维普资讯 http://www.cqvip.com

基于不确定性推理的个性化算法研究*)

白 勇[1] 蓝章礼[2]

(重庆电力高等专科学校 重庆 400053)[1] (重庆交通学院计算机学院 重庆 400074)[2]

摘 要 远程教育系统的个性化研究是目前网络教育的重要课题,而学生模型的建立是提供个性化服务的关键。为此,本文通过采用人工智能技术,以自然频率法为基础推算后验概率,并通过与可信度方法相结合进行运算,对学生用户模型的建立提出了一种算法。

关键词 远程教育,个性化,算法,可信度

Research of Individual Algorithm Based on Uncertainty Inference

BAI Yong[1] LAN Zhang-Li[2]

(Chongqing Electric Power College, Chongqing 400053)[1]

(Computer & Information College of Chongqing Jiaotong University, Chongqing 400074)[2]

Abstract The study of individuation is an important issue of the distance education. And how to build a student's model is the key of individual service. Based on the principle theory of AI, using natural frequency method to calculate the posterior probability, and uniting with the reliability, an algorithm of performing a student's individual user model is given.

Keywords Distance education, Individuation, Algorithm, Confidence

1 引言

现代远程教育的个性化是目前研究现代远程教育的重要课题,而要在现代远程教育系统中提供个性化的教学服务,其关键是要能够建立一个合理的、能体现不同学生特点的学生用户模型,并在该模型的基础上进行个性化的教学推荐。而学生的个性化用户建模和个性化推荐都需要一定的算法来进行实现,为此,本文对学生个性化用户建模提出作者的观点,将不确定推理中可信度的方法运用于学生用户建模中,在学生用户建模上找到一种新的算法。

2 采用的技术与方法

一个个性化的远程教育系统成功与否的关键在于能否根据学生的行为收集必要的学生信息,并根据这些行为信息对学生的用户模型进行建立和改进,这个建模的过程是实现个性化的关键,最后能够按照学生模型中的相关数据对呈现给学生的页面进行个性化的推荐。为此,建立如图 1 所示的学生个性化服务体系结构。

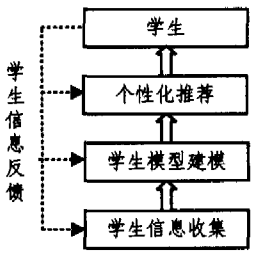

图 1 个性化教学体系结构

2.1 个性化用户建模方式的选择及理由

在学生模型的建立中,每个学生在学习中对各类资源的喜好程度、对每门课程的学习要求等都是不相同的,而且每个学生的兴趣和习惯也不是一成不变,因此每个学生的模型是变化的。在学生登录系统进行学习的过程中,他登录的时间可能并不长,内容也可能并不多,学习每一门课程的时间也不会太长,而且学习下一门课程时,对系统的要求可能又发生了改变,因此如果采取数据挖掘技术对每一个学生的特点进行挖掘将存在着数据量少、待挖掘出结果时其结果可能已经不再适合学生需要等问题。为此,本文认为学生模型的建立应满足以下几点要求:

1)应该采用自动用户建模。因学生对每门课程的要求不一样,再加之有的课程课时本来就少,不可能每门课都要求学生进行手工定制建模,而且学生的兴趣习惯有可能随着不同的课程而有所不同,所以用户建模必须采用自动用户建模,一是避免学生对系统的厌烦情绪,提高服务质量,二是提高用户建模的准确性;

2)用户模型必须具有可描述性。模型应该是一种面向算法的、具有特定数据结构的、形式化的用户描述;

3)用户模型的改变应方便易行;

4)用户模型应具备渐进性;

5)用户模型应具备时效性。用户通过教学系统学习时,对系统的要求是随时间不断变化的,系统应该能够及时察觉用户的这种变化,并能够及时地在页面上反映出来。

本文将在满足上述条件的情况下进行用户建模,其建模的过程是在系统运行的过程中根据用户行为不断进行的。但是,由于自动用户建模容易造成误判,因此,本文在自动获取用户信息的基础上,加入了对用户行为进行可信度判断的方法,以降低误判对系统的影响。

