<<电子商务

CtoC电子商务站点中的Web个性化推荐技术*

李树青
南京财经大学信息工程学院信息管理系　南京 210046

〔摘要〕分析CtoC电子商务平台中的用户行为特征,提出在此类站点中实施Web个性化推荐技术的基本原则和特点,并指出该种技术的三个明显特点。最后,结合Web个性化推荐技术的三种主要流程,对CtoC电子商务平台实施Web个性化推荐技术的方法作详细说明。

〔关键词〕CtoC　Web站点　个性化　推荐技术
〔分类号〕G202

Web Personalization Recommendation Technologies in CtoC E-commerce Websites

Li Shuqing
Department of Information Management, College of Information Engineering, Nanjing University of Finance & Economics, Nanjing 210046

〔**Abstract**〕Based on analyzing the characteristics of users' behavior in CtoC E-commerce websites, this paper proposes the basic principle and characteristics of Web personalization recommendation technologies in these Websites. At last, this paper introduces three main processes of Web personalization recommendation, such as data preprocessing, obtaining the personalized contents, recommending these contents, and explains the implementation of Web personalization recommendation technologies in CtoC E-commerce Websites in detail.

〔**Keywords**〕CtoC　Website　personalization　recommendation technologies

1 引　言

随着互联网技术在人类生产和生活各个方面的影响不断加深,传统的商业交易模式也在发生着深刻的变化,一大批网络公司应运而生,所推出的各种网络服务也极大地方便了商业用户的交流和沟通,如EBay、阿里巴巴、淘宝等。

从这些电子商务站点所提供的服务类型看,它们主要可以分为BtoB、BtoC和CtoC等不同类型。其中,BtoB类型和BtoC类型的发展历史较长,面向的用户群体主要是商业企业实体,因此,这些站点往往可以提供质量较高的信息内容和相关服务。与此相对的,作为直接面向终端用户并为他们搭建起沟通桥梁的CtoC类型站点,由于交易用户往往并非专业的买家和卖家,多为一般的个人用户,因此直接将传统BtoB类型和BtoC类型的成功经验借鉴过来,并非是一种构建CtoC电子商务站点的有效方法[1]。特别对于Web个性化推荐技术的应用而言,这种区别表现得更为明显。

2 CtoC电子商务站点中Web个性化推荐技术的特点

在过去的几年中,个性化推荐技术越来越受到人们的关注。然而,作为一门新兴技术,它主要是由较年轻的学科领域知识组成,相关的研究并不系统,所以人们对其认识也各不一样。如有人认为,个性化是一种能力。这种能力可以根据从用户偏好和行为特征中提取到的知识来对用户提供定制化的服务和内容[2];也有人认为,个性化技术综合使用了网络技术和客户信息,由于这些信息包含已收集来的信息和实时产生的信息,因此可以有针对性地定制商业站点以适应用户的交互行为,它能有效地降低交易处理时间和让用户得到更为满意的产品[3];还有的学者认为,个性化是根

*本文系2007年江苏省高校自然科学基础研究项目"基于Web个性化推荐服务的CtoC电子商务平台框架"(项目编号:07KJD520074)研究成果之一。
收稿日期:2008-11-18　修回日期:2009-01-21　本文起止页码:134-137,80　本文责任编辑:徐　健

据已有的用户偏好信息和交易活动中的当前用户行为,定制交流方法的一种能力。也就是说,个性化通过建立一种一对一的联系,理解用户的个别需求,帮助实现在特定环境下准确地表达用户需求特征,从而增强客户的忠诚度[4]。

然而,这些定义都强调了一个主要特征,即能够根据已知的用户信息来定制电子商务网站提供给用户的信息内容,包括产品、服务和交易方法等。由于与其他诸如 BtoB等类型的电子商务站点相比,CtoC电子商务站点中的 Web用户行为特征具有较为明显的不同,因此,它在实施 Web个性化推荐技术的时候,必须充分考虑这些用户行为因素。

