

基于 Web 日志和商品分类的 协同过滤推荐系统

□崔亚洲 段刚 [电子科技大学 成都 610054]

〔摘要〕传统的协同过滤算法一般以用户对商品的评分作为研究,考虑到显式的评分数据一般很少能够取到,本文从记录用户浏览历史的服务器端日志文件以及客户购买历史数据作为数据源,根据商品分类进行评分矩阵设计,最后由此得到最近邻居从而给予个性化推荐,试验证明推荐效果比较显著。

〔关键词〕协同过滤; 服务器端日志; 个性化推荐

〔中图分类号〕F760.2 〔文献标识码〕A 〔文章编号〕1008-8105(2006)03-0039-04

引言

近年随着 Internet 的发展,电子商务发展迅猛,为了更好地抓住用户的兴趣爱好,从而使得用户多次购物并保持对电子商务企业的忠诚,电子商务个性化推荐系统发展起来。推荐系统泛指任何将个性化推荐作为输出的系统或在大规模可选对象中依据个人喜好引导用户作出选择的系统,在用户浏览商品过程中,提供商品信息和建议,模拟销售人员帮助用户完成购买过程^[1]。推荐系统可以从三个方面来提高网站的销售能力:将浏览者转变为消费者;提高网站的交叉销售;提高客户的满意度^[2]。

协同过滤是较早运用在电子商务推荐系统中的方法之一,它的核心思想是用户倾向于购买具有相似兴趣爱好的用户群所购买的商品,基本的出发点是:客户是可以按兴趣分类的;客户对不同信息的评价包含了客户的兴趣度;客户对一个未知信息的评价将和其相似兴趣客户的评价相似^[3]。这三点构成了协同过滤的基础,用户对商品越感兴趣,则评价就越高。客户-商品评分矩阵如下,有 m 客户对于 n 种商品的评分形成 $m \times n$ 阶矩阵,用 $A(m, n)$ 表示, $R_{i,j}$ 代表 i 客户对 j 商品的评价:

$$A = \begin{bmatrix} R_{1,1} & \cdots & R_{1,j} & \cdots & R_{1,n} \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ R_{i,1} & \cdots & R_{i,j} & \cdots & R_{i,n} \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ R_{m,1} & \cdots & R_{m,j} & \cdots & R_{m,n} \end{bmatrix}$$

由于一方面客户商品评分数据少以及客户不能准确评分,另一方面随着客户、商品的不断增长,导致了用户评分数据的极端稀疏,结果是不能对某些客户进行推荐从而影响了推荐的精度。

文献[4]采用了维度约简技术(Dimensionality Reduction),利用一种单一值分解技术(Singular Value Decomposition)来约简推荐系统中顾客-商品矩阵数据的维数。文献[5]提出通过对稀疏数据的关联可以提高推荐质量。文献[6]提出基于 Web 使用挖掘和 Web 内容挖掘的推荐系统。本文通过分析服务器日志以及购买历史构建用户-商品类矩阵,来发现用户的最近邻予以商品推荐。

一、Web 日志分析

用户在商务网站购物时,会在服务器端的 Web 日志里留下足迹,可以利用这种隐含手段来得到一些有用信息,如:用户浏览了哪些网页,浏览了多少类的

* [收稿日期] 2005-09-14

** [作者简介] 崔亚洲(1979—)男,电子科技大学 2003 级管理学院硕士研究生;段刚(1964—)男,电子科技大学管理学院副教授。

商品?有多少商品被加入购物车?在结账前有多少商品从购物车中取出?有多少用户是先浏览商品然后再进行购物?有多少用户只浏览商品而不从购物?通过这些日志,可以容易地发现用户的兴趣度。

一般结构的日志文件的格式是:访问日期和时间、访问的IP地址、客户号、访问的网页URL名称、下载文件的大小、页面加载所耗时间、访问请求是否成功的状态值、访问者浏览器以及操作系统等。不同的Web Server 其日志格式也不一样,常用的Microsoft IIs6.0的日志格式如下:

```
# Fields : date time s - ip cs - method cs - uri - stem
cs - uri - query s - port cs - username c - ip cs ( User -
Agent ) sc - status sc - substatus sc - win32 - status

2005 - 03 - 24 23 :54 :12 202.119.200.87 GET /
images/blog_footer1.gif - 80 - 218.58.55.123 Mozil-
la/4.0 + ( compatible ; + MSIE + 6.0 ; + Windows + NT
+ 5.1 ; + SV1 ; + Maxthon ) 200 0 0
```