2.2 自然频率法

自然频率法[1]是利用采集到的简单自然数进行运算,得

*)基金项目:网络教育关键技术示范工程(2001BA101)。蓝章礼 讲师。

维普资讯 http://www.cqvip.com

到后验概率。自然频率方法只使用自然数而不使用小数进行运算。其公式为：

$$P(H/E)=\frac{e\&h}{e\&h+e\&-h}$$

在自然频率法中只用击中率和误报率，不用考虑基础率，其中击中率在本文中是指学生既喜欢又点击的次数，用 $e\&h$ 表示，误报率是指学生虽然进行了点击却不喜欢的次数，用 $e\&-h$ 表示，基础率是指点击的总数。该方法虽然形式上与标准概率方法不同，但在数学上是等价的。

在实际应用中，判断击中还是误报的方法可根据学生在某一页停留的时间进行，如果学生点击了某个页面并在该页面停留的时间大于 T（可由该课程的教师确定 T 的大小），则认为击中，统计入击中率。若点击了某个页面，但该页面停留的时间小于 T，即可认为学生虽然点击了该页，但并不喜欢它，没有认真进行阅读和学习，可看成是误报，统计入误报率。

2.3 可信度方法

可信度[2]是指根据经验对一个事物或现象为真的相信程度，用 $CF(H,E)$ 表示，其取值范围为 $[-1,1]$，若 $CF(H,E)>0$，则表示前提条件 E 的出现增加了 H 为真的概率，若 $CF(H,E)<0$，则表示前提条件 E 的出现减少了 H 为真的概率，若 $CF(H,E)=0$，则表示 E 对应的证据出现对 H 没有影响。其计算公式为：

$$CF(H,E)=\begin{cases}\dfrac{P(H/E)-P(H)}{1-P(H)}, & \text{若 } P(H/E)>P(H)\\ 0, & \text{若 } P(H/E)=P(H)\\ \dfrac{P(H)-P(H/E)}{-P(H)}, & \text{若 } P(H/E)<P(H)\end{cases}$$

由上式可知，计算可信度只需知道先验概率 $P(H)$ 和后验概率 $P(H/E)$，先验概率可将数据库中现有的概率作为本次的先验概率进行使用，后验概率则可直接采用上节所阐述的由自然频率法进行计算，利用这两个量计算可信度。

在系统长期的运行中，我们对一个不确定的事实的相信程度不可能只考虑某一次利用先验概率和后验概率运算得到的可信度，而是要综合考虑各次的可信度的结果进行运算，因此需要对各次的可信度值进行合成，其公式为：

$$CF_{1,2}(H)=\begin{cases}CF_1(H)+CF_2(H)-CF_1(H)\times CF_2(H) & CF_1(H)\text{和}CF_2(H)\text{均}\geqslant 0\\ CF_1(H)+CF_2(H)+CF_1(H)\times CF_2(H) & CF_1(H)\text{和}CF_2(H)\text{均}<0\\ \dfrac{CF_1(H)+CF_2(H)}{1-\min\{|CF_1(H)|,|CF_2(H)|\}} & CF_1(H)\text{和}CF_2(H)\text{异号}\end{cases}$$

在对学生行为的统计过程中，学生的每一次点击都是肯定的，因此就点击本身而言都是可信的，关键是点击过后是否真的喜欢才是不确定的。因此，可直接将 $CF(H,E)$ 赋值给 $CF(H)$。系统在第一次使用时，直接将 $CF(H)$ 赋给 $CF_{1,2}(H)$，即第一次运算时为 $CF_1(H)$ 赋 0，实际上第一次没有合成的对象，也没有必要进行合成运算，因此直接赋值。以后各次使用时，将上一次的结果 $CF_{1,2}(H)$ 赋给 $CF_1(H)$，本次运算的 $CF(H)$ 赋给 $CF_2(H)$ 进行运算。

3 系统模型的建立及分析

设某门课程的教学资源按表现形式分为 A、B、C、D 等几类提供给学生浏览学习，这种表现形式可以由系统开发者确定，如按文本、图片、交互动画、音频、视频等进行划分，或按难度等级进行划分，或按授课教师不同划分等。每个学生对每一类资源的喜好程度可能不同，学生这种喜好程度的不同反映在他对各类资源的点击和浏览的多少上，若某类资源学生很喜欢、觉得比较适合自己学习，那么他点击该类资源的次数就必然较多，系统在安排面向该学生的页面时，就应该更多地呈现他所需要的这类资源。那么，系统怎样从学生的学习活动中获取信息，推算学生的喜好并进行针对学生个性的资源安排呢？目前对人员行为的分析常采取数据库和数据挖掘技术进行，如 Web usage mining 技术，它通过对 Server Logs、Error Logs、Cookie Logs 等日志信息，以及用户的注册数据等进行挖掘，获得关于学生学习的信息，作为对学生提供教学服务的依据。这些方法需要在大量数据的基础上才能进行分析和挖掘，不能及时地分析学生的行为特点以便对教学资源进行调整，而且运算量较大，实际运用比较困难。为此，本文利用人工智能中不确定性推理的方法进行研究，采用自然频率法推算后验概率并和可信度方法相结合进行推理，提出一种算法，找到一种根据学生行为对学生模型进行建立和改进的方法，以期能在满足时效性和实用性上有效地解决远程教育系统中教学资源的个性化安排问题。

