

# 基于智能体的电子商务协作推荐系统

何 波, 陈 媛, 王华秋, 董世都

(重庆工学院计算机科学与工程学院, 重庆 400050)

**摘 要:** 传统的协作筛方法存在的主要问题是需要人为地提供评价, 该文对其进行了改进, 根据用户事务模式自动获取用户评价, 构建评价矩阵。针对现有个性化信息推荐系统存在的问题, 将 Agent 应用到电子商务个性化信息推荐, 引入设计的协作筛方法, 设计了基于 Agent 的电子商务协作推荐系统(ECCRS)。ECCRS 是基于服务器端的电子商务协作推荐系统, 它考虑了用户访问页面的时间特性, 不需要用户注册, 在推荐时考虑了页面的新颖性, 同时实现了离线处理、增量更新与在线推荐的结合。实验结果表明, ECCRS 采用的推荐方法是有效的。

**关键词:** 电子商务; 智能体; 协作推荐; 个性化主动信息服务

## E-commerce Collaborative Recommendation System Based on Agent

HE Bo, CHEN Yuan, WANG Huaqiu, DONG Shidu

(School of Computer Science and Engineering, Chongqing Institute of Technology, Chongqing 400050)

**【Abstract】** The paper improves the traditional collaborative filtering method which needs artificial evaluation. The advanced method can gain evaluation value automatically through user transaction pattern and establish evaluation matrix. Aiming at problems of current personalized information recommendation systems, this paper applies Agent and designs collaborative filtering method to E-commerce personalized information recommendation, and designed E-commerce collaborative recommendation system based on Agent, namely, ECCRS, which is based on server. ECCRS considers time of page calling and novelty of page. It can supply recommendation service for unregistered users and combined off-line processing, incremental updating and on-line recommendation. The experimental results indicate that the recommendation method of ECCRS is feasible.

**【Key words】** E-commerce; Agent; Collaborative recommendation; Personalized active information service

随着电子商务的迅猛发展, 电子商务用户的个性化服务成为一个重要的研究课题。在智能个性化主动信息服务中最重要的服务是个性化信息推荐。目前已有的一些电子商务信息推荐系统, 比较有代表性的有: WebWatcher 系统<sup>[2]</sup>, SiteHelper 系统<sup>[3]</sup>和Footprints 系统<sup>[4]</sup>等, 这些系统能够实现对用户的信息推荐, 但是仍然存在一些问题, 主要包括:

(1) 多数电子商务信息推荐系统没有考虑利用用户访问页面的时间特性。而时间特性是一种可利用的重要资源, 即如果用户对某页面感兴趣, 那么就会在该页面停留较长时间, 而如果用户对某页面不感兴趣, 那么就会停留较短的时间;

(2) 多数电子商务信息推荐系统针对的是注册用户, 较少考虑非注册用户的个性化信息推荐;

(3) 多数电子商务信息推荐系统没有考虑用户是否有新颖信息需求的偏好。

### 1 ECCRS 系统框图

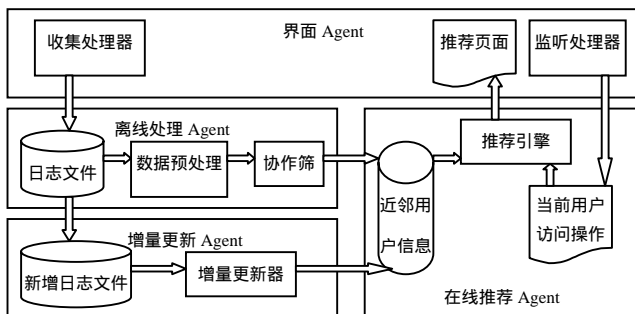


图 1 ECCRS 系统框图

ECCRS 是一个基于服务器端的电子商务协作推荐系统, 它考虑了用户访问页面的时间特性, 不需要用户注册, 在推荐时考虑了页面的新颖性, 同时实现了离线处理、增量更新与在线推荐的结合。ECCRS 包括界面 Agent、离线处理 Agent、增量更新 Agent 和在线推荐 Agent 几个部分, 其系统框图如图 1 所示。

