基于用户兴趣分类的协同过滤推荐算法①

陶 俊，张 宁

(上海理工大学 管理学院，上海 200093)

摘 要：在现代信息网络中，个性化的推荐系统已经成为用户和应用软件交互的关键部分。推荐算法是个性化推荐系统的核心，其中，协同过滤算法是至今应用最为成功的推荐算法之一。但传统的协同过滤算法没有考虑用户兴趣的多样性，对用户兴趣度量不准确，难以适用于用户多兴趣的推荐系统，提出了适应用户兴趣多样性的协同过滤算法并利用改进的模糊聚类算法搜索最近邻。最后采用实际的日志数据进行算法实验，实验结果表明该算法较其他推荐算法具有较优的执行效率和推荐精度。

关键词：个性化；协同过滤算法；兴趣分类；模糊聚类

Collaborative Filtering Algorithm Based on Interest-Class

TAO Jun, ZHANG Ning

(School of Management, University of Shanghai for Science and Technology, Shanghai 200093, China)

Abstract: In the modern information network, the personalized recommendation system has become a key part of users in software application. Recommendation algorithms are the core of personalized recommendation systems. Among them, the collaborative filtering is one of the most successful recommendation algorithm in application. However, the traditional collaborative filtering algorithm does not consider user's multiple interest and measure user's interest imprecisely, and can't be applied to recommendation system with kinds of interests. In this paper, a new method of collaborative filtering algorithm based on users' interest category is proposed using improved fuzzy clustering algorithm to search the nearest neighbors. Finally, the algorithm experiment is given with actual log-data. Results show that the proposed algorithm outperforms the other recommendation ones in efficiency and recommending accuracy.

Keywords: personalization; collaborative filtering algorithm; interest classification; fuzzy clustering

随着互联网和电子通讯的飞速发展，网络中的信息量急剧上升，如何帮助用户在海量的数据中找到对其有价值的信息，指导其决策行为已成为研究者们关注的热点。现今网络系统的一个新的服务方向就是如何快速有效的推荐给用户可能感兴趣的资源。个性化的推荐系统就在这种背景下产生出来的。对于推荐系统而言，推荐算法是其核心所在。目前的推荐算法有基于内容的过滤推荐、协同过滤推荐算法、基于人口统计学的推荐算法、基于知识的推荐算法以及混合推荐算法，其中协同过滤算法是目前应用最为成功的推荐算法之一。

协同过滤这一概念首次由 Goldberg、Nicols、Oki 及 Terry[1]在 1992 年提出，应用于 Tapestry 系统，该系统适合用户群量少且要求用户给予较多的显示评价信息。Tapestry 系统奠定了协同过滤推荐研究的雏形。目前协同过滤推荐算法主要分为两类：1)基于用户的协同过滤算法 用户对项目（资源）的评分比较相似，则他们对其他项目的评分也比较相似，从而找到具有相似兴趣的最近邻，形成推荐。2)基于项目的协同过滤算法根据用户对不同项目评分的相似性来估计该用

① 基金项目:国家自然科学基金(70971089);上海市重点学科项目经费资助(S30501)
收稿时间:2010-08-24;收到修改稿时间:2010-09-26

户对某个项目的评分，以此进行推荐。

协同过滤算法主要不足有三个方面[2,3]：一是稀疏性问题，即当推荐系统中数据量很大而用户的显示评分数据又很少时，难以计算相似性，而无法推荐；二是冷启动问题，当新项目(资源)刚进入系统时，没有用户对其评价，造成协同过滤无法推荐该资源。三是可扩展性问题，推荐系统中的用户和资源会随时间快速的增长，而协同过滤算法的复杂度和数据量呈线性关系增长，严重影响了执行效率，从而导致可扩展性较差。通过分析网络日志数据，本文提出对用户兴趣分类并用数据挖掘的方法获取用户潜在的兴趣，采用改进的模糊聚类算法对用户兴趣聚类，从聚类中搜寻最近邻而形成推荐并对算法进行仿真实验。

1　基于用户的协同过滤推荐算法

基于用户的协同过滤算法是目前应用最为广泛的，算法的基本思想是使用统计方法挑选出与目标用户喜好最相似的若干用户并将其感兴趣的项目推荐给目标用户。假如目标用户对项目的评价与他的“最近邻居”相似，而目标用户对某个项目的评价可以从其“最近邻居”的评价中综合得到。该算法可分为三个阶段[4]：