3.1 基本思想及相关模型

如果一个学生喜欢某一类的资源，那么，他对该类资源点击数必然多，而对他不喜欢的资源类型点击就必然少。如果他的行为百分之百可信，即越喜欢就点击得越多，点击数与喜欢的程度完全成正比，那么，我们可以设计一个简单的模型，对呈现给该学生的页面进行调整。以安排 A 类资源为例建立理想的教学资源安排模型，如图 2。

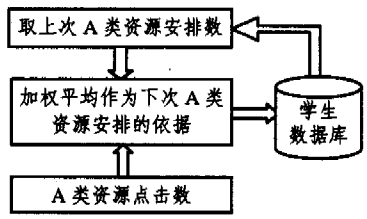

图 2 理想的教学资源安排模型

而实现上，学生的点击行为并不是百分之百可信的，它存在着随意性，有时候可能只是随意点击，是否真正喜欢连学生自己也不一定能确定，可信度不可能百分之百，也就是不能够按照图 2 的方法短简单地对教学资源进行安排。如果把学生的点击行为都按照完全可信的方法进行统计并将之运用到学生模型的建立中，就必然存在大量的误判，如果不排除或降低误判对系统的影响，这样建立起来的学生模型也是不可信的，有多大的实际价值也很难确定，所以需要加入可信度的方法，对学生的点击行为进行判断，建立一个合理的系统资源安排模型。为此本文建立如下算法。

算法的基本思想是：系统在确认学生身份并让其登录后，根据设定的时限 T，统计某类教学资源的击中率和误报率（击中和误报的判断可以利用登录到该页面的时间 T 进行判断，T 值可由教师确定），并利用自然频率法计算后验概率 $P(H/E)$；根据概率计算可信度 $CF(H,E)$，然后进行可信度的合成；用点击数和可信度求积获得本次应该安排的该类资源的安排数；最后根据计算的安排数和上次该类资源的实际安排数进行加权平均，输出下次拟安排的数目到数据库，学生在下一次登录教学系统时系统便根据数据库的记录自动进行教学资源的安排。

下面以安排 A 类资源为例建立系统资源安排模型，如图 3。

维普资讯 http://www.cqvip.com

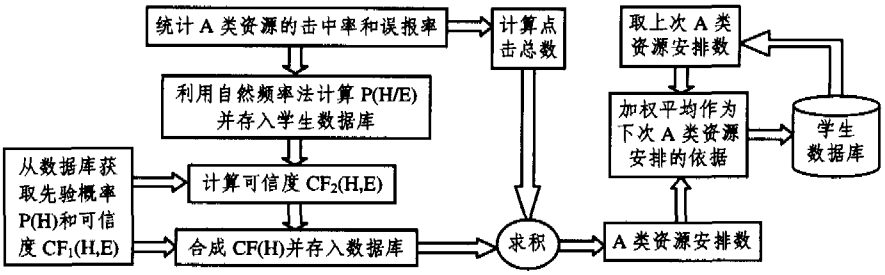

图3 基于可信度的资源安排模型

上述模型是系统对教学资源进行个性化安排的一种方法，其关键在于计算学生点击行为的可信度有多大。因为对于自动用户建模而言，统计学生的行为存在误判，要降低误判对系统自动用户建模的影响，就必须对其行为的可信度进行判断。其中最重要的数据为先验概率 $P(H)$、后验概率 $P(H/E)$、可信度 $CF(H)$和该类资源的安排数。这些数据都是在统计的基础上，利用人工智能中不确定性推理的方法进行运算而得到。

每一个学生要浏览的资源类型有多种多样，每类资源的安排数量、每个学生对每类资源的点击率、先验概率、后验概率及可信度加上学生的一些其它特点，就构成了每个学生的个性特点，就可以根据这些特点建立可以表达学生个性化的用户模型。针对先验概率和后验概率是在不同的时间进行不确定性推理得到的，每次的后验概率就是下次的先验概率的特点，建立如图4所示的学生模型。