结合 CtoC电子商务站点用户的诸多行为特点,该种类型站点所能采用的 Web个性化推荐技术有如下三个明显特点:

· 可以为匿名用户提供推荐内容,无需用户注册登录也能使用个性化推荐功能。从交易过程来看,CtoC电子商务站点所面向的 Web用户群并非商业企业实体,而是一般终端客户。具体来说,卖方用户可以发布所要销售的商品信息,为此,该用户必须是注册用户,并向该站点提供自己的联系信息。而买方用户则可以查询这些商品的发布信息,并选择所需的商品。一旦选择到合适的商品,买方用户就可以根据卖方用户所提供的信息直接与卖方用户取得联系。一般而言,这类站点并不要求买方用户必须注册。事实上,为了方便用户使用,很多站点都允许匿名用户浏览选择所需商品。因此,利用服务器日志中的用户会话信息,可以有效地获取匿名用户的访问特征,从而实现对此类买方用户的个性化推荐功能。

· 应该使用图片等多媒体信息和关键词结合的推荐内容。其中,图片所包含的信息真实性相对较强,买方用户可以直接根据图片内容获知对商品的喜好程度,但是图片本身并不易于检索。相反,关键词信息则可以通过简洁的词语来标明所售商品的重要特征,而且也可以允许快速搜索。但由于卖方用户的非专业性特点,或者由于某种主观因素的影响,卖方用户可能并不能或者不愿提供准确的关键词信息,这就产生一个问题,即买方用户通过关键词虽然可以快速搜索到所需商品,但仍然需要进一步通过图片或者更为详细的文字说明来确定该商品是否满意。因此,结合使用这两种方式可以给用户提供更为准确的推荐内容。

· 只负责内容推荐,不负责辅助交易行为的完成,甚至都无法实现主动推送功能。由于大部分 CtoC站点的买卖双方都没有完善的网络交易渠道,真实的货款和商品交易往往都不在该站点上进行,站点只是一个沟通的媒介。因此,在提供个性化推荐内容时,站点往往并不需要采用信息推送的推荐方法来向用户发布信息。更多的情况是,在用户浏览商品信息时,站点能够在浏览页面上向当前用户展示推荐的商品信息内容。

除此以外,CtoC电子商务站点中 Web个性化推荐技术还需考虑很多其他因素,比如可以综合运用多种个性化技术来处理不同的推荐内容和方法,以达到更好的推荐效果等。

3 实施 Web个性化推荐技术的方法

一般的 Web个性化推荐技术框架可以分为两个主要部分:①离线处理阶段,主要包含数据预处理工作,它主要用于产生用户事务信息和完成特定的挖掘任务等,所使用的方法有关联规则发现和 URL聚类分析等;另一个是在线处理阶段,利用得到的最频繁项和 URL的聚类结果,系统就可以根据当前用户的浏览行为产生动态的推荐内容,并以链接的方式将推荐内容在下一次用户请求时添加到网页上。

3.1 数据预处理

如前文所述,CtoC电子商务站点所采用的 Web个性化推荐技术需要支持对匿名用户的个性化推荐功能,为此,就需要从 Web服务器日志中识别匿名用户的会话信息以获取匿名用户的访问特征。传统的 Web服务器日志分析技术由于可能会受到本地缓存和代理服务器的干扰,因此,识别用户会话信息更有效的方法是利用 Cookie或者对 URL重写。但是由于个人隐私权的问题或者服务器的支持能力问题,这些技术也存在一定的局限性。为此,有学者提出利用引用项和代理域实现的启发式方法,来识别用户会话信息和推断丢失的引用信息[5]。

除了识别用户会话信息外,原始的日志信息还必须经过清洗并被转换为一组访问序列。访问序列的组成元素可以根据实际处理的要求来灵活选择。如由于用户的一次访问会产生对不同网页文件的多个请求,这种由用户一次访问生成的多个被请求网页集合被称为网页访问序列。同样,由用户通过提交查询产生的关键词或者访问指定商品所对应的关键词,也可以生成一种关键词访问序列。在 CtoC电子商务站点中,Web个性化推荐技术应该使用图片等多媒体信息和关

