文章编号：1001-2486（2002）03-0067-05

一种面向个性化服务的无需反例集的用户建模方法*

应晓敏，刘 明，窦文华
（国防科技大学计算机学院，湖南 长沙 410073）

摘 要：随着 WWW 信息的快速增长，查找用户感兴趣的信息变得越来越耗时耗力。个性化服务能为不同的用户提供有针对性的服务，日益受到研究者的重视。用户建模是实现个性化服务的关键技术。传统的需要正、反例集作为训练例集的用户建模方法容易干扰用户的正常浏览，或者由于推断失误而引入噪声。基于遗传算法和 k 近邻方法提出了一种无需反例集的用户建模方法，该方法被应用于个性化信息过滤中。实验结果表明，基于无需反例集的用户建模方法的信息过滤算法可以达到 73.91% 的过滤率和 94.44% 的过滤精度。无需反例集的用户建模方法是一种可行、高效的用户建模方法。

关键词：个性化服务；用户建模；遗传算法
中图分类号：TP393 文献标识码：A

A User Modeling Method without Negative Examples for Personalized Services

YING Xiao-min，LIU Ming，DOU Wen-hua
（College of Computer，National Univ. of Defense Technology，Changsha 410073，China）

Abstract：With the exponential growth of World Wide Web，it becomes more and more time and energy consuming for users to find what they're interested in. It leads to a clear demand for personalized services，which can provide different users with different services. User modeling is the key technology in implementing personalized services. Conventional user modeling methods with both positive and negative examples will either interfere users' normal browsing or bring in noises. A user modeling method without negative examples is presented. A hybrid of genetic algorithms and kNN classifier are utilized to search the words describing users' interests. The method is applied in personalized information filtering. The experiments show that the filtering ratio and precision can be 73.91% and 94.44% respectively，which demonstrates that our user modeling method is feasible and efficient.

Key words：personalized services；user modeling；genetic algorithms

互联网的飞速发展正在改变着人们的生活，WWW 已经成为人们交流和获取信息的重要媒介。据统计，到 2000 年 7 月，WWW 已经发展成为一个拥有 20 亿页面的分布式信息空间，而且这个数字仍在以每天七百万的速度增加[1]。然而资源的极大丰富也带来了资源使用的困难，人们发现在浩瀚的 Web 信息资源中查找和发现用户感兴趣的信息成为一件非常耗时耗力的事情。Web 信息的迅速增长使得原有的不区分用户的服务越来越难以满足人们的需求。为了将用户从信息的海洋中解脱出来，直接、快速地浏览感兴趣的内容，研究者提出了个性化服务。所谓个性化服务就是指对不同的用户采取不同的服务策略，提供不同的服务内容[2]。很显然，在提供个性化服务时，系统必须知道用户的兴趣、偏好和访问模式等用户信息，才可能“投其所好”，实现个性化服务。用户模型是关于用户兴趣、偏好、模式的可计算描述，它是个性化服务的基础。用户建模是指根据用户信息（如浏览内容、浏览行为等）归纳出用户模型。用户建模是实现个性化服务的关键技术。

1 用户建模

用户兴趣类可粗粒度地划分为用户感兴趣类（IC，Interesting Category）和用户不感兴趣类（NIC，

* 收稿日期：2001-12-27
作者简介：应晓敏（1975—），女，博士生。

Not Interesting Category)。这种粗粒度划分的建模一般都被视为一个二类归纳学习问题[3]，其算法大多要求用户同时提供正、反例集。

California 大学 Irvine 分校开发的 Syskill&Webwert[4] 系统要求用户对浏览过的页面标注 hot 或 cold，由此获得训练用户模型的正、反例集。然后根据这些训练样本计算出每个词出现在正、反例集中的概率。关键词及其出现在正、反例集中的概率构成了用户模型。

CMU 大学开发的 PWW[5] 系统采用的用户建模方法与 Syskill&Webert 类似，所不同的是 PWW 不要求用户对页面进行标注。它观察用户对页面中超链接的选择，推断用户选中的超链接所指向的链宿页面为用户感兴趣类，反之则为用户不感兴趣类。通过这种推断获取训练用户模型的正、反例集，而后计算每个词出现在正、反例集中的概率，以此构成用户模型。