1) 构建用户信息。用户的评价和偏好明确地表示为一个 $m*n$ 的项目评价矩阵 R，这里 m 是用户数，n 是项目数，R=[r_{ij}]，元素 r_{ij} 表示用户 i 对项目 j 的评分。在电子商务推荐系统中，元素 r_{ij} 既可表示用户是否购买商品，也可表示用户对商品的偏好程度。

2) 产生“邻居”。计算系统中目标用户与其他所有用户的相似度，找出 K 个最相似用户集—“最近邻居”。K-“最近邻居”根据相似度大小从大到小排列的“邻居”集合。

计算用户两个用户之间相似性首先要获取这两用户评分过的所有项目，然后利用某种相似性度量方法进行计算。度量用户相似性有多种方法，常见的有余弦相似性、相关相似性和修正余弦相似性[5]。

①余弦相似性（Cosine）

用户评分数据可以看作 n 是维项目空间上的向量，用户之间的相似性通过向量间的余弦夹角度量，若用户对某项没有评分，则将该项评分设为 0。设用户 i 和用户 j 在 n 维项目空间上的评分分别用 $\vec{i}$，$\vec{j}$ 表示，在用户 i 和用户 j 之间的相似性计算公式如式(1)所示：

$$sim(i,j)=\cos(i,j)=\frac{\vec{i}\cdot\vec{j}}{\|\vec{i}\|\times\|\vec{j}\|} \tag{1}$$

②相关相似性（Correlation）

设 I_{ij} 表示被用户 i 和用户 j 共同评分过的项目集，则用户 i 和用户 j 之间的相似性 $sim(i,j)$ 通过 Pearson 相关系数度量：

$$sim(i,j)=\frac{\sum_{c\in I_{ij}}(r_{ic}-\bar{r}_i)(r_{jc}-\bar{r}_j)}{\sqrt{\sum_{c\in I_{ij}}(r_{ic}-\bar{r}_i)^2}\sqrt{\sum_{c\in I_{ij}}(r_{jc}-\bar{r}_j)^2}} \tag{2}$$

式中，r_{ic} 表示用户 i 对项目 c 的评分，$\bar{r}_i$，$\bar{r}_j$ 分别表示用户 i 和用户 j 对项目的平均评分。

③修正余弦相似性（Adjusted Cosine）

余弦相似性的度量方法中并没有考虑不同用户的评分尺度问题，修正的余弦相似性度量方法通过减去用户对项目的平均评分来改善上述缺陷，设 I_{ij} 表示被用户 i 和用户 j 共同评分过的项目集，设 I_i，I_j 分别表示被用户 i 和用户 j 评分过的项目集，则用户 i 和用户 j 的相似度计算公式如式(3)所示：

$$sim(i,j)=\frac{\sum_{c\in I_{ij}}(r_{ic}-\bar{r}_i)(r_{jc}-\bar{r}_j)}{\sqrt{\sum_{c\in I_i}(r_{ic}-\bar{r}_i)^2}\sqrt{\sum_{c\in I_j}(r_{jc}-\bar{r}_j)^2}} \tag{3}$$

式中，r_{ic} 表示用户 i 对项目 c 的评分，$\bar{r}_i$ 和 $\bar{r}_j$ 分别表示用户 i 和用户 j 对项目的平均评分。最近邻居搜寻就是对每个用户 i，在整个的用户空间中查询用户集 $m=\{n_1,n_2,\cdots,n_s\}$，使 $i\notin m$ 且 n_1 与 i 相似度最高，n_2 与 i 相似度次之，依次递减排列。

3) 推荐。“ 最近邻居” 集产生后,可计算目标用户对项目的预测评价值进行 Top - N 推荐。通过预测项目评分值搜索最近邻居而产生推荐，预测评分计算方法如下：

$$p_{i,y}=\bar{R}_i+\frac{\sum_{j\in NN,y\in N} sim(i,j)(R_{j,y}-\bar{R}_j)}{\sum_{j\in NN,y\in N}|sim(i,j)|} \tag{4}$$

其中，Pi,y 表示目标用户 i 对项目 y 的预测评价值；$\overline{R}_i$ 为用户 i 的平均评分值；Pj,y 表示为目标用户 i 的最近邻居集的用户 j 对项目 y 的评价。在此，目标用户 i 的最近邻居集用 NN 表示。按照兴趣度预测值 Pi,y 的高低产生推荐集。

