基于密度的动态协同过滤图书推荐算法*

武建伟[a]，俞晓红[b]，陈文清[c]
（洛阳理工学院 a. 现代教育技术中心；b. 数理部；c. 电气工程与自动化系，河南 洛阳 471023）

摘　要： 针对协同过滤推荐技术在个性化服务应用中存在的服务质量和服务效率问题，提出一种基于密度的动态协同过滤图书推荐算法。在对读者的图书流通记录进行兴趣度模糊筛选基础上，利用扩展的密度聚类算法进行区域聚类，读者的兴趣模型依据聚类区域的密度与权重变化更新，动态进行协同过滤图书推荐。实验表明，该算法在提高推荐精确度上，优于传统的协同过滤推荐算法。

关键词： 协同过滤；个性化推荐；动态；相似度
中图分类号： TP391　　**文献标志码：** A　　**文章编号：** 1001-3695(2010)08-3013-03
doi:10.3969/j.issn.1001-3695.2010.08.053

Density-based dynamic collaborative filtering books recommendation algorithm

WU Jian-wei[a], YU Xiao-hong[b], CHEN Wen-qing[c]
(*a. Modern Education Technology Center, b. Dept. of Mathematics & Physics, c. Dept. of Electric Engineering & Automation, Luoyang Institute of Science & Technology, Luoyang Henan* 471023, *China*)

Abstract: In view of the problem of service quality and efficiency about the application of collaborative filtering recommendation technology to personalized service, this paper proposed a density-based dynamic collaborative filtering books recommendation algorithm. Based on the fuzzy filtering of reader's interest by the reader's books circulation records, the region clustering was done using the extended density clustering algorithm and the reader's interest model was renewed according to the variation of density and weight in clustering region, then achieved the dynamic collaborative filtering books recommendation. Experiments show that the algorithm is better than the traditional collaborative filtering algorithms in improving the recommendation accuracy.

Key words: collaborative filtering; personalized recommendation; dynamic; similarity

随着信息技术发展，数字图书馆的建设已经从基于信息资源的数字化进入了信息整合、服务个性化的发展阶段。个性化服务通过收集和分析用户信息来学习用户的兴趣和行为，并挖掘用户隐藏的兴趣和用户群体的行为规律，从而制定相应的信息过滤策略，为用户提供个性化的主动推荐服务。充分运用个性化服务技术能很好地解决信息过载问题，提高数字图书馆的服务质量和服务效率。

推荐系统是个性化服务中最重要的技术之一，协同过滤推荐是当前应用最成功的技术[1]，其基本思想是基于用户对资源的历史评价，依据评价相似的最近邻居的评分数据向目标用户进行推荐。协同过滤推荐算法主要有基于用户的协同过滤推荐和基于资源的协同过滤推荐[2~6]。数字图书馆用户的流通记录，在中图法分类树中分布区域，反映了用户的兴趣偏好。区域密度及区域权重大代表用户的长期兴趣，区域密度及区域权重小代表用户的短期兴趣。随着区域密度及其权重变化，用户的兴趣会发生漂移。本文提出一种基于密度的动态协同过滤推荐算法，首先根据图书流通记录的归还时间，利用模糊隶属函数筛选读者对借阅图书的兴趣度，扩展基于密度的 DBSCAN 算法进行区域聚类[7]，建立用户的兴趣模型，随着用户兴趣聚类区域密度及权重变化，进行动态的协同过滤推荐。

1　基于密度的动态协同过滤推荐算法

1.1　数据筛选

图书的流通记录从借阅、归还、续借三个方面反馈读者的行为。借阅：读者的借阅图书信息，反映读者的读书类别倾向。归还：读者借阅图书归还时间，反映读者对该书的态度或兴趣。续借：续借的书一定是读者非常感兴趣的。本文根据读者图书归还时间，利用模糊理论的隶属函数来计算读者对借阅的图书感兴趣的程度，并筛选掉读者不感兴趣的借阅记录。

