基于 Web 日志挖掘和相关性度量的电子商务推荐系统①

马 勇[1], 鲜 敏[1], 郑 翔[1], 黎远松[2]

[1](四川工程职业技术学院 电气信息工程系, 德阳 618000)
[2](四川理工学院 计算机学院, 自贡 643000)

摘 要: 基于 Web 日志挖掘的个性化推荐技术已在电子商务网站中广泛应用, 针对现有推荐系统的准确性不高等问题, 提出一种基于 Web 日志挖掘和相关性度量的个性化推荐系统. 首先, 提取用户的访问日志, 并对其进行预处理, 以获得精简的结构化数据. 然后, 对日志进行分析, 提取出特征序列. 再后, 根据特征的出现频率和页面停留时间, 计算出页面与交易文本文档的相关性. 最终, 利用夹角余弦公式计算出用户与页面的相关性, 并以此形成推荐列表. 实验结果表明, 该方案能够根据用户偏好精确的给出个性化推荐.

关键词: Web 日志挖掘; 推荐系统; 相关性度量; 电子商务

E-Commerce Recommender System Based on Web Log Mining and Correlation Measure

MA Yong[1], XIAN Min[1], ZHENG Xiang[1], LI Yuan-Song[2]

[1](Sichuan Engineering Technical College, Deyang 618000, China)
[2](School of Computing, Sichuan University of Science and Engineering, Zigong 430072, China)

Abstract: Nowadays, personalized recommender technology based on Web log mining has been widely used in the e-commerce website. For the issues that the existing recommender systems do not have high accuracy, a recommendation system for e-commerce based on web log mining and correlation measure is proposed. First, the user's access log is extracted, and the data is preprocessed to obtain the structured data. Then, the log is analyzed to extract the characteristic sequence. After that, the correlation between the page and the transaction text documents is calculated according to the occurrence frequency of characteristics and the page dwell time. Finally, the angle cosine formula is used to calculate the correlation between the user and the page, and thus form a list of recommendations. Experimental results show that the proposed scheme can accurately give personalized recommendation according to the user's preference.

Key words: e-commerce; recommender system; Web log mining; correlation measure

现今, 利用网络进行日常商业交易的互联网用户越来越多, 许多公司也利用网络来销售他们的商品和服务, 电子商务已改变了传统买卖的方式. 然而, 这也使公司和顾客面临着一些新的挑战. 对于一个特定的商品, 顾客将面临多个选择, 使其处于困惑和迷失状态. 对于网站管理员而言, 评估提供的商品和服务是否迎合顾客的需要已经变得至关重要. 处理这种问题的有效方案是为个人用户提供个性化推荐, 为顾客提供感兴趣的商品推荐单[1]. 目前, 已存在多种推荐系统, 可以分为两大类: 基于内容的系统和协同过滤系统. 基于内容的系统基于商品内容而生成推荐, 协同过滤系统利用客户与商品之间的交互而形成推荐[2]. 尽管这些推荐系统取得了明显的进步和广泛的应用, 但是它们仍然存在一些局限性, 其中一个问题是大多数推荐系统使用二进制事务(点击流)数据, 即表示一个特定的商品是否被购买. 然而, 仅利用这些数据不能够提供更好的推荐.

Web 挖掘[3]是数据挖掘技术在 Web 上的应用, 将传统的数据挖掘技术与 Web 相结合, 从网络上挖掘有用的信息. Web 挖掘技术的发展提升了电子商务推荐系统在企业运营中的应用价值. 基于 Web 挖掘的购物个性化推荐系统可以直接与用户交互, 模拟商店销售

① 基金项目:四川省高校重点实验室项目(2014WZY05);四川省智慧旅游研究基地规划项目(ZHY15-01)
收稿时间:2016-01-12;收到修改稿时间:2016-03-01 [doi:10.15888/j.cnki.csa.005341]