Stanford 大学开发的 LIRA[6] 系统与 Syskill&Webert 类似，要求用户在浏览的过程中对浏览的页面进行标注，标注值在 [- 5， + 5] 内，反映页面的兴趣度，而后采用相关性反馈学习用户模型。

为了得到训练例集，上述系统或者要求用户对浏览过的页面进行标注，如 Syskill&Webert 和 LIRA，或者根据用户对页面的访问情况进行推断（用户访问过的页面为正例，否则为反例），如 PWW。

然而，Carroll 和 Rosson 的心理学研究表明：即使用户知道他的工作在不久后会给他带来好处，用户仍然不愿意参与这个训练过程[7]。事实上，要求用户在浏览过程中对页面进行标注会干扰用户的正常浏览，降低系统的可用性。

同样，推断用户访问过的页面为用户 IC 页面、反之为用户 NIC 页面也可能给系统带来严重的后果，典型的有：

(1) 用户可能点击某个超链接，浏览链宿页面，结果发现该页面其实并非自己感兴趣的页面；

(2) 用户可能忽视某些自己感兴趣的页面，从而导致有些应该浏览的页面未被浏览过；

(3) 由于传统 Web 浏览器的界面结构易于导致用户进行深度优先搜索，用户容易忘记回退到过去的页面以浏览某些自己感兴趣的超链接。

为了减少对用户的干扰，降低因推测而引入的系统噪声，我们提出无需反例集的用户建模方法。

2 无需反例集的用户建模方法

2.1 算法描述

我们采用示例学习建立用户模型，用户模型由两部分组成：(1) 通过 FSS 算法选择出来的特征子集；(2) 代表用户各个子兴趣类的代表点。

设用户感兴趣类包括 l 个子兴趣类 $\boldsymbol{C} = \{C_1, C_2, \cdots, C_l\}$。所筛选出的特征子集为 $\boldsymbol{F} = \{f_1, f_2, \cdots, f_n\}$，其张成的特征子空间为 $\boldsymbol{\Phi} = f_1 \times f_4 \times \cdots \times f_n$。代表点集合为 $\boldsymbol{R} = \{r_1, r_2, \cdots, r_m\}$，其中 $r_i (i = 1, 2, \cdots, m)$ 为特征子空间 $\boldsymbol{\Phi}$ 中的特征向量，其所属的子兴趣类为 c_i，$c_i \in \boldsymbol{C}$。则用户模型 $\boldsymbol{M}$ 可以表示为 $\boldsymbol{M} = \{\boldsymbol{F}, \boldsymbol{\Phi}, \boldsymbol{R}, \boldsymbol{C}\}$。

无需反例集的用户建模方法是指建模过程不需要用户提供 NIC 页面，也无需对用户的兴趣进行推测的用户建模方法。建模过程只需将用户感兴趣的页面（或文件）作为训练例集，根据用户各个兴趣类的精细分类结构进行特征子集的选择，从而构成用户模型。

2.2 训练例集的获取

用户在浏览 Web 的过程中会保存一些自己感兴趣的页面和文件，而且为了便于将来浏览，往往还会对保存的页面和文件加以整理和分类。这些分好类的页面和文件可以作为训练用户模型的样本。这样获取训练例集的优点主要有：

(1) 不会中断用户的正常浏览，用户只需要提供保存这些分类页面的目录；

(2) 所用的页面均为用户 IC 页面，不会引入系统噪声；

(3) 由于保存了 IC 中的精细分类，从而可以充分利用正例集的结构信息。

2.3 特征子集选择

Mladenic[8] 对 Yahoo！的 49600 个页面进行了统计，结果表明：删除停止字（指无实际意义的词，

如中文的“的、地、得”，英文的“a、the”等）后有 32 万个词；再减去生僻词后还有 7 万个词。如此数量庞大的特征不仅会大大增加学习算法的计算代价，而且还会严重地影响归纳学习的质量[9]。