定义 1　图书借阅。读者借阅记录中所借阅图书的集合，读者借阅一本图书，表明对该图书具有兴趣的倾向意图。

定义 2　图书归还。图书归还时间反映了读者对图书感兴趣程度。如果刚刚借阅即归还，表明对该图书不是非常感兴趣，甚至是不感兴趣；如果图书被续借，表明对该图书非常感兴趣。定义读者图书归还集：$R_{\text{time}}=\{r_1,r_2,\cdots,r_n\}$。其中 $r_i=\dfrac{\text{return}_{\text{time}(i)}-\text{borrow}_{\text{time}(i)}}{T}$，$\text{borrow}_{\text{time}(i)}$、$\text{return}_{\text{time}(i)}$ 分别为读者某个借阅记录图书借阅的时间和归还时间，T 为图书借阅规定还书周期。

定义 3　用隶属函数 u_{like} 和 u_{dislike} 分别表示感兴趣与不感

收稿日期：2010-01-20；修回日期：2010-03-01　　基金项目：河南省教育厅自然科学基金资助项目(2008A520017)；洛阳理工学院青年基金资助项目(2009QZ27)

作者简介：武建伟(1976-)，男，河南洛阳人，讲师，主要研究方向为基于 Web 人工智能(wjw@lit.edu.cn)；俞晓红(1977-)，女，河南洛阳人，讲师，硕士，主要研究方向为非线性控制；陈文清(1970-)，男，河南信阳人，副教授，博士研究生，主要研究方向为人工智能与控制。

兴趣的模糊程度，$f_{\text{like}}(r_i)$ 为读者基于定义 2 对图书的感兴趣与不感兴趣的模糊值。梯形函数表示如图 1 所示。

图中 a 和 c 是隶属函数 u_{like} 和 u_{dislike} 的界定参数值。图书归还时间区 $r_a - r_{\min}$ 用来筛选出不感兴趣的书目，归还时间区 $r_{\max} - r_c$ 筛选出非常感兴趣的书目。如果借阅时间超过规定归还周期 T，则该图书的借阅时间信息无效，以最大值取代。则 $f_{\text{like}}(r_i)$ 定义如下：

$$f_{\text{like}}(r_i) = \begin{cases} 1 & \text{if } c < r_i \leqslant r_{\max} \\ (r_i - a)/(c - a) & \text{if } a \leqslant r_i \leqslant r_{\max} \\ 0 & \text{if } r_{\min} \leqslant r_i < a \\ \perp & \text{if } r_i < d_{\min} \end{cases} \tag{1}$$

1.2 兴趣模型建立

DBSCAN(density-based spatial clustering of applications with noise) 是一种基于密度的典型聚类算法，对于聚类中的每个对象，在给定的半径 *Eps* 领域中至少要包含最小数目 MinPts 个对象。引入密度可达的概念，一个簇是基于密度可达性的最大的密度相连对象的集合。不包含在任何簇中的对象被认为是“噪声”[8~10]。本文结合中图法分类，扩展 DBSCAN 算法，对读者的借阅记录在中图法分类中的分布进行区域聚合，建立读者的兴趣模型。

定义 4 区域质心(centroid)。设聚类区域 CR(clustering region) 空间中有 m 个数据点 $p_i(i = 1,2,\cdots,m)$，其各点到中图法分类树树根距离的均值为该区域空间的质心，则

$$\text{centroid}_{\text{CR}} = \frac{1}{m}\sum_{i=1}^{m}\text{distance}(p_i) \tag{2}$$

其中：$\text{distance}(p_i)$ 为点 p_i 到分类树树根欧氏距离。

定义 5 区域最大距离(Md) 设聚类区域 CR 空间中质心点与最远点的距离为 Md_{CR}，则

$$Md_{\text{CR}} = \max(\text{distance}(\text{centroid}_{\text{CR}}, p_i)) \tag{3}$$

定义 6 区域权重(α)。设聚类区域 CR 中所含读者借阅图书记录为 m，读者的所有图书借阅记录为 n，则区域权重 α 为

$$\alpha_{\text{CR}} = \frac{m \times \sum_{i=1}^{m} f_{\text{like}}(p_i)}{n \times \sum_{j=1}^{n} f_{\text{like}}(p_j)} \tag{4}$$

