序列模式挖掘在电子商务个性化服务中的应用

靳明霞，李玉华，管建军
（华中科技大学 计算机学院，湖北 武汉 430074）

摘 要：分析了电子商务发展面临的问题和个性化服务的特点，提出了Web使用挖掘技术在电子商务个性化服务中的应用方法，论述了基于Web挖掘的个性化服务研究。详细阐述了其挖掘过程，最后讨论了使用序列模式和分类相结合的技术得以实现个性化服务的方法。利用这些算法得到的个性化信息可以准确把握用户兴趣模式并对Web信息资源的组织方式进行有效更新，从而提高网络信息服务效率，为用户提供“一对一”的具备自适应性的智能个性化服务。

关键词：Web挖掘；序列模式；电子商务；个性化服务
中图分类号：TP39 **文献标识码**：A **文章编号**：1673-629X(2006)10-0233-04

Application of Sequential Patterns Mining Project on Electronic Commerce Personal Information Service

JIN Ming-xia, LI Yu-hua, GUAN Jian-jun
(Computer Coll., Huazhong Univ. of Sci. and Techn., Wuhan 430074, China)

Abstract: In this paper, analyze the exist problem on electronic commerce and the characteristics of personal information service. Then present Web usage mining on electronic commerce personal information service concrete usage in this domain. Discuss a method of sequential patterns combining with data classification mining to be used in the procedure. With the individual information obtained by these algorithms, can master the customer interests model precisely and change the framework of Web information resources so as to improve the efficiency of network information supply dramatically.

Key words: Web usage mining; sequential patterns; electronic commerce; personal information service

0 引 言

目前，电子商务的竞争比传统的商务竞争更加激烈。因此，若想在竞争中生存进而获胜，信息服务方式就必须从传统的“一对多”发展到“一对一”的个性化用户服务方式。随着电子商务中引入个性化用户服务方式，企业需要对Web环境下的客户资料数据进行深入的统计与分析，找出不同用户兴趣所在，透视隐藏在这些数据之后的更重要的用户兴趣模式信息以及关于这些数据的整体特征的描述并预测发展趋势等。序列模式挖掘的个性化服务，即是利用个性化技术将传统的数据挖掘(Data Mining)对象同Web访问信息结合起来，利用Web使用挖掘的方法抽取用户感兴趣的潜在有用模式与信息，然后基于这些模式和信息为用户提供“一对一”的具备自适应性的智能个性化服务。这些智能个性化服务可大大缩短用户在网络上的访问延迟，使得提供给用户的网络信息服务质量得到最大程度的提高。

收稿日期：2006-03-01

作者简介：靳明霞(1966-)，女，河南延津人，硕士研究生，高级讲师，研究方向为电子商务、数据挖掘；李玉华，副教授，硕士生导师，研究方向为计算机网络及应用、人工智能。

1 电子商务中的个性化服务及Web挖掘

1.1 电子商务发展面临的问题

随着因特网的发展和普及，越来越多的用户利用搜索引擎来搜索网上信息。尽管搜索引擎的发展已较成熟，但人们在使用中却发现要准确、快速地查找自己所需的信息越来越困难[1]。主要原因有两个：

(1) 一次搜索的检索结果(一系列URL地址)可能有成千上万条，而在这过于庞大的信息群中，有用信息只是其中的一小部分，并且常常发生收到或下载的信息难以消化的情况，即所谓的“认知过载”。

(2)目前的搜索引擎都是服务器端软件，用户需要严格按照各种引擎所要求的格式输入查询词，但种种限制使用户不知道如何贴切地表达自己的信息需求，也不知道如何更准确地寻找所需信息，即所谓的“迷航”。

1.2 个性化服务的特点和功能

为了适应用户不断增长的信息需求，研究人员纷纷从人工智能中寻找突破口。在许多探索性研究中，人们寻求一种将用户感兴趣的信息主动推荐给用户的服务方式，这便是个性化信息服务。个性化主动信息服务(Personalized Active Information Service, PAIS)[1]作为一种崭新的智能信息服务方式，应用前景广阔，十分引人注目[2]。个性化