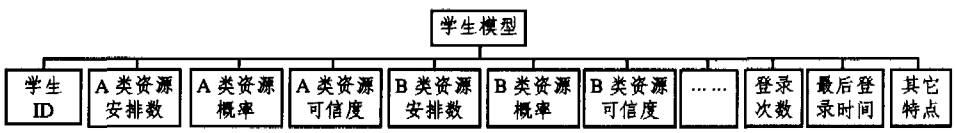

图4 学生模型

每个学生模型实际上就是他对每种资源的喜好程度(体现在安排数上)、击中概率和可信度的集合，这些都是在学生登录网页进行浏览的过程中根据教学资源安排模型的算法逐渐建立起来的，其实质是对每类资源的安排数量进行调整，即利用图3的模型进行改进。在学生学习的过程中利用学生的点击信息进行反馈，对每类资源的安排数量进行调整，也即是对学生模型进行改进。

根据图3资源安排模型及图1个性化教学体系结构，我们可以对个性化教学体系结构进行具体化，设计出具体的教学安排系统模型，如图5。

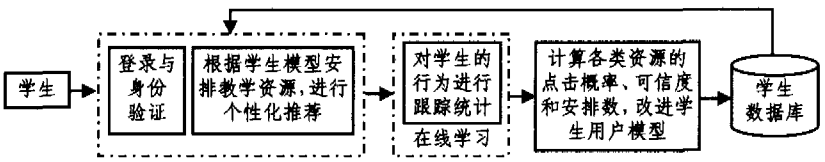

图5 个性化教学安排系统模型

由图5可以看出，学生在登录系统后，系统进行了个性化教学推荐，而学生在线学习时，系统则进行学生信息统计，学生离开系统时，系统按照图3所示的资源安排模型的方法进行学生用户模型的建立或改进，对每类资源的概率、可信度和安排数量进行计算，并记录入学生数据库，以此来自动地改进学生的用户模型，下次学生登录时系统直接读取数据库内有关该生的信息，根据数据库内的数据推荐学习的内容，安排个性化的页面。以此下去，反复对学生的信息进行统计、运算和改进，学生在线学习的点击实际上是对系统的一个反馈，系统正是利用这些反馈信息对学生的用户模型进行改进的，使之不断适应变化中的用户个性化需要。

3.2 算法模型的优点

根据以上对算法所阐述的基本思想和采取的模型及相应的计算公式进行分析，可得出该算法具有以下三个优点：

首先具有及时性，系统不需要采集大量数据进行挖掘和分析后再对教学资源进行调整，而是系统第一次使用后即可对资源的安排进行调整，以后在每次使用后逐步进行改进和调整；

其次是采集的数据量少，只对点击率 $e\&h$ 和误报率 $e\&h$ 进行统计。先验概率 $P(H)$ 和可信度 $CF_1(H)$ 第一次设为0，待第一次使用后进行计算确定，以后可直接调用；

第三是计算量小，后验概率 $P(H/E)$ 和可信度 $CF(H)$ 的计算都只采用简单的加减乘除便可得到，计算十分简单，运算量很小。

由于该算法具有以上优点，因此在实际应用中编程简单，构造的数据库结构简单，比较容易实现，笔者根据以上算法模型建立了一个实验系统，进行了实验，结果与预期的结果一致，具有较强的时效性和实用性。

参 考 文 献

1 赵晓东，傅小兰. 贝叶斯推理的改进方法[J]. 心理科学，2002，25(1)：96～97
2 王永庆. 人工智能原理与方法第一版[M]. 西安：西安交通大学出版社，1998
3 Reviere R. Rethinking open and distance education practices: unearthing subjectivities and barriers to learning. in Commonwealth of Learning (2002) Pan-Commonwealth forum on open learning, 29 July～2 August 2002
4 李华，何茜，吴中福. 基于Web的个性化学习系统研究[J]. 计算机工程与应用，2002，13：239～242
5 Passerini K, Granger M J. A developmental model for distance learning using the internet. Journal article in Computers and Education, 2000, 34(1): 1～15
6 Roblyer M D, Wiencke W R. Design and use of a rubric to assess and encourage interactive qualities in distance courses. The American Journal of Distance education, 2003, 17(2): 77～98