定义 7 区域密度(Den)。

$$\text{Den}_{\text{CR}} = \frac{\max(\text{num}(C_{\text{CR}}))}{\frac{1}{k}\sum_{i=1}^{k}\text{num}(C_{\text{CR}}(i))} \tag{5}$$

其中：k 为聚类区域 CR 所含图书类别 C_{CR} 的个数；$\text{num}(C_{\text{CR}(i)})$ 为类别中图书的个数；Den_{CR} 反映区域中节点分布。

定义 8 区域相邻系数(ε)。相邻两个区域空间质心点之间的距离与两个区域空间的区域最大距离之和的比值。即

$$\varepsilon_{\text{CR}(x),\text{CR}(y)} = \frac{\text{distance}(\text{centroid}_{\text{CR}(x)}, \text{centroid}_{\text{CR}(y)})}{Md_{\text{CR}(x)} + Md_{\text{CR}(y)}} \tag{6}$$

其中：$\text{distance}(\text{centroid}_{\text{CR}(x)}, \text{centroid}_{\text{CR}(y)})$ 为两个区域空间质心的欧式距离。$\varepsilon_{\text{CR}(x),\text{CR}(y)}$ 值越小，说明两相邻区域空间数据点的密度越大。

定义 9 兴趣模型。读者兴趣模型由聚类区域质心、聚类区域权重、聚类区域密度三个方面的特征共同组成。读者兴趣模型可以表示为

$$U = \{\text{centroid}_{\text{CR}(i)}, \alpha_{\text{CR}(i)}, \text{Den}_{\text{CR}(i)}\} \tag{7}$$

其中：$i = \{1,2,\cdots,l\}$ 为读者借阅记录依据中图法分类树满足参数 Eps 和 MinPts 的聚类区域。

定义 10 读者聚类。依据读者的兴趣模型，采用最近共享邻居节点聚类算法(SNN) 进行读者聚类[11]。SNN 算法是利用节点间最近共享邻居节点的个数作为相似度。读者之间兴趣模型中，两两计算聚类区域质心之间的距离，所有距离相近的两个聚类区域共享的记录个数之和作为两个读者之间的相似度。

$$\text{sim}(U, U') = \sum_{j=1}^{h}\text{num}(nn(U_{\text{CR}}(j)) \cap nn(U'_{\text{CR}}(j))) \tag{8}$$

其中：$nn(U_{\text{CR}}(j))$ 和 $nn(U'_{\text{CR}}(j))$ 分别为读者兴趣模型中一个聚类区域的节点集，h 为读者间距离相近的聚类区域个数。

1.3 推荐估值

读者借阅记录的聚类区域反映读者的兴趣特征，兴趣在聚类区域应该是渐变的，本文采用距离平方反比法进行推荐。距离平方反比是利用邻近记录的平方值来决定相关权值，距离越近者，权重值越大。根据距离点 p_0 的最近质心所属的邻居聚类区域，点 p_0 的推荐估值定义为

$$f_{\text{like}}(p_0) = \alpha_{\text{CR}(i)} \times \sum_{j=1}^{x}\beta_{\text{CR}(i),j} \times f_{\text{like}}(p_j) \tag{9}$$

其中：$\alpha_{\text{CR}(i)}$ 为点 p_0 最近邻聚类区域的区域权重；权值 $\beta_{\text{CR}(i),j} = f(d_{0,j}) / \sum_{j=1}^{m} f(d_{0,j})$，$f(d_{0,i}) = 1/d_{0,j}^2$，$d_{0,j} = \text{distance}(p_0, p_{\text{CR}(i),j})$ 为与点 p_0 最近邻聚类区域中离点 p_0 最近 x 个点的欧式距离。

1.4 兴趣漂移

读者的兴趣会随着时间而发生变化，一些原本感兴趣的图书会被渐渐遗忘，对其失去兴趣，并产生新的兴趣。读者兴趣的渐变称为兴趣漂移。读者兴趣模型中聚类区域的密度和权重变化反映读者兴趣的漂移。兴趣漂移使得读者兴趣模型也随之变化。