键词结合的推荐内容,所以两种访问序列都需要处理。

下面对此分别进行说明。

3.1.1 图片网页访问信息的处理 由于图片信息能够给买方用户提供较为完整和全面的印象,因此需要获取网页访问序列中的图片信息,即这些图片网页的URL。按照上述的预处理方法,我们可以得到出现在日志中的 n个彼此迥异的图片网页 URL,组成的集合可以表示为:URLs = $\{u_1, u_2, \cdots, u_n\}$。假设用户事务模式的总数量为 m,则所有的事务模式集合可以表示为:Trans = $\{t_1, t_2, \cdots, t_m\}$,其中,每个事务模式向量 t_i 都是一个 n维的向量,每个维度对应一个图片网页的 URL。对于向量元素权重的取值,传统的二值表示方法过于简单,无法反映用户的兴趣程度,因此有必要结合 CtoC站点用户的行为特点,重新设计事务模式向量的权重[6]。

笔者采用了一种基于用户兴趣度的赋值方法。所谓用户兴趣度,是指用户对某个网页的关注程度,它可以从三个方面进行测度:①如果当前用户对某图片网页的访问频率较高,则说明用户对此图片网页的关注程度较高。设 P为最大的网页出现次数,p_k 为网页 K出现的次数,则网页 K的用户兴趣度正比于 p_k/P。②如果当前用户对某图片网页的访问时间较长,则可以利用时间长度进行加权。用户对网页的访问时间可以直接从 Web日志信息中通过相邻网页的访问时间相减得到。设 T为最大的网页访问时间,t_k 为网页 K的访问时间,则网页 K的用户兴趣度正比于 t_k/T。③如果当前用户对某图片网页本身产生的链出网页点击率较高,则可以利用该点击数量进行加权。设 C为最大的网页链出点击数,c_k 为网页 K的链出点击数,则网页 K的用户兴趣度正比于 c_k/C。综上所述,结合用户兴趣度的多元向量权值可以表示为:

$$w_k = \frac{p_k}{P} * \frac{t_k}{T} * \frac{c_k}{C}$$

3.1.2 关键词访问信息的处理 利用查询所使用的关键词和浏览商品所对应的关键词,可以构造一种关键词访问序列。设 R为用户访问产生的操作集合,每一条访问记录 r包括:r.uid(用户 ID号),r.keyword(被访问关键词)和 r.time(关键词所对应的访问时间戳)。其中,后两项是集合属性,代表着该用户一次连续访问所涉及的全部关键词组合,即:r = (r.uid, {r.keyword, r.time})。如果用户没有 ID号,系统可以任意分配一个惟一的随机标识符。

通常,从日志中得到的关键词访问序列是一条相当长的关键词集合。如果需要从中发现用户的个性化信息需求,往往还需要进一步加以整理,如由于用户键入无用关键词和空白词语产生的关键词序列元素并无实际意义,因此需要在处理前将其删除。

这种关键词序列在内容上有着自己的特点:①具有包含词语语义的上下文环境,传统的词语相似度匹配通常需要进行词语语义分析以保证结果的有效,但在关键词序列中,词语都不是单独出现的,所以通过词语所在关键词序列的其他词语,一般能够限定词语的语义概念;②在一定程度上,词语的顺序也表达了用户一次查询的完整过程,代表了用户一个不断调整和定位所需商品内容的思路。不过,关键词访问序列和网页访问序列不同之处在于关键词序列元素之间没有前后对应的必然次序,因此无需对不连续的序列部分进行路径完善。