日志分析的工具较多,国外比较有名的有 Analog、Web Utilization Miner、pitkow^[7]等。由于日志文件存储了所有用户浏览情况,有许多访问页面对分析是没有作用的,如一些后缀为.jpg、.bmp、.gif的图片文件,这些文件在分析前需要从日志中清除。日志处理一般通过如下步骤:数据清洗、客户验证、会话识别、路径补全,经处理后的日志可以通过各类挖掘方法来发现用户的整个浏览模式。

二、推荐系统设计

本文用的数据源从日志以及购买历史两部分来处理,并根据商品分类情况对用户浏览、购买赋以不同的数值设置用户-商品类评分矩阵,根据评分矩阵进行最近邻的查找,最后从最近邻中找出频繁集而推荐。

(一) 分值设定

评分值根据用户浏览页面次数和购买的商品件数设定,购买的越多,其兴趣度越大,而浏览的次数越多,也可以说明其对该商品是比较感兴趣的。如下所示, i 表示用户, j 表示商品, m 和 n 表示用户浏览商品的次数和用户购买商品的个数, a 和 b 代表两种系数:

$$R_{ij} = \begin{cases} a * m & \text{浏览了 } m \text{ 次} \\ b * n & \text{购买了 } n \text{ 次} \\ 0 & \text{其它} \end{cases}$$

R_{ij} 代表评分值,分值越大,表示用户对该商品越感

兴趣。

这些值可以在不同的环境中进行不同设定,总的来说,客户浏览评分的最大值应该小于购买的分值,即无论你浏览了商品多少次,也比不上购买一件该商品的兴趣度大。将用户购买商品的次数评分以及浏览的次数评分累加,即可得到该用户对该商品的最后得分,总评分都是大于0的数值,值越大则表明该用户对该商品的兴趣越大。

(二) 商品分类

一般企业商品都按一定规则进行分类,如:鞋子包括运动鞋、皮鞋、凉鞋、拖鞋等;办公用品分为电脑、文件夹、笔、办公桌椅等。分类的情况视具体企业而定,同一种商品在不同公司也许就分在不同的类别中。《毕业那天我们一起失恋》A公司可以划到网络小说的类别中,B公司就可能归结到青春小说中。对于电子商务公司来说,好的分类不仅可以让用户节省时间快速找到自己需要的商品,还可以使页面布局清晰明了,给用户赏心悦目的良好体验。

(三) 构建用户-商品类评分矩阵

经处理后的日志再结合商品分类信息,就可以构建用户-商品类矩阵。根据数据量的大小,建立到第几层次的商品分类,一般越接近底层分类,则数据越准,但是矩阵项目会较多、稀疏性大从而会影响处理的时间和效率,越是概括性的分类数据,则矩阵规模就越小,处理时间越快但准确性会有所降低。可以在二者中找到一个中间点,使效率、精度达到最优。

(四) 最近邻的查找

矩阵建立好后,接下来就是要找出相同兴趣的用户,可以通过修正的余弦相似性来计算:设对项目 i 和项目 j 共同评分过的用户集合用 U_{ij} 表示, U_i 和 U_j 分别表示对项目 i 和项目 j 评分过的用户集合,则项目 i 和项目 j 之间的相似性 $\sin(i, j)$ 为:

$$\sin(i, j) = \frac{\sum_{c \in U_{ij}} (R_{i,c} - \bar{R}_i)(R_{j,c} - \bar{R}_j)}{\sqrt{\sum_{c \in U_i} (R_{i,c} - \bar{R}_i)^2} \sqrt{\sum_{c \in U_j} (R_{j,c} - \bar{R}_j)^2}}$$

$R_{i,c}$ 表示用户 i 对项目 c 的评分, $\bar{R}_i$ 、 $\bar{R}_j$ 表示用户 i 和用户 j 对项目的平均评分^[8]。

最近邻查询的目标就是对每一个用户 u ,在整个用户空间中查找按相似性由高到低排列的用户集合 $C = \{C_1, C_2, \dots, C_k\}, u \notin C$ 。

(五) 商品推荐

最近邻取出后,然后按频繁集进行推荐。设用户 u 购买的商品集合为 $S = \{S_1, S_5, S_7, S_{10}\}$,最近邻集

合 C 用户群的商品集合按出现频率高低排列为 $S = \{S_1, S_2, S_3, S_4, S_5, S_6, S_7, S_8, S_9, S_{10}\}$,则可以取前 5 个 u 没有购买的商品 $\{S_2, S_3, S_4, S_6, S_8\}$ 给予推荐。