规则 1 如果新的图书借阅记录落在读者兴趣模型的某个聚类区域内或外，使区域外一部分点被吸收，导致聚类区域的密度和区域权重发生变化，则更新读者的兴趣模型。

规则 2 由于规则 1 使区域相邻系数 $\varepsilon_{CR(x),CR(y)}$ 变得比较小，则两相邻区域进行合并，更新读者的兴趣模型。

规则 3 如果新的图书借阅记录落在读者兴趣模型的聚类区域外，并且产生新的聚类区域，则更新读者的兴趣模型。

规则 4 读者的兴趣应该是渐变的，兴趣相似的读者兴趣变化也应是相似的。读者的借阅行为没有符合规则 1 ~ 3 的，不进行兴趣模型更新，对相似邻居集合中兴趣模型更新最大的读者进行动态推荐。

2 实验结果及分析

2.1 实验数据集及评价标准

本文采用洛阳理工学院图书管理系统的流通日志数据，以一个学期读者借阅记录日志为单位作为训练数据集，下一个学期读者借阅记录日志作为预测集。根据本文提出的算法进行训练预测，采用平均绝对偏差(mean absolute error，MAE) 作为推荐质量度量方法，$\text{MAE} = \sum_{i=1}^{N}|p_i - q_i|/N$，$\{p_1, p_2, \cdots, p_N\}$ 为预测的读者兴趣值集合，$\{q_1, q_2, \cdots, q_N\}$ 为读者实际的兴趣值集合，MAE 越小，说明推荐算法的预测精度越高[12]。

2.2 实验结果及分析

依据本文提出的基于密度的动态协同过滤推荐算法，

计算两种方式的 MAE 值：a) 以读者的兴趣模型，基于密度的静态协同过滤推荐；b) 基于密度与权重变化，更新读者的兴趣模型，动态进行协同过滤推荐。实验过程中，目标用户的最近邻居个数从 5 增加到 26，间隔为 3。实验结果如图 2 所示。

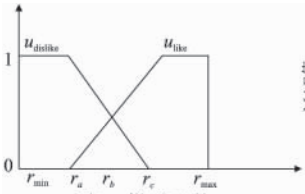

图1　梯形函数

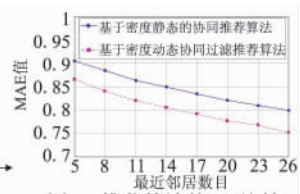

图2　推荐算法的MAE比较

从图 2 可以看出，本文提出的基于密度的动态协同过滤推荐算法的推荐精度要比静态的高。

3　结束语

随着数字图书馆图书信息资料的数量级数增长，信息过载和信息迷向问题的突显，面向用户个性化需求构建的整合将成为未来的趋势[3]。本文提出的利用读者兴趣聚类区域密度及权重变化，动态协同过滤图书推荐算法，实验表明在提高数字图书馆的服务质量和服务效率上，优于传统的协同过滤推荐算法。

参考文献：

[1] 曾春，邢春晓，周立柱. 个性化服务技术综述[J]. 软件学报，2002，13(10)：1952-1961.

[2] SARWAR B，KARYPIS G，KONSTAN J，*et al*. Item-based collaborative filtering recommendation algorithms[C]//Proc of the 10th International World Wide Web Conference. New York：ACM Press，2001：285-295.

[3] 章炯，李华. 基于资源类的时间加权协作过滤算法[J]. 计算机应用研究，2009，26(6)：2107-2109.

[4] HERLOCKER J，KONSTAN J，TERVEEN L，*et al*. Evaluating collaborative filtering recommender systems[J]. ACM Trans on Information Systems，2004，22(1)：5-53.

[5] 周军锋，汤显，郭景峰. 一种优化的协同过滤推荐算法[J]. 计算机研究与发展，2004，41(10)：1842-1847.