信息服务是在对用户及其需求了解的情况下,即通过用户研究,从数量庞大、增长迅速、类型复杂的网络信息中提取出用户真正需要的那一小部分提交给用户,是以"用户为中心"的服务原则在网络环境下的具体体现。它有以下功能:

(1)记忆型。记忆型通过在系统中记录使用者的信息,当使用者再次登录该网站时,系统利用用户过去的历史数据,给用户必要的提示和帮助。具体功能包括:向登录用户致意;为用户建立个性化书签和分配用户个性化的存取权限等。

(2)引导型。引导型是指系统通过提供替代的浏览选项,协助引导使用者更快更容易地获取所寻求的信息。这类个性化服务不但能增加使用者的忠诚度,而且可以减轻用户在大型网站里所面临的"数据超载"和"信息迷航"问题。功能包括:向用户进行超链接的推荐;为用户导航。

(3)定制服务型。这类系统可以按照用户的知识、兴趣和偏好对网页的内容、结构进行个性化设定,达到对数据负荷进行管理,使用户和网站的交互简单化和个性化。具体功能包括:个性化的网站布局设计;个性化的内容定制;个性化的超链接定制;个性化的定价和营销。

(4)工作任务辅助支持型。这类系统能按照用户特点,致力执行特殊的动作程序,给用户的工作辅助以帮助和支持。这是最先进的个性化功能,可以在客户端或服务器端实现。具体功能包括:个性化的行动助理;个性化的疑问解答和个性化的谈判助手。

1.3 实现个性化服务的 Web 数据挖掘

数据挖掘(Data Mining)是从大量的、不完全的、有噪声的、模糊的和随机的数据中提取人们事先不知道的,但又是潜在有用的信息,发现其后隐含的规律性,将其模型化,完成辅助决策作用的过程[3]。

Web 挖掘就是将数据挖掘的技术应用到 Web 数据上,以发现客户浏览模式规律的过程。可以运用关联、分类、聚类等技术手段,从中提取出可以指导市场策略的有用数据,通过收集、分析和处理从网上获取的有关消费者消费行为的数据,确定特定消费群体或个体的消费习惯、爱好、倾向,进而预测出消费者下一步的消费行为,有针对性的提供服务。

Web 挖掘(Web Mining)从性质上可分为 3 类:Web 内容挖掘、Web 结构挖掘及 Web 使用挖掘(见图 1)。Web 内容挖掘是从文档内容或其描述中抽取内容,分为文本挖掘(包括 text,HTML,XML 等格式)和多媒体挖掘(包括 image, audio, video 等媒体类型);Web 结构挖掘是从 WWW 的组织结构和链接关系中推导知识;Web 使用挖掘是从 Web 的访问记录中抽取感兴趣的模式。面向电子商务的 Web 挖掘主要包括 Web 内容挖掘和 Web 使用挖掘[4]。通过 Web 内容挖掘,可进行电子商务海量商品信息采集;通过 Web 使用挖掘,可辅助商家理解用户行为,从而改进站点结构,调整销售策略,提供个性化服务。

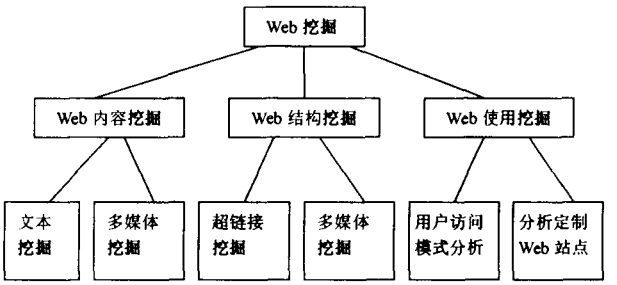

图 1 Web 挖掘分类

2 Web 使用记录的挖掘过程

2.1 数据准备

数据准备包括数据采集和数据预处理。

数据采集所采集的对象为 Web 挖掘的源数据,包括:Web 服务器日志、客户登记信息和交易数据库[5]。Web 服务器日志是由客户访问所产生的服务器日志数据,可分为 Server logs, Error logs 和 Cookie logs。