三、实证研究

以电子商务公司 A 为例 ,从服务器取日志文件 ,从后台的 SQL Server2000 数据库里取销售历史数据 ,考虑到该公司数据量大以及减少因各类促销而引起销售量的变化等因素 ,我们取 2004 年 3 月 1 日到 2004 年 8 月 30 日这半年促销较少的月份数据作为测试集 ,前 3 个月作为训练集 ,后 3 个月数据作为测试集 ,包括日志文件、客户购买数据两部分的数据。

对数据进行清洗处理 ,对于日志文件 ,由于存储了所有客户浏览历史 ,我们只需要客户浏览商品页面的数据 ,只取格式为“ /product . asp/product . asp ? productid = 商品代码 ”的页面 ,这些页面就是客户浏览过的商品页面。经过处理后 ,我们得到以下形式的日志文件 :

表 1 处理后的服务器日志

IP 地址	访问时间	浏览路径(相对路径)	浏览次数	操作系统
168 . 192 . 100 . 20	2004/03/01	/product . asp/product . asp ? productid = 35681	1	Windows2000
...	...	...	...	...
202 . 103 . 252 . 13	2004/04/10	/product . asp/product . asp ? productid = 224564	3	WindowsMe
...	...	...	...	...
202 . 112 . 15 . 85	2004/05/31	/product . asp/product . asp ? productid = 5688	2	WindowsNT
202 . 112 . 15 . 85	2004/05/31	/product . asp/product . asp ? productid = 174426	2	WindowsNT

对于用户的购买历史 ,我们去掉一些减价促销的商品 ,该数据中主要有 1 元、5 元商品促销以及价格为 0 元的附赠品 ,取得格式如下 :

表 2 用户购买信息

用户号	对应 IP	购物日期	商品 id	购买数量	价格
115521	202 . 115 . 15 . 85	2004/04/01	244524	5	15 . 34
...	...	...	...	...	...
127485	168 . 192 . 100 . 20	2004/05/06	287954	1	22 . 00
...	...	...	...	...	...
66885	202 . 103 . 117 . 49	2004/06/30	124568	2	10 . 40

商品分类的数据(以图书为例) ,最低分层为 5 层 ,对商品数据进行分析后 ,发现近 80.47% 的商品

所属分类都是到第 4 层 ,所以分类数据取到第 4 层 ,这样既可以减少部分数据又不会让分析精度降低太多 ,部分分类数据如下 :

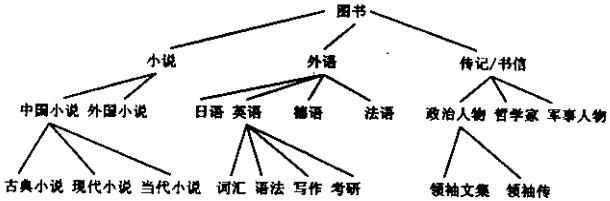


图 1 图书分类

接下来进行评分设置 ,经统计分析后得到 ,用户一般会浏览 4 种商品才购买 1 种商品 ,每种商品购买数少于 10 件的占总购买数的 83.6% ,而从兴趣度方面考虑 ,总浏览商品的分值应该小于购买商品的分值 ,综合以上情况考虑评分如下 :

$$R_{ij} = \begin{cases} 0.2m & \text{浏览商品 } m \text{ 次 } 0 < m < 5 \\ 0.9 & \text{浏览商品 } m \text{ 次 } ,m \geq 5 \\ 1n & \text{购买商品 } n \text{ 个 } 0 < n < 10 \\ 10 & \text{购买商品 } n \text{ 个 } ,n \geq 10 \\ 0 & \text{其它} \end{cases}$$

R_{ij} 代表评分值 , i 表示用户 , j 表示商品。

经汇总分类处理后 ,得到关于用户 - 商品类的评分矩阵 ,部分数据如下 :

表 3 用户 - 商品类评分数据

用户 id	" 当代小说 " 分值	...	" 英语考研 " 分值	...	" 华语流行 " 分值
115521	15	...	8.9	...	12.2
...	...	...	...	...	...
127485	21.4	...	14	...	11
66885	11	...	10.2	...	9.6
...	...	...	...	...	...
205894	17.4	...	10.6	...	10