### 2 ECCRS 系统结构

#### 2.1 界面 Agent

界面 Agent 是用户与系统交互的接口, 主要包括收集处理器和监听处理器。收集处理器的主要任务是收集用户的访问记录并将其加入日志文件。监听处理器的主要任务是监听当前用户访问操作序列, 通常采用 URL 重写、暂存 Web 服务器的日志文件等方法来进行监听, 获取当前用户访问操作序列。在线推荐 Agent 产生的推荐页面集也通过界面 Agent 提供给用户。

#### 2.2 离线处理 Agent

离线处理 Agent 主要包括数据预处理器和协作筛。数据预处理器的主要任务是对服务器的日志文件进行数据预处理

**基金项目:** 重庆市科委科技攻关基金资助项目(CSTC.2004AA2001-8277-9); 重庆工学院青年基金资助项目

**作者简介:** 何 波(1978 - ), 男, 硕士、讲师, 主研方向: 信息推荐, 信息安全, 数据挖掘; 陈 媛, 硕士、副教授; 王华秋, 博士生、讲师; 董世都, 硕士、讲师

**收稿日期:** 2006-06-14 **E-mail:** hebo@cqit.edu.cn

理。数据预处理的主要步骤包括：数据净化，用户识别，用户访问操作识别，路径完善和用户事务模式识别等。经过数据预处理后，可获得用户事务模式，并且生成每个用户的事务模式文件，每一个用户事务模式文件包含了若干个用户事务模式。由于 ECCRS 要考虑用户访问页面的时间特性，所以得到的是带有时长的用户事务模式。带有时长的用户事务模式如定义 1 所示。

设  $L$  为用户访问操作集合，每一条访问记录  $l \in L$  包括：用户主机地址  $l.ip$ ，用户 ID 号  $l.uid$ ，被访问 Web 页 URL 地址  $l.url$ ，访问时间戳  $l.time$  和访问 Web 页的时长  $l.timelength$ 。

**定义 1** 带有时长的用户事务模式定义为三元组，形式化表示如下：

$$t = \langle ip, uid, URL \rangle; URL = \{(l_1^t.url, l_1^t.time, l_1^t.timelength), \dots, (l_m^t.url, l_m^t.time, l_m^t.timelength)\}$$

$$\text{其中：} l_k^t \in L, ip_k^t = l_k^t.ip, uid_k^t = l_k^t.uid, 1 \leq k \leq m; l_k^t.timelength = l_k^t.time - l_{k-1}^t.time, 1 \leq k \leq m-1.$$

协作筛<sup>[5]</sup>是一种共享用户间的兴趣和评价的重要技术，采用的方法是收集用户对项目的兴趣偏好，匹配出具有相似兴趣偏好的用户，对这些相似用户共享彼此的兴趣和评价。它能够为用户提供协作推荐。

论文设计的协作筛方法主要有 3 个步骤，描述如下：

#### (1) 自动获取用户评价

传统的协作筛方法的一个主要问题是需要人为地提供评价。而论文设计的协作筛方法对其进行了改进，根据用户事务模式自动获取用户评价，构建评价矩阵。用户对一个页面的评价由用户访问页面的次数、相对时长、访问新度和重视程度多个因素决定，这些因素由定义 2~定义 4 分别表示。用户对一个页面的评价由定义 5 表示，评价矩阵的由定义 6 表示。

**定义 2** 一个用户  $u$  访问一个页面 WebPage 的次数  $Count(u, WebPage)$  定义为：在一个固定的时间段  $T$  内，一个用户访问 WebPage 的次数。为了去掉那些用户偶然访问的页面，要求  $Count(u, WebPage) \geq 2$ 。