特征子集选择（FSS，Feature Subset Selection）是指从一个大的候选特征集合中选择一个“好”的子集来一致地描述已知例集，“好”的评价标准是该子集中的特征最具有代表性[10]。特征子集选择实质上是一个组合优化问题。假设从 N 个原始特征中挑选 n 个特征，则所有可能的组合数为：

$$C_N^n = \frac{N!}{(N-n)!\,n!} \tag{1}$$

在我们的应用中，$N = 10,125$，即使删去出现频率小于 3 的特征，依然有 3384 个特征；而且文档中真正起描述作用的特征数目相对较少，即 $n \ll N$。根据式（1）我们不难看出，FSS 的组合数是一个惊人的数字。最优特征子集选择因为搜索空间太大而难以实现。文献［10］证明了最优特征子集选择是一个 NP 难题。

GA 是一种模拟生物界自然选择和遗传进化的高度并行、随机、自适应的搜索算法。它具有全局搜索能力，全空间并行搜索，并将搜索重点集中于性能高的部分，效率高而且不易陷入局部最优，善于搜索复杂问题和非线性问题[11]。在本文中，我们基于 GA 和 kNN 分类器在特征空间中搜索近优特征子集。

将例集分为代表点集合与训练例集两部分。代表点集合构成 kNN 分类器，训练例集在候选特征子集张成的空间中通过 kNN 分类器的分类精度作为该候选特征子集的适应值。为了在分类精度和特征子集包含的特征数目之间取得平衡，可以定义适应度函数为

$$fitness(s) = precision(s) / percentage(s) \tag{2}$$

或者

$$fitness(s) = (1-\omega)\,precision(s) + \omega(1 - percentage(s)) \tag{3}$$

其中，s 为候选特征子集，$fitness(s)$ 为 s 的适应度函数，$precision(s)$ 为用 s 表示的训练例集通过 kNN 分类器的分类精度，$percentage(s)$ 是指 s 包含特征的个数占特征总数的比例，ω 为 $precision(s)$ 和 $percentage(s)$ 的协调因子，$\omega \in [0,+1]$，反映在期望分类精度与期望特征数目之间的一种权衡。

实验中我们发现，式（2）定义的适应度函数在特征维数较高时对包含特征个数多的子集惩罚太重，导致 GA 算法过早收敛。相对而言，式（3）定义的适应度函数不仅更利于搜索最优的特征子集，而且通过对协调因子 ω 的定制，使得算法更加灵活。

3 实验

3.1 实验数据及其预处理

在实验中，我们搜集了代表用户三个子兴趣类（猫科动物、农业和信息过滤）的 150 个网页。其中 30 个页面作为代表点，其余页面作为训练例集。在用 GA 搜索特征子集之前，所有页面经过预处理，表示成二进制编码的形式。首先将 HTML 页面转换成 TXT 文档；然后建立停止字表，删去所有 TXT 文档中的停止字；对所有的词取词根，使具有相同含义但语法形式不同的词归一成同一个词根，如 computing 和 computed 归一成 comput；统计所有文档的词根总数，形成长度为 N 的词根列表，N 就是原始特征总数；对原始特征进行压缩，删去出现频率小于 3 的原始特征，形成长度为 N' 的词根列表 *StemList*；最后每个文档与 *StemList* 比较，若有该词根，则该位为 1，否则为 0，形成长度为 N' 的基因串。

3.2 FSS 实验结果

在实验中，原始特征总数为 $N = 10,125$，词根列表长度 $N' = 3384$。GA 的初始种群个体数为 50，交叉概率为 0.7，变异概率为 0.001，采用随机全局选择算子[12]（SUS，Stochastic Universal Sampling），单点交叉算子和基于适应值的最优保持策略。适应度函数取式（3），协调因子 $\omega = 1/2, 2/3$ 和 $3/4$，前 20 代种群的平均适应值、平均分类精度和平均特征数分别见图 1、图 2 和图 3。