通过修正的余弦相似性来查找用户 66885 的最近邻集合 ,按相似性由高到低排列得到 {44675 , 205894 , 88721 , 134589 , 5265 } ,再根据这 5 个客户所购买的商品去除 66885 本身购买的商品 ,最后按出现频率高低排列的有如下 5 种商品 :功夫(DVD) ,大内密探零零发(DVD) ,鹿鼎记(DVD) ,唐伯虎点秋香(DVD) ,望夫成龙(DVD) 。可以看到都是周星驰的作品 ,而客户自己更是购买了周星驰主演的其它电影 :月光宝盒、仙履奇缘、逃学威龙系列、算死草等。

最后随机的取出 1000 个用户 ,运用该方法以及传统的协同过滤方法来对数据进行验证 ,主要分析后 3 个月中对于该用户被推荐商品的购买情况 ,设购买率 = $\frac{\text{购买商品数}}{\text{被推荐的总商品数}}$,发现当推荐数为 5 时 ,推荐

效果最佳,得出结果如下:

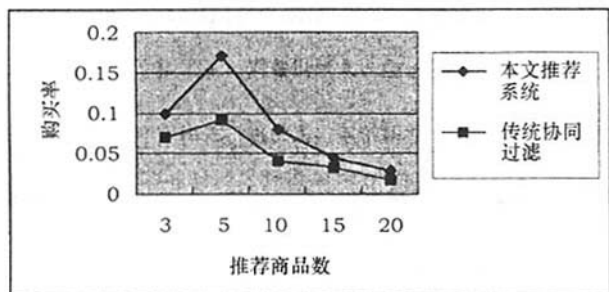


图2 推荐效果图

四、结束语

推荐系统目前已运用于不少商务网站,但是运用的结果较差,主要是:数据本身问题,多而杂的数据集难以通过简单方法发现有用规则^[9];一般是基于商品的推荐而没有考虑到客户个性化的需求,如近几天销售最多的商品排行。导致不能满足用户意愿,不能吸引用户购买商品。

本文针对协同过滤的缺点所在的矩阵稀疏性以及冷开始问题^[10],提出一种新的数据收集以及评分设定,考虑购买以及浏览的两部分数据,并且对于不同情况予以不同的评分设定,从而能够较好地弥补这两方面的不足,试验也证明效果显著。

未来的研究方向可以从以下方面展开:一是提高系统效率。在网站购物过程中,用户的任一步等待时间过长,都可能会导致用户的直接退出;二是将推荐结果进行多方面应用,在用户浏览网页时在线推荐以

及线下的 email 推荐等;三是将推荐结果与营销手段结合,实现真正的一对一营销。

参考文献

- [1] RUD O P. 数据挖掘实践[M]. 朱扬勇,左子叶,张忠平等,译[M]. 北京:机械工业出版社, 2003. 266 - 275.
- [2] LIN W Y, ALVAREZ S A, RUIZ C. Collaborative recommendation via adaptive association rule mining[C]. International Conference on Knowledge Discovery and Data Mining, 2000.
- [3] DEMIRIZ. A enhancing Product Recommender Systems on Sparse Binary Data[EB/OL]. <http://www.rpi.edu/demira/research.htm>, 2003.
- [4] BADNUL M S, GEORGE K, JOSEPH A K et al. Application of dimensionality reduction in recommender system - a case study[C]. In Proc of the WebKDD 2000 Workshop at the ACM SIGKDD2000, Boston, 2000.
- [5] MOBASHER B, DAI H, LUO T et al. Discovery of aggregate usage profiles for web personalization[C]. In Proc of the WebKDD 2000 Workshop at the ACM SIGKDD2000, Boston, 2000.
- [6] 王实, 高文, 李锦涛. 基于分类方法的 Web 站点实时个性化推荐[J]. 计算机学报, 2002, 25(8): 845 - 852.
- [7] 邢东山, 沈钧毅. 基于 Web 日志的因特网协作推荐系统的研究[J]. 西安交通大学学报, 2002, 36(12): 1271 - 1274.
- [8] 冯是聪. 基于 Web 挖掘的个性化技术研究[J]. 计算机工程与设计, 2004, 24(1): 4 - 6.
- [9] 邓爱林. 电子商务推荐系统介绍[EB/OL]. <http://www.dmgroun.org.cn/ppt11.ppt>, 2004.
- [10] 黄光球, 靳峰, 彭绪友. 基于兴趣度的协同过滤商品推荐系统模型[J]. 微电子学与计算机, 2005, 22(3): 5 - 8.