**定义 3** 一个用户  $u$  访问一个页面 WebPage 的相对时长  $TimeLength(u, WebPage)$  用式(1)表示。

$$TimeLength(u, WebPage) = (TotalTimeLength(u, WebPage) / Size(u, WebPage)) / \max(TotalTimeLength(u, WebPage) / Size(u, WebPage)) \quad (1)$$

其中  $TotalTimeLength(u, WebPage)$  为用户  $u$  访问 WebPage 的时长之和； $Size(u, WebPage)$  为用户访问 WebPage 的次数。而  $\max(TotalTimeLength(u, WebPage) / Size(u, WebPage))$  为用户  $u$  访问过的 WebPage 中的最大值。

$TimeLength(u, WebPage)$  是一个相对值，表征一个用户对 WebPage 的访问时间偏好特性。

**定义 4** 一个用户  $u$  对一个页面 WebPage 的访问新度  $new(u, WebPage)$  用式(2)表示。

$$new(u, WebPage) = \frac{\sum_{i=1}^{Count(u, WebPage)} (Time_i(u, WebPage) - Time(StarLog))}{Count(u, WebPage)} \quad (2)$$

其中， $Time_i(u, WebPage)$  表示用户  $u$  在固定的时间段  $T$  内第  $i$  次访问 WebPage 的时刻， $Time(StarLog)$  表示开始记录 Log 的时刻。

$new(u, WebPage)$  是一个相对值，表征一个用户对

WebPage 的访问时间的鲜程度，如果一个用户最近频繁访问 WebPage，则该值较大，否则较小。

**定义 5** 一个用户  $u$  对一个页面 WebPage 的评价值  $Evaluation(u, WebPage)$  用式(3)表示。

$$Evaluation(u, WebPage) = (\log_2 Count(u, WebPage)) \times (TimeLength(u, WebPage) + new(u, WebPage)) \quad (3)$$

**定义 6** 评价矩阵是各个用户对所有页面评价值的  $p \times q$  二维矩阵，其中  $p$  为用户数， $q$  为所有页面的数量；每个列表示多个用户对一个页面的评价，每个行表示一个用户对所有页面的评价，每个行列交叉的元素为一个用户对特定页面的评价。

如果一个用户没有访问某个页面 WebPage，则该用户对 WebPage 的评价为 0。

#### (2) 计算用户间的相似度

对于评价矩阵  $r_{p \times q}$ ，它的每一个列为多个用户对一个页面 WebPage 的评价值(共  $q$  个 WebPage)，它的每一行为一个用户对所有 WebPage 的评价值(共有  $p$  个用户)，那么用户  $a$  和用户  $u$  的相似性采用 Pearson 相关性系数来计算，用式(4)表示。

$$similar(u, a) = \frac{\sum_{i=1}^q ((r_{a,i} - \bar{r}_a) \times (r_{u,i} - \bar{r}_u))}{\sqrt{\sum_{i=1}^q (r_{a,i} - \bar{r}_a)^2 \times \sum_{i=1}^q (r_{u,i} - \bar{r}_u)^2}} \quad (4)$$

其中， $r_{a,i}$  表示评价矩阵中用户  $a$  所在行的第  $i$  列评价值； $\bar{r}_a$  表示评价矩阵中用户  $a$  所在行评价值的均值，满足  $-1 \leq similar(u, a) \leq 1$ 。

Pearson 相关性系数从一个线性回归模式导出，依赖一些假定，如关系必须是线性的。

#### (3) 选择近邻用户

当计算出一个用户与其他所有用户的相似度后，选择一些具有较大相似度的用户作为该用户的近邻用户。常用的选择近邻用户的方法有两种：

1) 阈值法：设置一个用户相似度阈值，选择与一个用户的相似度大于或等于阈值的用户作为近邻用户。设置高的阈值，将得到相似度较大的近邻用户，但是也往往导致近邻用户很少或没有，不利于协作推荐。

2) 最优的  $k$  个近邻法：从大到小选择  $k$  个与用户相似度较大的用户作为近邻用户。如果  $k$  值选择过大，会造成某些近邻用户与该用户的相似度过小，在协作推荐时引入过多噪音。