[6] 李春，朱珍民，叶剑，等. 个性化服务研究综述[J]. 计算机应用研究，2009，26(11)：4001-4005.

[7] ESTER M，KRIEFEL H P，SANDER J，*et al*. A density-based algorithm for discovering clusters in large spatial databases with noise[C]//SIMOUDIS E，HAN J W，FAYYAD U M. Proc of the 2nd International Conference on Knowledge Discovery and Data Mining. Portland：AAAI Press，1996：226-231.

[8] QIAN Wei-ning，GONG Xue-qing，ZHOU Ao-ying. Clustering in very large databases based on distance and density[J]. Journal of Computer Science and Technology，2003，18(1)：67-76.

[9] 周水庚，周傲英，曹晶，等. 一种基于密度的快速聚类算法[J]. 计算机研究与发展，2000，37(11)：1287-1292.

[10] 向坚持，刘相滨，资武成. 基于密度的 K-means 算法及在客户细分中的应用研究[J]. 计算机工程与应用，2008，44(35)：246-248.

[11] ERTOÖZ L，STEINBACH M，KUMAR V. Finding clusters of different sizes，shapes，and densities in noisy，high dimensional data[C]//Proc of the 2nd SIAM International Conference on Data Mining. 2003.

[12] SARWAR B M，KARYPIS G，KONSTAN J A，*et al*. Application of dimensionality reduction in recommender system：a case study[C]//Proc of ACM WebKDD 2000 Workshop. 2000.

[13] 马文峰，杜小勇. 数字资源整合的发展趋势[J]. 图书情报工作，2007，51(7)：66-70.

(上接第 3012 页)

5　结束语

本文研究政策导向型多 agent 协同系统是一种可控制的柔性的多 agent 协同系统，该系统建立在政策管理技术之上，能在动态的、分布式的环境中实现资源和服务的协同。它通过动态更新由 agent 解释的政策规则来改变它们的行为，而无须改变系统的软件配置和编码，使 agent 的行为决策变得更加规范与可控，因此具有高度的灵活性、扩展性。接下来，进一步提高实时交互的稳定性和有效解决政策冲突是下一步研究的方向。

参考文献：

[1] 廖备水，高济. PDC-agent 支持的动态自组织系统[J]. 计算机辅助设计与图形学学报，2006，18(2)：217-224.

[2] SLOMAN M. Policy driven management for distributed systems[J]. Journal of Network and Systems Management，1994，2(4)：333-360.

[3] 胡军. 面向自治计算的基于政策的多 agent 系统体系研究[D]. 杭州：浙江大学，2006.

[4] 朱从民，黄玉美，上官望义. 移动机器人 Java agent 控制系统设计[J]. 计算机工程与应用，2009，45(5)：74-77.

[5] SMITH M K，WELTY C，MCGUINESS D L. OWL Web ontology language guide[M]. [S.l.]：W3C，2004.

[6] USZOK A，BRADSHAW J M，LOTT J，*et al*. New developments in ontology-based policy management：increasing the practicality and comprehensiveness of KAoS[C]//Proc of the 9th IEEE International Workshop on Policies for Distributed Systems and Networks. 2008：145-152.

[7] DAVY S，JENNINGS B，STRASSNER J. Using an information model and associated ontology for selection of policies for conflict analysis[C]//Proc of the 9th IEEE International Workshop on Policies for Distributed Systems and Networks. 2008：82-85.

[8] 胡军，付亚军. 一种基于概念分解的政策精化方法[J]. 计算机应用研究，2009，26(5)：1650-1653.

[9] PARASHAR M，BHAT V，LIU H，*et al*. Automate：enabling autonomic applications on the grid[J]. Cluster Computing，2006，9(2)：161-174.

[10] DIGNUM V，VAZQUEZ-SZLCEDA J，DIGNUM F. Omni：introducing social structure，norms and ontologies into agent organizations[C]//Proc of PROMAS 2004. 2005：181-200.

[11] 史忠植，林芬. 主体网格智能平台 AGrIP 构建及其应用[J]. 智能系统学报，2006，1(1)：17-23.