3.2 获取推荐内容

对于关键词访问序列,比较适合的处理方法是使用诸如 Apriori算法为代表的关联规则挖掘算法。该算法主要用于发现共现频繁项,这里的项即是关键词,项集基于关键词来表示,多个项构成的频繁项集可以表示为:R = $\{r_1, r_2, \cdots, r_k\}$,其中,每个 r_i 都代表了一个频繁项集,其支持度可以利用频繁项集中项所在的事务模式数量占总事务模式数量的比重来表示。支持度的阈值需要在挖掘前指定并应用于算法当中,以缩小搜索空间。只有满足最小支持度的项集才能有效。

关联规则用于发现频繁共现项之间的关系,在 Web事务模式中,关联规则可以表达基于用户查询或者导航行为产生的不同关键词之间的关联关系。尽管存在一定的缺陷,频繁项和关联规则却是一种提供个性化推荐结果的较为有效的方案[7]。

对于由商品图片所在网页 URL组成的网页访问序列,利用事务聚类则可以产生 URL推荐集合,也就是网页 URL的聚类集合。传统的协同过滤技术主要匹配当前的用户模式和服务器在以前时间段从其他用户那里获得的相似聚类模式信息,相似的技术也可以被应用于 Web个性化推荐的计算当中,即对用户事务模式进行有效聚类。此处和协同过滤技术不同之处在于不需要显式的划分等级和与用户交互。用户事务模式可以被映射到以 URL引用向量表示的多维空间。标准的聚类算法通常按照项组之间的距离对项进行归组。在 Web事务中,每个事务聚类代表着一组根据网页 URL共现度计算得到的相似事务。

前文已经说明,假设用户事务模式的总数量为 m,

则所有的事务模式集合可以表示为: Trans = { t_1, t_2, …, t_m },其中,每个事务模式向量 t_i都是一个 n维的向量。为了对这些事务进行聚类,需要对事务之间的距离进行测量,常见的测量方法是计算向量的余项夹角值。这里推荐的方法并未考虑在一个事务中网页 URL的引用次序,主要原因在于其主要应用范围是用于改善网站站点拓扑结构设计以减少网络流量,对 Web的推荐意义不是很大,。

按照上述定义可以计算相似度矩阵,以作为聚类的基础。但是,基于距离的聚类算法在处理高维数据时会产生维数灾难。因此,可以在预处理阶段将低支持度的网页 URL去除,同时,也可以将不属于站点内的网页 URL信息去除[8]。具体的聚类方法有很多,如多元 k值算法 (multivariate k-means algorithm)[9]。不管哪种方法,最终都会产生一组聚类结果集,表示为: C = { c_1, c_2, . . . , c_k },其中,每个 c_i都是 Trans的一个子集。从本质上看,每个聚类都代表了一组具有相似访问模式的用户。但是,仅仅依赖聚类事务模式并不能直接代表用户模式的累计情况。因此,需要计算每个聚类 C的中心向量。类中心向量可以看成是一组网页 URL对应中间值的集合,对于类中每个 URL的中间值,可以利用出现该 URL的事务数量和聚类中事务总数的比值来表示。最后,可以去除具有低支持度的 URL,得到和事务聚类相关的 URL聚类结果。

3.3 结果推荐

推荐引擎是个性化推荐系统中的在线处理模块,它根据使用挖掘得到的信息,计算用户当前会话的推荐结果集。一般而言,在生成推荐集的时候,需要考虑如下因素,如需要选用合适的匹配算法,利用每个聚类与频繁项和当前活动会话的相似程度来得到推荐结果。再如,推荐的网页 URL不应该是用户已经访问过的网页 URL,等等。

对于关键词访问序列而言,从频繁项集直接计算出推荐结果,需要将当前用户会话信息中的关键词和项集关键词集合进行匹配以发现推荐集。如果使用大小为 w的滑窗,只需直接对尺寸大于 w + 1的项集进行计算,判断它们是否满足预定义的支持度和含有当前关键词信息,每个推荐的关键词权值可以利用对应关联规则的置信度来判断,该置信度对应一个唯一的候选关键词。如果规则满足指定的阈值,就可以将其追加到关键词推荐列表中。上述做法中有一点值得注意,那就是有时会难以找到足够大的项集用于提供推荐信息,这对于小型 CtoC站点而言尤为常见,因为这些站点的会话长度通常都很小。在这种情况下,除了减少阈值的处理方法外,也可以通过减少滑窗尺寸来进行处理。当然,这种做法也会导致无法全面考虑用户会话产生的历史访问信息。