传统基于用户的协同过滤算法要求有较多的用户评分数据且算法效率也较低。

2 Web日志分析处理

协同过滤算法需要整理用户评分数据、计算相似性、寻找最近邻从而完成推荐。而大多的网络用户很少对资源进行显示的评分，这需要 Web 挖掘算法来获取隐性的评价数据。本文以学校网络中心网络日志作为数据源。在此主要分析和利用用户行为记录表，其包括用户名、目标 IP、应用类型、访问时间、URL 及网站等信息。用户行为记录表格式描述如下表 1 所示。

表 1 用户行为记录表简要格式

用户	目标	IP 应用类型	访问时间	URL	网站
u1	IP1	访问网站	2009.11.11 18:12:27	URL1	NET1
u2	IP2	http 下载	2009.11.11 18:41:22	URL2	NET2
u1	IP3	娱乐	2009.11.11 19:35:43	URL3	NET3
u2	IP4	聊天交友	2009.11.11 19:56:08	URL4	NET4
u1	IP5	访问网站	2009.11.11 20:02:14	URL5	NET5

从表 1 中可以看出，用户 u1 在三十分钟内共访问了 NET1、NET2 以及 NET5 的 3 个网站，且这三个站点分别属于两种不用的应用类型，表明了用户 u1 的兴趣不是唯一的。用户的兴趣可以用其选择的项目（访问的网站）来反映且项目的类型不同也体现了用户兴趣的不同。在实际数据中，用户名是以 IP 地址表示的，为了计算方便将用户名解析编号后计算各个用户的度(用户访问的不同站点的数目)。用户编号的 sql 脚本如下：

```
create procedure makeusercode()
    BEGIN
      declare num1 int default 1;
      declare flag int default 0;
      declare userid int(10);
      declare usercode cursor for select distinct host_id from au1
            union
          select distinct host_id from au2
            .....
            union
          select distinct host_id from aun;
          declare continue handler for not found set flag=1;
          open usercode;
          repeat
          fetch usercode into userid;
          inset into code(num,userid) values(num1, userid);
          set num1=num1+1;
          until flag=1 end repeat;
    END
```

其中，au1 至 aun 为用户行为记录表；计算用户度的脚本在此略。

3 算法改进

依据上述分析的基础上，对传统协同过滤算法进行如下改进：首先按照用户访问站点的类型对其兴趣分类；其次对同一个用户预测最近邻时要区分预测项目的类别(页面的应用类型)以保证预测的准确性；再次利用改进的模糊聚类算法对相似用户进行聚类，生成最近邻，以提高算法的精度和效率；最后按照用户兴趣在每类项目中所占的权重分配相应的该类项目的推荐数目[6-8]。

算法的设计步骤如下：

1) 用户兴趣分类 用户的兴趣可以通过其浏览的网站反映，按照网站的应用类型分类，每种应用类型至少包括一个网站（项目）。由于日志数据中已经按照应用类型进行分类了，则对用户兴趣分类实现较为简单。以表 1 为例简要说明：用户 u1 访问的网站有 NET1、NET3 及 NET5，它们的应用类型分别为访问网站和娱乐，则用户 u1 的兴趣分成两类，表示为 CI_1=2。

2) 构建用户兴趣矩阵 根据用户访问网站的记录映射用户兴趣，计算用户兴趣度，建立用户兴趣矩阵。考虑到推荐的时效性，用户最近的浏览记录对推

荐越有利，从 2 个月的日志中间断的截取 4 个时间段，每段的时间周期为 3 天。由于本文挖掘和分析的是用户隐式评分信息，评分信息将由用户浏览页面的行为间接反映。为了直观描述用户对项目的兴趣度，将用户对项目的兴趣值划分成 0 到 5 共 6 个标准，划分方法是为每个标准设置一个阈值，当用户 i 对项目 j 的访问次数超过某个阈值时，评定相应的值。假设提出以 ri,j 表示用户 i 对项目 j 的评价值。在一段时间内，考虑有些用户兴趣较分散，有些较集中，会对推荐造成一定的影响，在量化评分时按照用户兴趣的种类数做适当的修正。用户评价值 $r_{i,j}=\frac{\alpha}{CI_i}\cdot S_{i,j}$，其中 $S_{i,j}$ 为用户 i 对项目 j 的兴趣值；α 为调节参数，可根据需要调整，一般取 1。计算对于用户 i 浏览过类型 $CI_i\ (i=1,2,\cdots k)$ 中的网站（项目）总数，据以上的方法统计用户 i 对每个应用类型的评分值。最后，计算出用户对所有应用类型的兴趣值，形成用户的兴趣矩阵。