人员向用户提供商品推荐，帮助用户找到所需商品，从而顺利完成购物过程. 目前，学者提出了多种基于 Web 挖掘的网页推荐系统，例如，文献[4]提出一种个性化的最优搜索引擎，根据用户的反馈文本，获取检索信息来训练反向神经网络，最终实现电子商务网页的公正排名. 然而，该方案需要大量的反馈样本，对没有反馈的新商品不能很好地进行推荐. 文献[5]提出一种基于语义 Web 数据挖掘技术的网页推荐算法，分析 Web 日志数据，发现用户访问模式，同时挖掘商品的购买顺序以及时间，从而构建推荐模型. 然而其使用空间向量模型来表示文本，这使得推荐的准确度较低，且推理和判断能力较弱. 文献[6]描述了一种基于 Web 字典的网页推荐算法. 该算法根据早期用户访问网页的内容和时间来确定该网页的重要性，改进搜索引擎算法的时间和空间复杂度. 其在大型 Web 数据库上的实验表明其一定程度上能为用户提供所需的网页. 然而，该方案只能应用于注册用户，对于非注册用户不能形成推荐.

本文提出一种基于 Web 日志挖掘和相关性度量的个性化推荐系统. 对于所有用户，提取其访问日志，通过预处理获得精简结构化数据. 然后提取出特征序列，并根据特征的出现频率和页面停留时间，利用夹角余弦公式计算出用户与页面的相关性，并以此形成推荐列表. 实验结果表明，本文方案比现有方案能够更好的给出个性化推荐.

1 提出的推荐系统

通常，当用户与门户网站进行交互时，会将用户点击数据保存到原始日志文件中. 通过数据预处理和数据清理单元，能够从原始日志文件中提取有价值的信息，并将它转换为结构化形式，进一步用于发现模式，为电子商务网站的用户提供商品推荐. 图 1 描述了本文提出的推荐系统框架，下面将详细描述本文推荐系统中的各个阶段.

1.1 数据采集

在这一阶段，采集并储存所有被浏览网页的导航数据. 本文利用通用日志文件格式来保存数据，记录重要属性，即 IP 地址、时间戳、状态代码、URL、Http 方法(GET 和 POST)、用户代理和推荐人 URL，用作进一步分析. 获得的数据在本质上是非高度结构化且不一致[7]，因此，必须进行预处理.

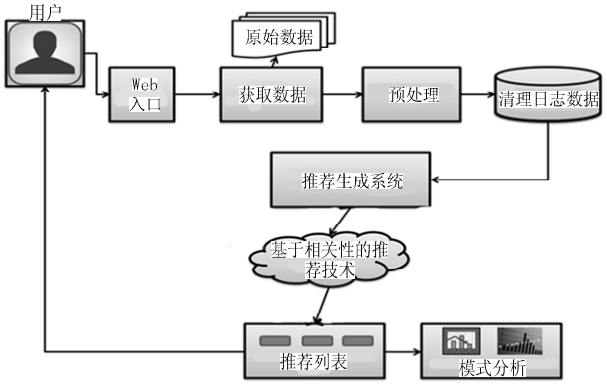

图 1 本文推荐系统框架

1.2 数据预处理

数据预处理用于消除不一致和冗余的数据，其过程如图 2 所示.

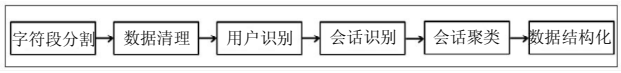

图 2 数据预处理阶段

(1) 字段分割: 通过使用分隔符(如空号)来区分一种属性与另一种属性.

(2) 数据清理: 日志中记录的原始数据通常非常大，存在很多对于挖掘任务没有意义的数据记录和数据项. 其中包括用户会话中一同下载并记录在日志中的图形文件 (gif, jpg, jpeg)、音频文件(mp3,mid,wmv)、视频文件 (rm)、格式文件(css,js)等记录[8]. 其次是搜索引擎产生的用户请求访问失败的记录，这类访问的返回代码为 404(请求的页面没有找到)、301(永久删除)、500(内部服务器错误). 另外，还有其他无关的日志，例如后缀为: css, map, js 等文件. 通过清理可以大大缩减数据总量，提高会话识别的精度.

(3) 用户识别: 其目的是分析究竟有多少不同的用户访问. 由于用户端高速缓存、代理服务器、防火墙以及动态地址池的存在，使得这一过程的实现较为困难. 一般的方法是采用启发式规则，以用户 IP 和代理来唯一确定用户，即用户 IP 地址和代理同时相同时为同一个用户. 当 IP 和代理都相同时，则利用引用日志和网站拓扑结构判断请求访问的页与过去访问的网页是否存在链接，如果不存在任何链接，就认为同一台机器上存在两个用户[9].