论文采用阈值可调的阈值法或  $k$  值可调的最优  $k$  个近邻法来选择近邻用户，可以得到较为准确的近邻用户。

### 2.3 增量更新 Agent

增量更新 Agent 主要包括增量更新器。增量更新器的任务是实现近邻用户信息的增量更新。采用的方法是：根据时间戳，将新的用户事务模式导入新增日志文件，对新增日志文件进行数据预处理，增量更新用户事务模式，在此基础上，实现近邻用户信息的增量更新。

### 2.4 在线推荐 Agent

在线推荐 Agent 主要包括推荐引擎。推荐引擎的任务是根据离线处理 Agent 和增量更新 Agent 得到的近邻用户信息和界面 Agent 监听到的当前用户访问操作序列进行在线推荐。

在收集到当前用户访问操作序列以后，需要评估近邻用户访问过的所有页面，选择适合的页面作为推荐项。评估候

选页面需要考虑 3 个因素：

(1)当前用户与候选页面对应用户的相似度。它反映了当前用户和其他用户的相似程度，是进行协作推荐的最重要因素，相似度越大越应该优先考虑。根据离线处理和增量更新部分得到了用户间的相似度，就可直接获取当前用户与候选页面对应用户的相似度。

(2)近邻用户对候选页面的评价。它反映了近邻用户对候选页面的评价情况，由用户访问页面的次数、相对时长、访问新度和重视程度多个因素决定，是进行协作推荐的重要因素，评价越大越应该优先考虑。根据离线处理和增量更新得到了评价矩阵，就可直接获取近邻用户对候选页面的评价。

(3)候选页面与当前用户访问操作序列的物理距离。从新颖性的角度看，推荐一个物理链接上远离当前用户访问操作序列的页面是优先考虑的对象。论文定义链接距离因子表示这种选择策略，如定义 7 所示。

**定义 7** 链接距离因子：给定页面拓扑有向图  $G=(V,E)$ ，设  $r$  为当前用户访问操作序列， $v \notin r$  为当前的一个候选页面； $dist(v,r,G)$ 表示  $v$  到  $r$  中的 URL 之间的最小物理链接路径距离， $v$  关于  $r$  链接距离因子计算公式为

$$idf(v,r)=\log(dist(v,r,G))+1 \quad (5)$$

如果  $r$  属于当前用户访问操作序列，则定义链接距离因子为 0。对最小物理链接路径距离取对数计算，目的是让链接距离因子不要过度影响推荐值。

设  $u$  为当前访问用户， $r$  为当前用户访问操作序列， $a$  为  $u$  的一个近邻用户， $v \notin r$  为  $a$  访问过的一个候选页面，根据评估候选页面需要考虑的 3 个因素，候选页面  $v$  的推荐值计算公式如下：

$$Rec(r,v)=\left(\sqrt{Evaluation(a,v) \times idf(v,r)}\right) \times similar(u,a) \quad (6)$$

如果  $v$  属于当前用户访问操作序列，则链接距离因子为 0，因而其推荐值也为 0。设定最小推荐值阈值，对所有候选页面都计算推荐值，将大于或等于 的候选页面作为推荐项，由所有推荐项构成的集合称为推荐集。推荐集由下式表示：

$$Recommend(u,r)=\{v \mid Rec(r,v) \geq \theta\} \quad (7)$$

### 3 系统实现及实验

实验的平台是 DELL 服务器，其配置为 P4 2.4GHz，内存 512MB，Windows2000 Server 操作系统，SQL Server 2000

(上接第 195 页)

和可见光图像的融合结果，其中图 1(a)红外图像，图 1(b)为可见光图像，图 1(c)为融合图像。通过目测法可以看到，融合图像将红外目标和可见光背景有效地结合起来，充分利用红外波段和可见光波段的信息互补性，达到便于观察的目的。