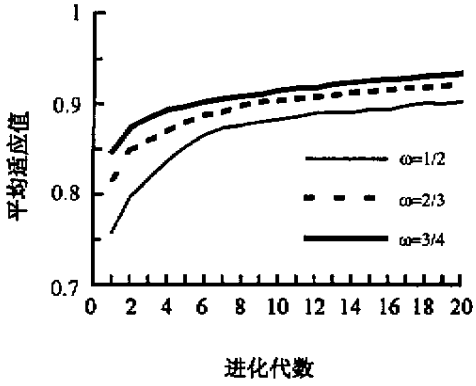

图 1 协调因子 ω 取不同值时
前 20 代种群的平均适应值

Fig.1 Mean fitness of the first 20 generations while ω equals different numbers

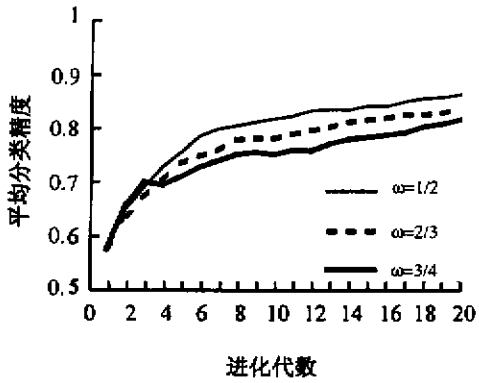

图 2 协调因子 ω 取不同值时
前 20 代种群的平均分类精度

Fig.2 Mean classification precision of the first 20 generations while ω equals different numbers

从图 1 可以看出，GA 从第 18 代开始逐渐收敛，平均适应值趋于平稳。ω 的取值对算法的收敛速度没有明显的影响。图 2 和图 3 分别是每代平均分类精度和平均特征数目的进化曲线。从图中可以看出，当协调因子 $\omega \to 1$ 时，平均分类精度减小，而平均特征数增大。这是因为根据式（3），随着 ω 的增大，分类精度在适应度函数中的比重减小，特征数目在适应度函数中的比重增大。

3.3 信息过滤实验及结果

3.3.1 信息过滤算法描述

对于 $\forall w$（w 为网页），w 可以表示成特征子空间 Φ 中的一个特征向量 w。下面给出信息过滤算法，T 为信息过滤阈值。

FOR i = 1 TO m DO

similarity$(\mathbf{w},\mathbf{r}_i) = \mathbf{w}\cdot\mathbf{r}_i$

ENDFOR/ * i * / / * 计算 w 与代表点 $\mathbf{r}_i$ 的相似性 * /

FOR i = 1 TO l DO

FOR j = 1 TO m DO

IF $c_j = C_i$

THEN

$p_j = 1$

ELSE

$p_j = 0$

ENDIF

ENDFOR/ * j * /

$$\text{similarity}(\mathbf{w},C_i) = \sum_{k=1}^{m} \text{similarity}(\mathbf{w},\mathbf{r}_k)\cdot p_k$$

ENDFOR/ * i * / / * 计算 w 与 C_i 的相似性 * /

IF max (similarity$(\mathbf{w},\mathbf{C})$) > T

THEN

w 属于 max(similarity$(\mathbf{w},\mathbf{C})$)所在的类别

ELSE

$w \notin \mathbf{C}$；

执行过滤功能；

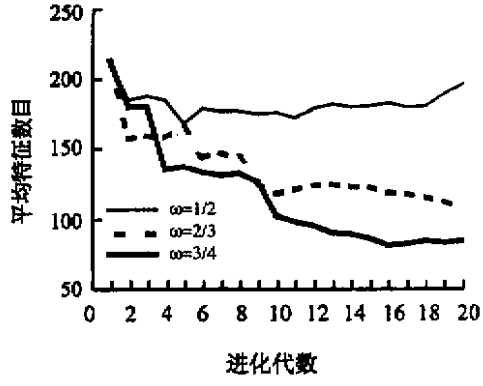

图 3 协调因子 ω 取不同值时
前 20 代种群的平均特征数目

Fig.3 Mean feature numbers of the first 20 generations while ω equals different numbers

```
ENDIF        /* 判断 w 所属的类别 */
```