进行数据挖掘的时候,一般并不是对原有的数据进行挖掘,而是先要对数据进行预处理。日志文件处理得好坏直接影响挖掘算法产生的结果,其处理过程是保证 Web 使用挖掘质量的关键。日志文件的处理包括以下几个方面:数据清洗、用户确定、用户访问序列确定、用户会话确定、访问路径完善等。目的是合并数据,将多个文件或多个数据库中的数据进行合并处理,选择数据,提取数据,清洗过滤,剔除一些无关记录,将文件、图形、图像及多媒体等文件转换成可便于数据挖掘的格式等。

2.2 数据挖掘模式

经过清洗和预处理,便可选择合适的数据挖掘方法和数据挖掘算法,从 Web 提供的信息数据库中发现一些隐藏的知识和信息。常用的挖掘方法有路径选择、关联分析、分类规则、聚类分析、序列分析等[6]。

(1)访问路径分析。访问路径是用户在网络上浏览时从一个网页到另一网页的路径。访问路径分析就是寻找频繁访问路径的方法,即通过对 Web 服务器日志文件中客户访问站点的访问次数分析,挖掘出频繁访问路径。

(2)关联规则发现。目的就是为了挖掘出藏在数据间的相互关系。在电子商务中关联规则的发现就是找到客户以网站上各种文件之间访问的相互联系。通过关联规则,可以发现如果客户购买了某一种商品,那么他也可能会购买另一种商品。这有利于商家更好地组织站点,为顾客减少过滤信息的负担。

(3)分类规则的发现。分类分析法的输入集是一组记录集合的几种标记。首先为第一个记录赋予一个标记,即按标记分类记录,然后检查这些标定的记录,描述出这些记录的特征。在电子商务中通过数据挖掘对客户进行细分,以便提供个性化服务。

(4)聚类分析。在电子商务的数据挖掘中,对日志的聚类分析是一种很好的方法。聚类分析与分类分析不同,

它的输入集是一组未标定的记录。聚类分为对客户群体的分类和 Web 页面的聚类。其中客户群体的聚类在电子商务和用户提供个性化服务中起着重要的作用。通过分组聚类出具有相似浏览行为的客户,并分析客户的共同特征,更好地帮助电子商务的用户了解自己的客户,向客户提供更适合的用户。

(5)序列模式的发现。其测重点在于分析数据间的前后的因果关系。就是在时间有序的事务集中,找到那些"一些项跟随另一项"的内部事务模式。发现序列模式能够便于电子商务的组织者预测客户的访问模式,对客户提供个性化服务,网站的管理员可将访问者按浏览模式分类,在页面上只展示该浏览模式的访问者经常访问的链接,而用"更多内容……"指向其它未展示的内容。

3　应用在个性化推荐中的序列模式

3.1　用户访问模式的分析

可以把网站看成相互链接的文档。因此,某些结点被多次访问可能是因为它所在的位置而不是因为它的内容[7]。例如,用户在访问一个网页的时候,经常会回退(backward)之后再选择新的网页。此时的回退页面被多次访问,就是因为它所在的位置而不是它的内容,这种回退是无意义的。因此,为了找到有意义的用户访问模式,假设回退都是为了点击别的页面而不是为了浏览,因此只关注前向(forward)模式。特别是当同一用户回退到已经浏览过的网页,此时前向浏览终止,这一前向参考路径(forward reference path)就成为最大前向参考(maximal forward reference)。当得到一条最大前向参考后,回退到前一个结点,寻找新的最大前向参考。

在获得的最大前向参考集合中,具有较高出现频率的一些最大前向参考,称之为热点访问模式(frequent traversal patterns),或者最大参考序列(large reference sequences)。最大参考序列的求解,类似于最大集合中关联规则的发现,唯一的不同在于最大参考序列是有序的,而最大集合仅仅是项的集合。在此所讨论的 Web 使用记录挖掘侧重于序列模式的挖掘,具体应用了最大前向参考算法和完全扫描算法来实现。