(4) 会话识别: 就是用户在规定时间内(或称一次浏览内)对服务器的一次有效访问，通过其连续请求的

页面, 可以获得他在网站中的访问行为和浏览兴趣. 通常采用时间窗口模型, 以用户访问时间作为划分会话的分界, 一般间隔时间取 30 分钟[10].

(5) 会话聚类: 目的是对属于唯一用户的会话进行分组, 将独立用户的会话组织起来. 会话信息为本文提供了用户在特定期间的整套活动信息.

(6) 最后, 在数据格式编排阶段, 本文以表格的形式存放这些会话数据.

1.3 提取特征序列

在该步骤中, 从上述获得的结构化日志中提取特征信息. 对于每次会话, 本文都构造一个字符串序列, 称为特征序列. 将会话中的提取特征信息, 如商品名称(p_id)、种类名称(c_id)、商品访问频率(f)和页面上停留的时间(tsp)保存到该字符串序列中. 图 3 为将遍历日志提取的特征信息转换成字符串形式的实例.

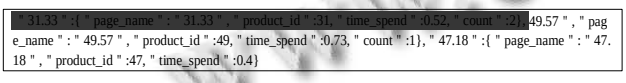

图 3 日志特征信息的字符串形式

图 3 中, 灰色的部分突出显示的 31.33 为对象名称, 其中, 31 为 p_id, 33 为 c_id, 0.52 为页面停留时间(tsp), 2 为特定会话中的商品访问次数.

1.4 基于相关性度量生成推荐

本文通过 Web 日志挖掘, 计算出页面与当前用户会话的相关性, 并以此排序页面形成推荐列表. 如果两个页面的相关性相同, 则通过判断用户之前在页面的停留时间, 将停留时间较长的加入推荐列表. 本文生成推荐列表的流程如图 4 所示.

在上文对单个 Web 页面进行特征序列提取后, 本文可以得到单个页面的特征序列集合 $T=\{t_1,t_2,t_3,...,t_n\}$, 并将一次交易中涉及到的页面集合表示为 $P=\{p_1,p_2,p_3,...,p_m\}$. 根据特征序列 t 和页面 p 的关系, 特征序列可以根据空间向量初步形成矩阵关系. 本文设定 $tw<t_j,p_i>$ 表示特征序列第 j 个特征 t_j 和第 i 个页面 p_i 的权重关系, 其根据特征序列在页面中出现的频率和特征序列 t_j 与页面 p_i 的相关性计算获得, 表达式如下:

$$tw<t_j,p_i>=tf\times w_j \tag{1}$$

上式中, tf 为特征序列在一定时间内出现的频率, w_j 为特征序列 t_j 和页面 p 的相关性, 表达式为:

$$w_j=\sum_{j=1}\log\frac{pt_j(1-ps_j)}{ps_j(1-pt_j)} \tag{2}$$

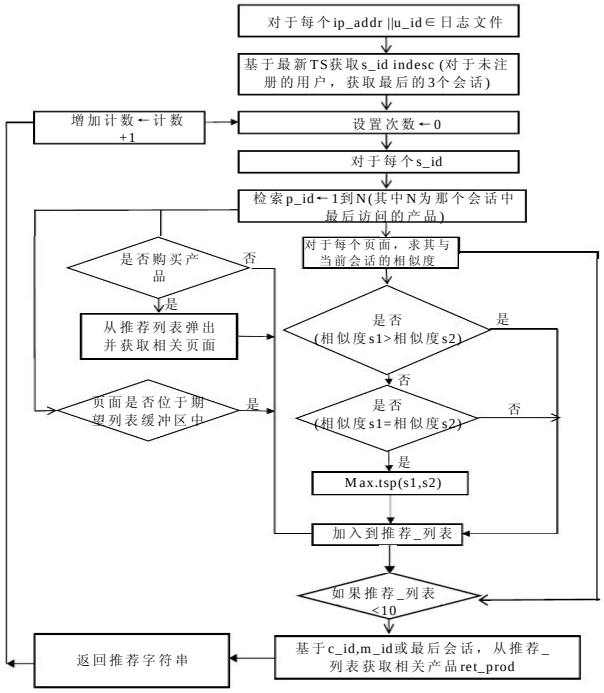

图 4 本文生成推荐列表流程

上式中, pt_j 为特征序列 t_j 在相关页面中出现的概率, ps_j 为特征序列 t_j 在非相关页面中出现的概率.

