基于 Web 挖掘的客户行为识别和推荐方法研究

刘伟江

（吉林大学 商学院，吉林 长春 130012）

摘要：电子商务的普及和发展在给用户带来大量信息的同时也使用户陷入信息过载的困境，因此如何抓住用户访问网站的机会，使访问者迅速找到自己所需要的产品，并最终使访问者变成满意的消费者是很多网站追求的目标。文章从易诚手机网站上 2008 年 6 月 1 日至 2008 年 8 月 31 日三个月的浏览日志的数据出发，首先利用 Web Log Explorer 软件找出顾客浏览产品的完整路径，然后把页面中的产品信息与用户访问网站过程中所形成的路径信息结合起来，通过关联规则方法总结出顾客的特点，以此为基准结合对新顾客路径信息的识别向新顾客推荐产品，从而使对新顾客的产品推荐更有针对性。期望这种方法能在提高浏览者的满意度的同时增强网站的赢利能力，从而达到双赢的效果。

关键词：Web 挖掘；顾客行为识别；推荐系统

中图分类号：C939　**文献标识码**：A　**文章编号**：0257-0246（2009）12-0078-04

引　言

电子商务以其方便快捷、不受时间和空间限制等诸多优点逐渐在全球流行起来。随着电子商务规模的不断扩大，电子商务网站上商品的数量日益增多，而且品种日新月异，同时访问电子商务网站的人数也快速增长。电子商务网站为用户提供越来越多选择的同时，其结构也变得更加复杂；同时，用户常常会迷失在大量的商品信息空间中，无法顺利找到自己需要的商品。① 那么，如何在电子商务网站上有效地提高商品的吸引力，提高用户对网站的忠诚度，从而使企业获得尽可能好的效益？如何使用户面对电子商务网站提供的上万甚至上百万种商品时，更好地选择到自己满意的商品？这就是所谓的信息过载问题，它已经成为现代电子商务所面临的主要问题之一。尽管搜索引擎可以让用户完全个性化地选择关键字，但返回的信息对于每一个用户仍然是相同的，也就是表达了主流的兴趣，倘若对某关键词或者关键词组，用户的兴趣和主流兴趣不一致，那么很难从返回的搜索结果中找到自己感兴趣的内容，而推荐系统② 是能分析以前的使用行为并为解决新情况提供推荐的信息系统，它可以根据用户的偏好来向用户推荐相应信息，因此它是一种比搜索引擎更进一步地提供个性化服务的智能代理系统，它能从 Internet 的大量信息中向用户自动推荐符合其兴趣偏好或需求的资源。推荐系统这种能够帮助用户迅速快捷地获取所需要的信息的特性，在提高用户满意度、增强客户价值的同时，也会

基金项目：吉林大学“985 工程”“经济分析与预测哲学社会科学创新基地”项目。

作者简介：刘伟江，吉林大学商学院副教授，研究方向：电子商务。

① H. Berghel, “Cyberspace 2000: Dealing with Information Overload,” *Communications of the ACM*, Vol. 40, No. 2, 1997, pp. 19-24.

② Ting-Peng Liang, “Recommendation Systems for Decision Support: An Editorial Introduction,” *Decision Support Systems*, 2008 (45), pp. 385-386.

给电子商务网站带来丰厚的商业利益。正因为如此，对为决策支持提供服务的推荐系统的研究越来越引起人们的关注。目前关于推荐系统的研究主要包括三个方面的内容：以人口统计特性为基础的过滤、协同过滤和以内容为基础的过滤。① 以人口统计特性为基础的过滤是分析用户的偏好并根据用户的人口统计特性如性别、年龄、收入等来作推荐；协同过滤是确定用户的偏好并根据用户的类型来进行推荐。也就是说，对一个用户的行为的预测是通过与其有类似行为的用户的行为推断出来的；以内容为基础的推荐是根据产品的特定属性来进行的推荐等。Wang 等人通过使用历史导航记录开发一个模型来支持智能网站导航。② letizia 设计了一个客户方代理，通过观察客户过去的访问行为来预测该用户下一次要访问的页面，③ 这种方法的一个限制就是没有利用大多数用户的信息而只利用一个客户的信息来预测他未来的行为。Borges and Levene 把日志数据当做一个带有权重的图，权重代表用户与网站交互的可能性，然后通过关联规则从导航图中提取导航模型。④ 他们的工作都是把网站中的每个页面当作一个节点，一个页面到另一个页面的链接当作边。本文在此基础上把页面中的产品信息与用户访问网站过程中所形成的路径信息紧密结合起来，进而通过关联规则方法总结出顾客的特性，以此为基准结合对新顾客路径信息的识别向新顾客推荐产品，从而使对新顾客的产品推荐更有针对性。期望这种方法能在有助于提高浏览者的浏览质量的同时提高网站的赢利能力，从而达到双赢的效果。