对于图片网页的访问序列而言,从事务聚类直接计算出推荐结果,需要将当前用户会话信息和事务聚类进行匹配以发现推荐集。对于每个网页 URL的聚类结果,都可以代表访问站点的几种不同的用户类型组合。一旦有一个新的用户会话开始,推荐的目标就是逐步将部分用户会话信息和合适的聚类进行匹配,最终给出用户可能愿意接受的推荐集合。所以,第一个任务就是要根据和当前用户会话的相似程度,来计算每个聚类的得分值。为了计算匹配程度,有必要对聚类向量和会话向量的尺寸进行规范化,如果直接将没有规范化的向量用于计算,由于大小不一致,较大的聚类一般会得到较小的权值。当然,由此也能看出,只有获得足够多的用户会话信息才能让大尺寸聚类获得更好的分值。最后可以设定一个统一的阈值,以过滤那些不满足要求的聚类结果。

4 结 语

Web个性化推荐技术可以有效地改进用户访问站点的体验,更为深远的影响表现在可以增加用户的忠诚度,有助于在产品供应商与用户之间建立一种长期的联系,而这对于 CtoC电子商务站点而言,就意味着会留住和吸引更多的用户来访问,从而为站点的进一步发展提供了很好的契机。然而,我们也要注意到,由于现阶段互联网基础环境的限制,笔者所述的 CtoC电子商务站点 Web个性化推荐技术仍然存在着诸多需要改进和完善之处,如匿名用户的有效识别与推荐算法的缩放性等,这些都是在下一步研究中需要解决的问题。

参考文献:

[1] 庚佃友. CtoC电子商务盈利模式研究. 商场现代化, 2006 (12): 133.

[2] Adomavicius D, Tuzhilin A. Personalization technologies: A process-oriented perspective. WIRTSCHAFTSINFORMAOTIK, 2006, 48 (6): 449 - 450.

[3] Personalized Gifts and Personalized Gift Ideas from Personalization Mall [2008 - 10 - 15]. http: //www. personalizationmall com /.

[4] Jill Dyche. CRM Handbook. Boston: Addison-Wesley Co., 2002: 112 - 167.

(下转第 80页)

份分析法、因子分析法、熵值法等[24]。

我国现有区域知识竞争力评价方法多采取主成份分析法、因子分析法。对知识区域竞争力评价模型的研究有林善浪,王健构建的知识竞争力层次模型;曹如中等构建的知识竞争力决定因素钻石模型和城市知识竞争力决定因素循环链;相丽玲等构建的知识竞争力模型。

4 存在的问题与发展趋势

·知识竞争力的理论研究依赖于经济学与管理学对竞争力理论的研究,而目前经济学与管理学对竞争力理论的研究流派众多,观点不一。知识竞争力的理论研究需要建立一个理论框架,对"知识竞争力"进行命题与验证。

·知识竞争力评价体系的构建与测度,传承于知识经济的评价与测度、国家竞争力的评价与测度、国家创新体系的评价与测度。而目前国际三大竞争力指标体系在对"知识能力"的测度理念与方法上的存在分歧,国内关于知识竞争力相关理论与实证研究基本是描述性的,缺乏理论依据与数据支持。知识竞争力的应用研究需要梳理相关的评价体系与测度方法,使知识竞争力评价体系建立在科学的基础之上。

参考文献:

[1] 潇雨. 知识是竞争力. Market Observation, 1999(1): 17.

[2] 艾米顿. 知识经济的创新战略:智慧的觉醒. 金周英等,译. 北京:新华出版社, 1998.