在获得特征序列 t 和交易页面 p 的关联矩阵之后, 本文还考虑了另外一个关键因素, 即页面停留时间(tsp). 为了确定交易 s 和页面 p 之间的访问关系, 本文基于页面的停留时间 tsp 来构建一个访问权重 $pw<p_i,s_j>$, 其表达式为:

$$pw<p_i,s_j>=\frac{tsp_i}{Content_i} \tag{3}$$

上式中, $Content_i$ 表示访问的第 i 个页面的文本长度, s_j 为交易事务集 $S=\{s_1,s_2,s_3,...,s_q\}$ 中的一次交易, tsp_i 为第 j 次交易中用户在第 i 个页面上的停留时间. $pw<p_i,s_j>$ 表示了停留时间与访问页面的文本长度大小之间的比值.

然后进行交易事物聚类, 目的是在相关性的基础之上对目标数据进行分类, 把相关性接近的数据凝聚在一起. 本文利用层次聚类方法将以往相似交易的文本聚类到一起, 形成一个文本文档集 pc, 即每个文档集对应一种相似商品类. 每个页面 p 在文本文档 pc 中的权重可表示为 $pcw<p_i,pc_i>$, 用于计算会

话相关性, 其表达式为:

$$pcw < p_i, pc_i >= \frac{\sum_{p_i \in P}\sum_{j=1}^{n} tw < t_j, p_i >}{\sum_{p_i \in P}\sum_{j=1}^{n} tw < t_j, p >} \quad (4)$$

当一个用户进行访问时, 产生当前会话 U. 令 $w_u < p_i, U >$ 表示页面 p_i 在当前用户会话中的权重, 其为所有特征序列在页面 p_i 中权重总和, 与所有特征序列在所有页面集 p 中权重总和的比值.

那么, 就可以计算当前用户会话 U 和文本文档 pc 之间的相关性. 本文采用夹角余弦公式来计算该相关性, 表示式如下:

$$Sim(U, pc) = \frac{\sum_{i=1}^{m} w_u \times pcw}{\sqrt{\sum (w_u)^2 \times \sum (pcw)^2}} \quad (5)$$

通过上述计算过程获得用户会话 U 和已经产生的相似交易文本文档 pc 集之间的相似度, 相似度越大, 说明当前用户需求与该文档集所对应的商品越相近, 所以本文以相似度作为最终推荐值, 并以此排序商品页面, 产生最终的网页推荐序列.

2　实验及分析

利用 XAMPP 服务器、phpMyAdmin 和代码编辑器 3IDE 执行本文提出的系统. 利用基于 MVC 的开源电子商务系统 Opencart 来构建网络购物环境. 本文为客户提供了不同商品的门户网站, 所使用的数据集为实时数据, 包含来自购物网站“asia-shopping”的 1121 个记录, 其中包括 100 名用户和 20 个商品. 图 5 为预处理后的结构化日志文件.

id	request	date	file_path	referrer	session_id	ip	server_protocol
1	GET	2015-11-28	data/demo/ipod_classic_1.jpg	http://pjshopping.asia/	1	116.143.229.227	HTTP/1.0
2	GET	2015-11-28	http://pjshopping.asia/index.php?route=product/pro...	http://pjshopping.asia/	1	116.143.229.227	HTTP/1.0
3	GET	2015-11-28	http://pjshopping.asia/index.php?route=common/home	http://pjshopping.asia/index.php?route=product/pro...	1	116.143.229.227	HTTP/1.0
4	GET	2015-11-28	data/demo/apple_cinema_30.jpg	http://pjshopping.asia/index.php?route=common/home	2	116.143.229.227	HTTP/1.0
5	GET	2015-11-28	http://pjshopping.asia/index.php?route=product/pro...	http://pjshopping.asia/index.php?route=common/home	2	116.143.229.227	HTTP/1.0
6	GET	2015-11-28	http://pjshopping.asia/	http://pjshopping.asia/	3	116.143.229.227	HTTP/1.0
7	GET	2015-11-28	http://pjshopping.asia/	http://pjshopping.asia/	3	116.143.229.227	HTTP/1.0

图 5　结构化的日志数据

本文以召回率、精度和准确度为评估系统有效性指标. 这些性能指标的计算需要 4 个参数, 即真阳性(TP)、假阳性(FP)、真阴性(FN)和假阴性(TN)率, 这些参数的定义如表 1 所示.