基于 web 挖掘的顾客行为识别和推荐方法研究

本部分由三部分组成，即数据输入、推荐方法和数据输出。具体情形如下：

1. 数据输入

本文采用用户的浏览路径作为隐式浏览输入信息，即从易诚手机网站（www. encity. cn）上 2008 年 6 月 1 日至 2008 年 8 月 31 日三个月浏览日志的数据出发，利用 Web Log Explorer 软件进行处理，进而得到用户访问网站的路径。在所获得的全部路径中，我们去掉只包含一个页面的路径，认为那是用户偶然访问的结果，从而获得 942 个包含 2 个或 2 个以上页面的路径。

表 1 最优路径统计结果

ID	Hits	Visitors	Path
1	114	79	/ProDetail. asp? id=642 -> /ProDetail. asp? id=550 -> /ProDetail. asp? id=580
2	156	68	/ProDetail. asp? id=550 ->/ProDetail. asp? id=583 -> /ProDetail. asp? id=636 -> /ProDetail. asp? id=580 -> /ProDetail. asp? id=471
3	132	61	/ProDetail. asp? id=700 -> /ProDetail. asp? id=668 -> /ProDetail. asp? id=731 ->/ProDetail. asp? id=725
4	103	51	/ProDetail. asp? id=550 -> /ProDetail. asp? id=643 -> /ProDetail. asp? id=587
5	102	39	/ProDetail. asp? id=642 ->/ProDetail. asp? id=707

由于本文只关注产品的浏览情况，所以清理了所有中间的过渡页面，因此统计路径的结果只包含产品信息页面，这样有助于以后分析消费者的特性。

① M. J. Pazzani, "A Framework for Collaborative, Content-based and Demographic Filtering," *Artificial Intelligence Review*, Vol. 13, No. 5-6, 1999, pp. 393-408.

② Y. Wang, W. Dai, Y. Yuan, "Website Browsing Aid: A Navigation Graph-based Recommendation System," *Decision Support Systems* Vol. 45, No. 3, 2008, pp. 387-400.

③ H. Lieberman, Letizia, "An Agent that Assists Web Browsing," Intl. joint Conf. on Artificial Intelligence, proc. 14th, 1995.

④ H. Berghel, "Cyberspace 2000: Dealing with Information Overload," *Communications of the ACM* Vol. 40, No. 2, 1997, pp. 19-24.

2. 推荐方法

本模块是推荐系统的核心部分，主要是在对所采集数据的统计分析基础上，结合手机的特性，通过关联规则的方法总结顾客行为特征，为未来推荐打下基础。所采用的方法如下：按每条路径被访问的次数的多少为原则来对路径进行排序（即每个路径一共被多少IP点击过），从中选择出客户点击数最多的5条路径进行分析，结果见表1：

把路径与相应产品相对应，从而转换为对应的产品（即手机型号）浏览顺序：

表2 产品浏览顺序

ID	Visitor	Path
1	79	5310XM-U600-E848
2	68	U600-U700-G600-E848-D908
3	61	N78-N82-I908-A1600
4	51	U600-5610XM-W580
5	39	5310XM-7310

路径本来是有序的，但由于顾客的喜好不同和网站设计等原因，可能对于同一类喜欢的产品会有不同的浏览顺序，所以这里我们将所有相同产品的路径都合计在一起成为一类产品，然后从中找出产品的共性。对于产品，本文从价格、品牌、上架时间、功能和外观设计等5个方面进行考察。上述产品参数如下：

表3 产品特性

产品ID ProDetail. asp? id	型号	品牌	上架时间	价格	外观设计
550	U600	三星	2007-04-23	1405	滑盖，触摸键盘
583	U700	三星	2007-08-16	1465	滑盖，触摸键盘
642	5310	诺基亚	2007-10-30	1380	音乐手机，小巧
580	E848	三星	2007-08-16	1345	滑盖，镜面
636	G600	三星	2007-10-06	1670	滑盖
471	D908	三星	2006-08-06	1075	滑盖
700	N78	诺基亚	2008-06-08	2365	直板，智能手机，GPS
668	N82	诺基亚	2008-05-04	3145	值班，智能手机，GPS
731	I908	三星	2008-08-12	4395	智能手机
725	A1600	摩托	2008-07-12	3255	触摸屏，手写
643	5610	诺基亚	2007-10-02	1315	音乐手机，滑盖
587	W580	索爱	2007-08-16	1525	滑盖，音乐手机
707	7310	诺基亚	2008-06-15	1140	音乐手机，小巧