最大前向参考算法关注的是作为独立个体的用户的访问模式,对于这类模式的分析能够有效地帮助企业建立"一对一"的服务模式,为用户提供个性化的服务。虽然企业能够为每一个用户建立"一对一"的服务模式,但是企业的这种投资是否是有效的投资?客户对企业的价值差别很大,有的客户比别的客户更有利可图,有的客户目前无利可图,有的客户目前和将来都无利可图。因此,企业所提供的"一对一"的服务应该建立在对客户的分类基础之上,即根据客户的价值提供与之相对应的服务。为此,应用完全扫描算法,求解用户作为一个群体的访问模式,并根据企业的策略选取其感兴趣的访问模式,来支持电子商务环境下的客户关系管理的客户分类。针对不同类型的客户,便可以应用不同的营销策略和提供不同的服务水平。

3.2　最大前向引用参考算法

在日志数据库中,包含了路径的起点(source)和终点(destination)。对于一个新的路径,如果没有起点,就认为起点为空。最大前向算法的思路是,根据用户 id 检索日志数据库,查找该用户的访问路径集合{(s1,d1),(s2,d2),…,(sn,dn)},依时间顺序排列,然后应用最大前向参考算法找到最大前向参考。设 DF 为存储所得的最大前向参考的数据库。该算法步骤如下:

Step1 : 初始化。设 i=1,路径 Y 为空。其中路径 Y 用来储存当前的前向参考路径。同时,将标志位 F 设为 1,表示前向路径。

Step2: 令 A=s1,B=d1.

If A is equal to null then

/* 这是一条新的路径 */

Begin

Write out the current string Y (if not null)to the database DF;

Set string Y = B ;

Go to Step5.

End

Step3:

If B is equal to some reference (say the j - th reference) in string Y then

/* 出现了与以前的参考重复的参考路径 */

Begin

If F is equal to 1 then write out string Y to database D_F;

Discard all the reference s after the j - th one in string Y;

F=0;

Go to Step5.

End

Step4: Otherwise,append B to the end of string Y.

/* 仍然在前向路径上 */

If F is equal to 0,set F =1.

Step5: Set i=i+1. 如果序列还没有扫描完毕,转到 Step2。

例如,对于如图 2 所示的路径,根据最大前向参考算法,可以得到前向参考路径如表 1 所示,其中 Y 表示路径,D_F 为最大前向参考数据库。

表 1　图 2 的最大访问路径

标号	Y	D_F 输出	标号	Y	D_F 输出
1	AB	-	9	ABEF	ABEFG
2	ABC	-	10	ABEFH	-
3	ABCD	-	11	A	ABEFH
4	ABC	ABCD	12	AI	-
5	AB	-	13	AIJ	-
6	ABE	-	14	AI	AIJ
7	ABEF	-	15	AIK	AIK
8	ABEFG	-			

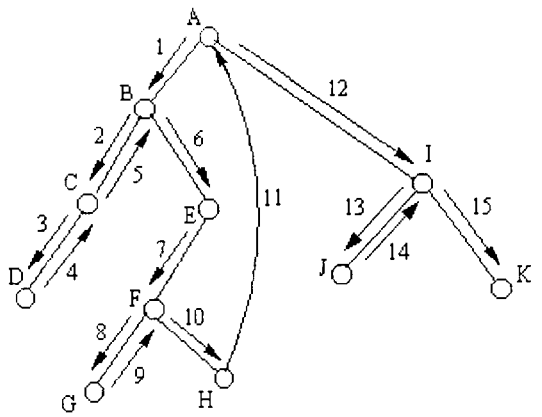

图 2 一个用户访问路径图

此例的分析如下,用户的访问路径的时序集合为:{(A,B),(B,C),(C,D),(D,C),(C,B),(B,E),(E,F),(F,G),(G,F),(F,H),(H,A),(A,I),(I,J),(J,I),(I,K)}。当第四步从 D 到 C 出现第一次回退,产生第一个最大前向参考 ABCD。因为继续回退,就从上一层(AB)开始寻找新的最大前向参考。可以看到,第九步回退产生第二个最大前向参考 ABEFG。以此类推,可以找到这个用户的所有前向参考 ABEFH, AIJ, AIK。