表 1　性能参数定义矩阵

	系统推荐的商品	系统未推荐的商品
期望的商品	真阳性(TP)	假阴性(FN)
不期望的商品	假阳性(FP)	真阴性(TN)

召回率为系统正确推荐的商品占用户所需商品的比例.

召回率=[redacted]　(6)

精度为系统正确推荐的商品占系统所推荐商品总数的比例.

精度=[redacted]　(7)

准确度为推荐系统所作出的正确判断数占所有判断的比例.

准确度=[redacted]　(8)

将本文方法与文献[5]和文献[6]方案进行比较, 图 6 描述了各种方法所获得的平均精度、召回率和准确度值. 可以看出, 本文方法的推荐性能最好, 其中召回率约为 81.5%, 精度约为 89.5%, 准确度约为 86%. 这是因为, 本文从 Web 日志中挖掘出用户访问的特征序列, 包括页面停留时间, 同一特征出现频率等信息, 并根据这些信息计算出该用户与商务网站页面的相关性, 根据相关性排序网页, 给出推荐. 所以本文方案能够很好地根据用户之前的偏好给出准确的推荐.

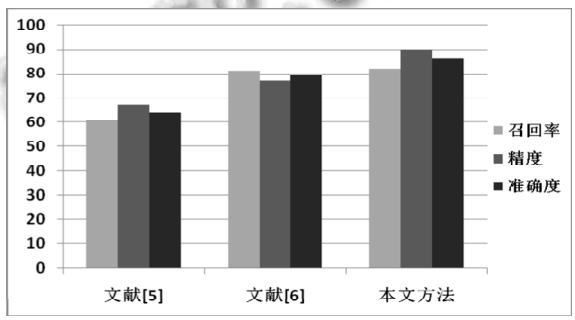

图 6　各种方案的性能比较

3　结语

本文提出一种基于 Web 日志挖掘和相关性度量的个性化推荐系统. 对用户访问日志进行预处理, 消除不一致和冗余的数据. 利用 Web 日志挖掘技术提取特征序列, 通过相关性分析计算出用户与页面的相关性, 并以此形成推荐列表. 与现有方案在召回率、精度和准确度方面进行比较, 结果表明, 本文方案获得了较

高的准确度.

在今后的工作中, 将考虑应用关联规则挖掘技术, 对用户访问的路径进行关联性分析, 进一步提高本文推荐系统的性能.

参考文献

1 解男男,胡亮,努尔布力,等.基于 Web 日志挖掘的网页推荐方法.吉林大学学报(理学版),2013,51(2):267–272.

2 黄伟建,桑志超,杜巍.电子商务环境下的 Web 数据挖掘系统架构设计.河北工程大学学报(自然科学版),2014,31(2): 83–85.

3 Carmona CJ, Ramírez-Gallego S, Torres F, et al. Web usage mining to improve the design of an e-commerce website: OrOliveSur.com. Expert Systems with Applications, 2012, 39(12): 1243–1249.

4 Karthik M. Secure web mining framework for e-commerce websites. International Journal of Computer Trends & Technology, 2013, 4(5): 321–334.

5 Siddiqui AT, Aljahdali S. Web mining techniques in e-commerce applications. International Journal of Computer Applications, 2013, 69(8): 39–43.

6 Wang TZ. The Ontology recommendation system in e-commerce based on data mining and web mining technology. Advances in Electronic Commerce Web Application & Communication, 2012, 35(2): 124–132.

7 宋淑彩,祁爱华,王剑雄.面向 Web 的数据挖掘技术在网站优化中的个性化推荐方法的研究与应用.科技通报,2012, 28(2):117–119.

8 郭晓晨.电子商务中的 web 数据挖掘应用研究.长春理工大学学报,2012,7(7):56–59.

9 Verma N, Malhotra D, Malhotra M, et al. E-commerce website ranking using semantic web mining and neural computing. Procedia Computer Science, 2015, 45(3): 42–51.

10 Yu CY, Shan J. The application of web data mining technology in e-commerce. Advanced Materials Research, 2014, 24(7): 1503–1506.