式下的数据交换[J].计算机工程与科学,2004,26(8):11-13.

[3] 何国辉.基于XML的电子政务系统设计[J].微计算机信息,2006,22(3):151-154.

[4] 张　繁,蔡家楣.电子政务系统中的数据交换和共享服务平台设计[J].计算机工程与应用,2003,39(7):226-229.

[5] 梁　娟,熊桂喜,李　静.电子政务中基于XML的关系异构交换技术[J].计算机与数学工程,2006,34(12):60-63.

(上接第214页)

参考文献:

[1] Eirinaki M, Vazirgiannis M. Web Mining for Web Personalization[J]. ACM Transactions on Internate Technology, 2003, 3(1):1-27.

[2] Baglioni M, Ferrara U, Romei A. Preprocessing and Mining Web Log Data for Web Personalization[J]. Advances in Artificial Intelligence, 2003(2):237-249.

[3] 杨小兵.聚类分析中若干关键技术的研究[D].杭州:浙江大学,2005.

[4] 吕亚兵.WEB站点日志数据挖掘的研究与实现[D].武汉:武汉理工大学,2006.

[5] 梁　伟,张慧颖.电子商务推荐系统中推荐模型的研究[J].计算机工程与应用,2006,42(36):183-186.

(上接第217页)

3 结束语

CCM构件是一个自包含的构件实体。文中对CORBA构件模型和基于CCM构件组装技术进行了研究,并在PLM中供应链系统(SCM)的销售管理子系统中应用。基于构件的开发遵循构件开发-应用组装-应用部署的过程,同时下一步要考虑系统的演化,以及自动组装过程的研究。

参考文献:

[1] 韩跃科.基于Java/CORBA的政务信息发布系统研究[J].计算机技术与发展,2007,17(6):203-205.

[2] Wallnau K C, Carney D. Building systems from commercial components[M]. [s.l.]:Addison Wesley,2002.

[3] 张　驰.基于接口匹配的构件组装[J].计算机应用,2007,27(6):1420-1422.

[4] 李　睿,徐　红,曾应员.构件化技术在学生成绩查询统计系统中的应用[J].计算机技术与发展,2007,17(6):214-216.

[5] 徐东升,张　驰.CORBA构件接口扩展技术与描述[J].微电子学与计算机,2007,24(4):12-14.

[6] 任洪敏,钱乐秋.构件组装及其形式化推导研究[J].软件学报,2003,14(6):1066-1074.