最大前向参考算法,过滤了回退所代表的客户访问模式,集中讨论前行访问模式来挖掘有趣的访问模式。

获得了所有用户的最大前向参考之后,就可以使用完全扫描算法(FS 算法)求解最大参考序列。如基于 DHP 的完全扫描法,每一次用 FS 算法对数据库的扫描和整理,都会减小下一次的扫描量,提高了扫描的效率。在此主要讨论最大前向参考路径的算法。

4 模式分析

对挖掘出的知识模型进行分析解释,讨论比较评价指标,从而推荐一个高效的营销管理模式,推动电子商务向个性化和智能化发展。

当前,有些系统为获得最高的正确率而不断进行优化,而实际上,市场的变化是如此之快,任何预测模型都不可能一直是正确的。也许,当数据本身因为数据的复制和转移而被破坏时,系统的设计者却在为提高一点利润而考虑系统的正确率,也可能商业模型定义的不好,尽管模型的预测能力强,但预测的商业方向不对。为了避免上述情况的发生,全面地衡量 Web 挖掘,应该从以下 3 个方面来考虑:

(1)正确率。Web 挖掘工具产生的模型必须尽可能正确。但要认识到,不同技术之间正确率的差别可能是因为随机取样的方法不同造成的,也可能是应用模型的市场本身隐藏的某些动态特征造成的。

(2)可解释性。Web 挖掘工具必须能够清楚地向最终用户解释模型是如何发现知识的,并且这些知识和常识应该很容易被测试和确证。它还要能够用清晰的方式解释利润或投资收回率的计算。

(3)集成度。Web 挖掘工具必须与当前的商业过程以及公司内部的数据和信息流相结合。如果要对数据进行复制和大量的预处理,产生错误的地方可能就比较多。如果各个过程紧密地结合在一起,产生错误的可能性就小得多。当满足这 3 个指标时,Web 挖掘工具产生的获取高额利润的模型就可能在比较长的一段时间内保持稳定。

5 结束语

探讨了电子商务中的 Web 挖掘技术,并具体使用了序列模式加分类模式。所提出的算法可使得 Web 信息服务提供者,根据用户网络浏览行为正确把握其兴趣所在并可动态地对其行为进行预测,根据这些个性化信息调整 Web 信息资源的组织方式,最大效率地为用户提供方便快捷且实用的个性化服务,不仅实现了"一对一"的个性化电子商务服务,同时也利用分类对客户群体进行了划分,从而照顾了商家的利益,使得数据挖掘切实起到了扩大营销、提供销售策略的目的。

参考文献:

[1] 方美琪. 电子商务概论[M]. 北京: 清华大学出版社, 1999. 86-107.

[2] Balabanovic M. An Adaptive Web Page Recommendation Servide[A]. In: 1st International Conference on Autonomous Agents[C]. Marina del Rey: [s. n.], 1997.

[3] 韩家炜, 孟小峰. Web 挖掘研究[J]. 计算机研究与发展, 2001,38(4):405-141.

[4] 王书舟,高中文. Web 使用挖掘技术在电子商务中的应用[J]. 微机发展,2003,13(12):41-43.

[5] 邢东山,沈钧毅. Web 使用挖掘的数据采集[J]. 计算机工程,2002,28(1):39-42.

[6] 赵立江,何钦铭. 一个个性化 Web 推荐系统的研究与实现[J]. 武汉理工大学学报,2004,28(5): 681-684.

[7] 石 晶,龚震宇,裘杭萍. 基于 Web 使用挖掘的个性化服务系统[J]. 电子科技大学学报,2002,31(4):400-403.

(上接第 232 页)

及实践[J]. 微机发展,2005,15(9):13-15.

[2] 康天增. 神经网络的原理和应用[J]. 机电设备,1996(5):33-36.

[3] 陈建宏. 矿山工程界线的光滑方法与自动追踪算法研究[J]. 湖南科技大学学报(自然科学版),2004(12):6-7.

[4] 尼尔森. 人工智能[M]. 郑和根译. 北京:机械工业出版社, 2003.22-24.

[5] 黎新懿,赵景亮. 用 Visual LISP 开发 AutoCAD2004 应用程序[M]. 北京:科学出版社,2003.