3) 寻找最近邻并推荐　对用户兴趣矩阵按不同的兴趣分类分别利用改进的模糊聚类算法进行聚类，从聚类结果中寻找最近邻。该步的关键就是计算用户之间的相似性，对于相似性计算方法有很多，综合考虑采用基于修正余弦相似性的计算方法。在应用类型 $c\ (c=1,2,\cdots,k)$ 中，计算目标用户 i 与类型 c 中用户 j 之间的相似度 $sim(i,j)_c$。

$$sim(i,j)_c=\frac{\sum_{c=1}^{k}(r_{i,c}-\bar{r}_i)(r_{j,c}-\bar{r}_j)}{\sqrt{\sum_{c=1}^{k}(r_{i,c}-\bar{r}_i)^2}\sqrt{\sum_{c=1}^{k}(r_{j,c}-\bar{r}_j)^2}} \quad (5)$$

对于用户 i 而言，把计算出的所有相似值按照从大到小选出若干个作为其最近邻居集。

以下将计算用户目标项目的评价值：设目标用户 i，计算各个应用类型 $c\ (c=1,2,\cdots,k)$ 中用户未进行隐式评分 j 项目的评分预测值 $p(i,j)_c$。

$$p(i,j)_c=\bar{r}_i+\frac{\sum_{j=1}^{n}sim(i,j)_c\cdot(r_{j,c}-\bar{r}_j)}{\sum_{j=1}^{n}sim(i,j)_c} \quad (6)$$

推荐时考虑不同项目所占推荐权重，按照推荐权重分配该项目的推荐数目。可定义推荐权重为 $w_{i,j}$：

$$w_{i,j}=\frac{N_{i,j}}{N_i CI_i} \quad (7)$$

式(7)中，CI_i 为用户 i 访问网站类型的数目；$N_{i,j}$ 表示该用户 i 访问网站 j 的次数；N_i 为用户 i 访问各类项目的总次数。考虑每个应用类型的预测评价值和每个网站（项目）在各自应用类型中的所占的权重加权后，推荐每种应用类型的项目的数目。

改进的算法的复杂度较传统协同过滤算法低很多，实际推荐系统会随着时间变化，数据量急剧增大，效率降低。采用用户聚类，将相似兴趣的用户进行聚类分组，搜索最近邻用户时，只要从相应的用户分组中搜索推荐，并且聚类还可以离线进行，降低了算法的执行时间。该推荐算法较适合网页的推荐，不需要获取较多的用户信息，且对于大量用户参与的情况也能适应，只需分析网络日志获取用户浏览行为数据进行隐式评分即可。

4　算法实验

为了验证算法的有效性，本文采用学校网络中心 Web 日志数据作为实验数据集。为了实验方便，搜集并处理了近两个月的用户行为记录数据，得到用户数目 232 个，网站数目 1241 及应用类型 8 个近 100000 条记录。将数据集的 90%作为训练集构建兴趣矩阵，其余的作为测试集。

本文采用信息检索领域中评估系统效果的准确率(Precision)标准来衡量传统算法和本文算法的精度[9]。

$$\mathrm{Precision}=\frac{Number\_of\_Hits}{N} \quad (8)$$

在上式子中，$Number\_of\_Hits$ 表示推荐准确数，N 表示推荐总数。

在设计的 5 组试验中，项目的最近邻数目 K=20，分别观察推荐总数 N 从 10 到 30 的不同情况下的算法结果比较。图 1 为部分测试数据的模糊聚类结果；图 2 中的推荐结果为 5 组实验的平均结果。