3.3.2 实验及结果

基于无需反例集的用户建模方法生成的用户模型，我们对搜索引擎 Yahoo！返回的结果进行了过滤实验。我们在 Yahoo！上输入查询“lion”（lion 为用户感兴趣的猫科动物之一），搜索引擎返回 748 个匹配网站。其中绝大多数网站虽然包含“lion”这个关键词，但实质上与动物“lion”毫无关系。例如在前 80 个网站中，只有 11 个网站与动物“lion”相关，还有 12 个网站因为关闭或过期而无法访问，其余网站均为以“lion”命名的公司或组织的网站，如 The Lion King WWW Archive，Red Lion Hotels & Inns，Lion City Hotel 等。基于无需反例集的用户建模方法生成的用户模型，我们对前 80 个匹配网站进行了过滤实验。共有 54 个页面被过滤，其中包含 3 个与动物“lion”相关的页面。过滤率为 73.91%，过滤精度为 94.44%。

4 结论和讨论

传统的用户建模方法需要正、反例集，容易对用户的正常浏览造成干扰，或者由于推测样本所属的类别而引入噪声。本文提出了一种无需反例集的用户建模方法，该方法无需用户 NIC 页面，既不会造成对用户的干扰，也不会引入噪声样本。相对于传统用户建模方法，该方法具有更好的系统可用性。

采用 GA 和 KNN 分类器进行特征子集的选取效率较高，不易陷入局部最优，而且能够搜索到描述用户兴趣的特征组合。但在实验中我们发现，采用 GA 搜索出的特征子集不仅包含了描述用户兴趣的特征，而且也包含了一些不能很好地描述用户兴趣的特征。这些特征的存在降低了过滤率和过滤精度。在进一步的工作中，我们考虑从以下几个方面改进以提高用户模型的准确度：

(1) 不再随机抽取样本作为 kNN 分类器的代表点，而是选择最能代表子兴趣类的样本作代表点；

(2) 在采用 GA 搜索较优特征子集之前，先对特征空间进行降维，缩小 GA 的搜索空间，提高搜索的效率。

参考文献：

[1] Sizing the Internet. July 10，2000 [Z] . http：//cyveillance.com/newsroom/press/ 000710.asp.

[2] 高文，刘峰，黄铁军等 . 数字图书馆——原理与技术实现 [M] . 北京：清华大学出版社，2000.

[3] Webb G，Pazzani M，Billsus D. Machine Learning for User Modeling [J] . User Modeling and User-Adapted Interaction，2001，11：19-29.

[4] Pazzani M，Billsus D. Learning and Revising User Profiles：The Identification of Interesting Web Sites [J] . Machine Learning，1997，27：313-331.

[5] Mladenic D. Text-Learning and Related Intelligent Agents：a Survey [J] . IEEE Intelligent Systems，July/August 1999，14 (4)：44-54.

[6] Balabanovic M，Shoham Y，Yun Y. An Adaptive Agent for Automated Web Browsing [J] . Journal of Visual Communication and Image Representation，1995，6 (4) .

[7] Carroll J，Rosson M. Paradox of the Active User [A] .Interfacing Thought：Cognitive Aspects of Human-Computer Interaction，MIT Press，1987：80-111.

[8] Mladenic D，Grobelnik M. Word Sequences as Features in Text-learning [A] . In Proceedings of the 7th Electrotechnical and Computer Science Conference (ERK ’98) . Ljubljana，Slovenia：IEEE Section.

[9] Caruana R，Freitag D. Greedy Attribute Selection [A] . In Proceedings of Machine Learning ’94，New Brunswick，NJ，1994，28-36.

[10] 陈彬，洪家容，王亚东 . 最优特征子集选择问题 [J] . 计算机学报，1997，20 (2)：133-138.

[11] 王凌 . 智能优化算法及其应用 [M] . 北京：清华大学出版社，2001.

[12] Baker J E. Reducing Bias and Inefficiency in the Selection Algorithm [A] . In Proceedings of the Second International Conference on Genetic Algorithms，1987：14-21.