Collaborative Filtering Recommendation System Based on Web Log and Commodities Classification

CUI Ya-zhou DUAN Gang

(Univ. of Elec. Sci. & Tech. of China Chengdu 610054 China)

Abstract Traditional collaborative filtering algorithm generally focuses on how the user rates the commodities. Considering that it's difficult to get the explicit rating data, this article takes the data source from the server log files of the users' browsing recording and the customers' purchasing history data, designs the rating matrix according to the classification of goods, and obtains the recent neighbors from the matrix, and consequently gives the personalization recommendation. The experiment shows that the approach is successful.

Key Words collaborative filtering; sever log; personalization recommendation

作者: 崔亚洲, 段刚, CUI Ya-zhou, DUAN Gang
作者单位: 电子科技大学, 成都, 610054
刊名: 电子科技大学学报(社会科学版)
英文刊名: JOURNAL OF UNIVERSITY OF ELECTRONIC SCIENCE AND TECHNOLOGY OF CHINA(SOCIAL SCIENCES EDITION)
年, 卷(期): 2006, 8(3)
被引用次数: 2次

参考文献(10条)

1. RUD O P;朱扬勇;左子叶;张忠平 [数据挖掘实践](#) 2003
2. LIN W Y;ALVAREZ S A;RUIZ C [Collaborative recommendation via adaptive association rule mining](#) 2000
3. DEMIRIZ A [enhancing Product Recommender Systems on Sparse Binary Data](#) 2003
4. BADNUL M S;GEORGE K;JOSEPH A K [Application of dimensionality reduction in recommender system-a case study](#) 2000
5. MOBASHER B;DAI H;LUO T [Discovery of aggregate usage profiles for web personalization](#) 2000
6. 王实;高文;李锦涛 [基于分类方法的Web站点实时个性化推荐](#)[期刊论文]-[计算机学报](#) 2002(08)
7. 邢东山;沈钧毅 [基于Web日志的因特网协作推荐系统的研究](#)[期刊论文]-[西安交通大学学报](#) 2002(12)
8. 冯是聪 [基于Web挖掘的个性化技术研究](#)[期刊论文]-[计算机工程与设计](#) 2004(01)
9. 邓爱林 [电子商务推荐系统介绍](#) 2004
10. 黄光球;靳峰;彭绪友 [基于兴趣度的协同过滤商品推荐系统模型](#)[期刊论文]-[微电子学与计算机](#) 2005(03)

本文读者也读过(7条)

1. 侯亚丽. 袁方. HOU Ya-li. YUAN Fang [Web日志挖掘中的数据预处理技术](#)[期刊论文]-[河北大学学报\(自然科学版\)](#) 2005, 25(2)
2. 张海强. 胡学龙. ZHANG Hai-qiang. HU Xue-long [一种基于引用日志文件的启发式会话识别算法](#)[期刊论文]-[扬州大学学报\(自然科学版\)](#) 2007, 10(3)
3. 欧阳一鸣. 汪曦东. 郭骏. 刘红樱. Ouyang Yiming. Wang Xidong. Guo Jun. Liu Hongying [Web使用挖掘数据预处理中的会话构造](#)[期刊论文]-[计算机工程与应用](#)2005, 41(25)
4. 邢东山. 沈钧毅 [基于Web日志的因特网协作推荐系统的研究](#)[期刊论文]-[西安交通大学学报](#)2002, 36(12)
5. 柳胜国 [Web日志挖掘数据预处理方法研究](#)[期刊论文]-[现代图书情报技术](#)2004(12)
6. 吴强. 梁继民. 杨万海 [Web日志挖掘预处理中的用户识别技术](#)[期刊论文]-[计算机科学](#)2002, 29(4)
7. 陈钰. 张功亮. 阚述贤. 周茜. 赵晓飞. 朱悠悠. Chen Yu. Zhang Gongliang. Kan Shuxian. Zhou Xi. Zhao Xiaofei. Zhu Youyou [一种基于领域本体的用户建模方法](#)[期刊论文]-[计算机与数字工程](#)2011, 39(2)

引证文献(2条)

1. 王玉祥. 乔秀全. 李晓峰. 孟洛明 [上下文感知的移动社交网络服务选择机制研究](#)[期刊论文]-[计算机学报](#) 2010(11)
2. 白娟. 周官志 [基于C/S架构的第三方个性化推荐系统构建](#)[期刊论文]-[商业时代](#) 2011(26)

本文链接: http://d.wanfangdata.com.cn/Periodical_dzkjdxxb-shkx200603010.aspx