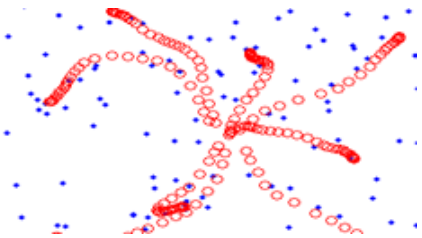

图 1　分类兴趣模糊聚类结果

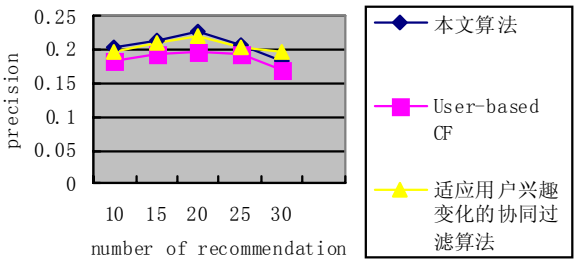

图 2 算法实验结果

从图 2 可以看出，本文改进的协同过滤算法较传统协同过滤和适应用户兴趣的系统过滤算法有明显的推荐精度。实验还发现改进的算法的执行时间有较大的提高。改进的推荐方法不仅考虑了用户最近兴趣对推荐的影响，着重是实际系统中用户兴趣的多样性的特征，从而有力的提高了推荐精度。

5 结语

本文主要分析了传统协同过滤推荐算法的不足和实际用户兴趣的多样性的特点，提出改进传统协同过滤算法的具体措施。文章采用真实日志数据进行仿真实验，实验结果表明改进的算法在推荐效率和推荐精度上都有明显的优势。随着个性化推荐的发展，对推荐算法在实时性和复杂度的要求将是以后推荐算法研究的重点。

参考文献

1 Goldberg D, Nichols D, Oki BM, Terry D. Using collaborative filtering to weave an information tapestry. Communications of ACM, 1992,35(12):61−70.

2 Sang YY, Liu PG, Li Y. A collaborative filtering algorithm fitting user interest evolution. Journal of the China Society for Scientific and Technical Information, 28(1):109−113.

3 Xing CX, Gao FR, Zhan SI, Zhou LZ. A collaborative filtering recommendation algorithm incorporated with user interest change. Journal of Computer Research and Development, 2007,44(2):296−301.

4 王茜,王均波.一种改进的协同过滤推荐算法.计算机科学,2010,37(6):226−227.

5 Sarwar B, Karypis G, Konstan J, Riedl J. Item-Based Collaborative Filtering Recommendation Algorithms. Proc. of the 10th Inernational World Wide Web Conference, 2001: 285−295.

6 Konstan J, Miller B, Maltz D, et al. GroupLens: Applying collaborative filtering to usenet news. Communications of the ACM, 1997,40(30):77−87.

7 Resn Ick P, Var Ian HR. Recommender systems. Communications of ACM, 1997,40(30):5628.

8 Zeng C, Xing CX, Zhou LZ. Similarity meansure and instance selection for collaborative filtering international. Journal of Electronic Commerce, 2004,4(8):115−129.

9 杨芳,潘一飞,李杰,等.一种改进的协同过滤推荐算法.河北工业大学学报,2010,39(3):82−87.

（上接第 29 页）

参考文献

1 田日才.扩频通信.北京:清华大学出版社,2007.

2 Roger L, Peterson RE, Ziemer DE, et al. Introduction to Spread Spectrum Communications. 北京:电子工业出版社,2006.2−28.

3 曹志刚,钱亚生.现代通信原理.北京:清华大学出版社,1992.

4 朱近康.CDMA 通信技术.北京:人民邮电出版社,2001.

5 Ziemer RE, Peterson RL. Introduction to Digital Communication. Prentice Hall, Inc. 2001.

6 Shannon CE. A mathematical theory of communication. Bell System Technical Journal, 1948,(27):379−423,623−656.

7 Seay TS. Hopping Patterns for Bounded Mutual Interference Infrequency Hopping Multiple Access. Proc. of hte 1982 IEEE MILCOM Conference, Boston, Massachusetts.

8 扬小牛,楼才义.软件无线电原理与应用.北京:电子工业出版社,2001.

9 周润景,图亚,张丽敏.基于 Quartus 的 FPGA/CPLD 数字系统设计实例.北京:电子工业出版社,2007.40−96.

10 徐光辉,程东旭,黄如.基于 FPGA 的嵌入式开发与应用.北京:电子工业出版社,2006.916−84.

11 潘松,黄继业.EDA 技术与 VHDL.北京:清华大学出版社,2005.

12 李光军,孟宪元.可编程 ASIC 设计及应用.北京:电子科技大学出版社,2000.

13 Ashenden PJ.VHDL 设计指南.北京:机械工业出版社,2005.

14 黄智伟,陈琼.FPGA 系统设计与实践.北京:电子工业出版社,2005.294−329,85−122.