[3] 彼得·圣吉. 第五项修炼:学习型组织的艺术与实务. 郭进隆,译. 上海:上海三联书店, 1998.

[4] 吴惠阳. 透视上海知识竞争力. 上海经济, 2005(5): 47-49.

[5] 王江,金占明. 核心知识竞争力与企业多元化战略. 企业管理, 2005(10): 134-137.

[6] 姚国琴. 知识竞争力与世界知识经济格局. 学习论坛, 2002(6): 35-37.

[7] 陈舒,王健. 知识竞争力及其指标体系研究. 财经界, 2007(8): 259.

[8] IMD World Competitiveness Center. IMD World Competitiveness Year-book 2005. Lausanne, Switzerland: IMD, 2005.

[9] Ricardo D. The Principles of Political Economy and Taxation. Homewood: Irwin, 1963.

[10] Ohlin B. Interregional and international trade. Cambridge, Mass: Harvard University Press, 1933.

[11] 约瑟夫·熊彼特. 经济发展理论——对于利润、资本、信贷、利息和经济周期的考察. 何畏译. 北京:商务印书馆, 1990: 290.

[12] Romer P M. Increase returns and longrun growth. Journal of Polltical Economy, 1986, 94(10): 15-18.

[13] 金培. 竞争力经济学. 广州:广东经济出版社, 2003(5): 440-466.

[14] 李品媛. 企业核心竞争力研究——理论与实证分析. 北京:经济科学出社, 2003(4): 45-47.

[15] Leonard-Barton D. 知识与创新. 孟庆国,侯世昌,译校. 北京:新华出版社, 2000: 57-59.

[16] 陆淳鸿. 企业竞争优势理论演进评述. 经济问题, 2007(4): 24-25.

[17] 郭斌. 界面管理:企业创新管理的新趋向科学学研究, 1998(1): 60-67.

[18] 易法敏. 核心能力导向的企业知识转移与创新研究. 北京:中国经济出版社, 2006(6): 22-23.

[19] 迈克尔·波特. 国家竞争优势. 李明轩,等译. 北京:华夏出版社, 2002: 17-19.

[20] 林善浪,王健. 知识竞争力及其评价指标体系研究. 科技进步与对策, 2008(2): 106-109.

[21] 张川蕾. 中国区域知识竞争力的综合评价及对策建议. 国际经济合作, 2008(4): 49-53.

[22] 相丽玲,汤亮亮,薛全胜. 区域知识竞争力的构成要素及其模型. 情报理论与实践, 2008(4): 515-517.

[23] 曹如中,胡伟强,戴昌钧. 城市知识竞争力决定因素评价研究. 中国科技论, 2008(2): 116-119.

[24] 李莉,高志刚. 关于区域竞争力内涵,指标体系,评价方法的研究述评. 新疆职业大学学报, 2005(3): 8-11.

作者简介〕相丽玲,女,1962年生,教授,在读博士,硕士生导师,发表论文 30余篇,出版专著 7部。

成 佳,女,1983年生,硕士研究生,发表论文 2篇。

(上接第 137页)

[5] Mobasher B, Cooley R, Srivastava J. Automatic personalization based on web usage mining. Communications of the ACM, 2000, 43(8): 142-151.

[6] 李树青,崔北亮. 搜索引擎系统中的 Web个性化信息推荐技术. 情报杂志, 2006(9): 84-87.

[7] 韩家炜,孟小峰,王静,等. Web挖掘研究. 计算机研究与发展, 2001(4): 405-414.

[8] 张日崇. 基于 Web的个性化挖掘方法[学位论文]. 吉林:吉林大学, 2004.

[8] 刘远超,王晓龙,刘秉权. 一种改进的 k-means文档聚类初值选择算法. 高技术通讯, 2006(1): 11-15.

作者简介〕李树青,男,1976年生,讲师,发表论文 13篇,出版著作 2部。