参照表3中的数据，我们可以很容易得出表2中路径的主要特性：

路径一：价格分别是1380、1405和1345，属于价格偏好型，价格变化幅度很小，在1400左右。

路径二：属于品牌偏好型，都是三星的产品，价格相差幅度也不大。

路径三：属于新产品偏好型，上架时间都差不多，属于新型号，都是2008年才上市的产品，而且和浏览的时间最多只间隔4个月。价格幅度相差很大，而且不限品牌。

路径四：滑盖手机、音乐手机，但这个参数很难量化。

路径五：外观小巧玲珑，没有数据量化这种外观上的相同点。

在此基础上，我们用关联规则来对路径进行分类。关联规则挖掘是数据挖掘研究中的一个重要领域，常见的基于关联规则分类算法通常在训练数据集上生成关联规则的全部集合，然后选择一个高质量的规则子集作为分类规则集去分类和预测测试数据集。关联规则的基本形式是通过以“if前件then结果”为形式的规则来实现，其中规则的前件可用合取范式 $R=(r_1 \wedge r_2 \wedge r_3 \wedge \cdots \wedge r_K)$ 表示，R也称

作规则集。

按照上述的方法，首先对所有路径进行量化，价格幅度则用标准差进行计算，大于 300 为大，小于 300 为小。上架时间距今少于 6 个月为新，否则为旧。我们得到如下规则：

r_1：（标准差＝小）→价位敏感类

r_2：（标准差＝大）∧（商品＝新）→喜新类

r_3：（标准差＝大）∧（同一品牌＝是）→品牌喜好类

r_4：（标准差＝大）∧（商品＝旧）∧（同一品牌＝否）→其他类

从中可以看出，我们将消费者分为四类：第一类是价位敏感型：关注某一价位的产品；第二类是品牌爱好型：关注某一品牌的产品；第三类是喜新型：关注最新上架的产品；还有外形和功能上的相同点，因为难以量化，所以本文将其归第四类即其他类。一般来说消费者不是单一类型的，顾客的浏览可以同时兼顾上面提到的前三种类型，但考虑到顾客在浏览时会表现出某一方面的特别喜好，所以会根据消费者这一时刻的表现将其划归到某一具体的类别中。

按上述规则将 942 个路径进行分类，从路径的分类结果中得知，大部分顾客都属于价位敏感型，这跟本网站的主要购物群体是大学生有关，通常他们会在购物时为自己设一个心理价位，并依照这个价位来挑选其中自己喜爱的产品，也就是顾客先是价位敏感，然后才是看品牌、功能或外观等。

3. 数据输出（顾客行为识别和推荐）

通过上面的分析可以得出每类顾客的路径特性，在此基础上我们就可以通过对在线访问者（新顾客）的浏览行为进行识别，确定该访问者的分类，然后针对访问者的分类特性对其进行推荐。一般来说每个价位都有很多产品，如果访问者浏览了 2 个页面，我们就可以锁定用户所能接受的价位段在哪里，然后再向访问者推荐这个价位段浏览次数最多的和购买量最多的产品，而不需要重新配合历史路径数据，然后再推荐相应的产品。以往的推荐系统都是比较离线数据和在线数据再作出推荐的，而随着客户的浏览，离线处理的结果总在变化，这样做的效率比较低。

本文的作法并不需要离线的历史数据，只要对在线的浏览行为对比分析就很容易得出客户的特性并以此作出推荐。当浏览者只浏览第一个产品的时候，并不能对该浏览者作出识别，需要做产品的关联分析。如当浏览者进入网站时，如果访问的第一个产品是 U600，那么首先将推荐 U700、E848 等这些关联度高的产品给浏览者。而当浏览者再点击 E848 的时候，就可以分析浏览者当前浏览产品的价格标准差和品牌、上架时间等情况，并对该浏览者进行识别。如判断为价位敏感型，就推荐标准差为该浏览者当前历史浏览产品的标准差以内的销售最多和被浏览最多的产品给该浏览者，按浏览者当前的历史浏览产品标准差不断调整价位段内的推荐产品。如果浏览者当前浏览的产品标准差超过 300，则按规则继续识别浏览者的类别，并在相应类别中推荐销售量最高和浏览次数最多的产品。

此外，在对在线用户进行客户行为识别和分类的基础上，可以针对不同的浏览者进行不同的营销策略，特别是针对价位敏感型的浏览者可以实行价格歧视，根据在线的识别，得出当前浏览者浏览的产品组价格标准差实行差别定价，以赠送代金卷或中奖积分等策略来让浏览者变成购买者；而对品牌喜好型和喜新型的浏览者也作出相应策略，如赠送品牌纪念品或最新的小礼物等方法来吸引这些浏览者使其变成购买者。这些方法的使用将会更有针对性地为不同的浏览者提供个性化的营销策略。

责任编辑：蔡中为