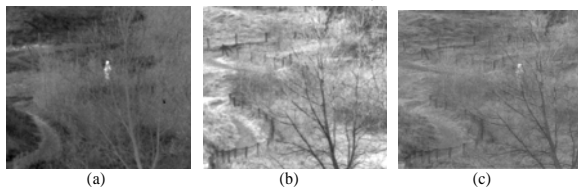


图 1 基于双正交小波的融合结果

### 5 结论

本文利用双正交小波的对称性和良好的滤波性能构造出对称/反对称双正交矩阵滤波器，对其经过平衡处理后，结合 Burt 融合方案，对可见光图像和红外图像进行了融合实验，

数据库管理系统。编程工具是 C++ Builder5.5。实验数据取自建立站点的访问 Log 文件，总的的数据量是一个月的访问记录。经过预处理，获得大约 800 个用户事务模式。

对在线推荐方法进行实验，得到的结果如图 2 所示。

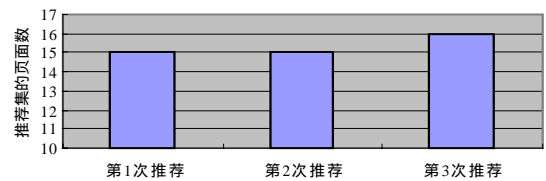


图 2 在线推荐方法的实验

从图 2 可以看出，随着用户访问过程的深入，即推荐过程的深入，推荐集的页面数没有减少，反而还稍微增加。推荐集页面数的增加意味着推荐范围的拓宽，增加了用户选择新信息的可能性。这说明 ECCRS 给出的在线推荐方法能够为用户提供较多的推荐结果，适合有新颖性信息需求的用户，ECCRS 采用的在线推荐方法是有效的。

### 4 结束语

Web 个性化信息推荐日益成为一个重要的研究课题。ECCRS 有效地解决了已有个性化信息推荐系统存在的一些问题，实现了离线处理、增量更新与在线推荐的结合。

#### 参考文献

- 1 周立柱, 邢春晓, 曾春. 个性化服务技术综述[J]. 软件学报, 2002, 13(10): 1952-1961.
- 2 Joachims T, Freitag D, Mitchell T. Web Watcher: A Tour Guide for the World Wide Web[C]//Proceedings of the 15<sup>th</sup> International Joint Conference on Artificial Intelligence. Morgan Kaufmann Publishers, 1997.
- 3 Ngu D, Wu X. SiteHelper: A Localized Agent that Helps Incremental Exploration of the World Wide Web[C]//Proceedings of the 6<sup>th</sup> International World Wide Web Conference, Santa Clara, California, USA. 1997.
- 4 Alan W, Maes P. Using History to Assist Information Browsing[C]//Proc. of RIAO'97: Computer-assisted Information Retrieval on the Internet, Montreal. 1997.
- 5 王实, 高文, 杜建平, 等. 基于代理服务器的协作浏览[J]. 计算机科学, 2002, 29(2): 40-42.

结果表明该方法能够获得较好的融合效果。

#### 参考文献

- 1 Tan H H, Shen L X, Tham J W. New Biorthogonal Multiwavelets for Image Compression[J]. Signal Processing, 1999, 79(1): 45-65.
- 2 Hardin D P, Marasovich J A. Biorthogonal Multiwavelets on  $[-1,1]$ [J]. Appl. and Comp. Harm, 1999, 7(1): 34-53.
- 3 Lebrun J, Vetterli M. Balanced Multiwavelets: Theory and Design[J]. IEEE Trans. on Signal Processing, 1998, 46(4): 1119-1125.
- 4 Burt P, Kolczynski R. Enhanced Image Capture Through Fusion[C]//Proc. of the 4<sup>th</sup> Int. Conf. on Computer Vision, Berlin, Germany. 1993: 173-182.
- 5 Burt P J, Adelson E H. The Laplacian Pyramid as a Compact Image Code[J]. IEEE Transactions on Communications, 1983, 31(4): 532-540.