第 25 卷第 2 期　　四川理工学院学报(自然科学版)　　Vol. 25　No. 2
2012 年 4 月　　Journal of Sichuan University of Science & Engineering(Natural Science Edition)　　Apr. 2012
文章编号:1673-1549(2012)02-0055-05

基于数据挖掘的自动化推荐系统算法

朱文忠

(四川理工学院计算机学院, 四川　自贡　643000)

摘　要:结合人工神经网络中的自适应共振理论(ART)及数据挖掘(Data Mining)技术来建构一个可自动聚类族群特征且能挖掘出关联特质的自动化在线推荐系统。探讨如何有效地运用数据挖掘技术从大量的数据库中挖掘出完整知识,以推荐适当的信息给使用者,帮助他们在浩大的信息流中找到真正需要、有用的文件或信息。整合 ART 及数据挖掘技术,并针对推荐系统的特性提出一种改进的 ART 算法(MART 算法)。实例验证了算法的有效性。

关键词:自动化推荐系统;人工神经网络;自适应共振理论;数据挖掘
中图分类号:TP311　　**文献标识码**:A

随着互联网的快速发展,网络信息量迅速增长,信息种类也日趋繁多,有书籍,研究论文,网络交流论坛,个人网站等,用户通过互联网了解所有种类的信息几乎是不可能的,然而他们必须花费越来越多的时间,从大量的信息流中剔除掉不相关的资源,检索到对自己有价值的信息。由于信息量的不断增大,这个检索的过程将会越来越烦琐。在互联网领域中,推荐系统(Recommender System)[1-2]作为解决这类问题的方法之一,已被广泛的应用于电子商务网站中[3-5],通过推荐商品和提供信息的方式帮助浏览者决定哪些商品需要购买。建议网站总交易记录中商品的销售量推荐商品。或者,基于对所有消费者以往的购买行为的分析,来进行推荐。被推荐的信息类型包括给用户建议合适的商品、提供个性化的商品信息和总结商品的群体反馈信息等等。

在线自动推荐机制的框架如图 1 所示。框架结构依赖于,由组偏好中产生的,并通过 ART 神经网络预处理的知识。当一个用户发起一个服务请求,系统会通过识别用户的偏好种类和找出适合用户偏好的相关规则来处理用户的个性化信息。当规则被激活并且合适的知识找到以后,推荐就会在线的展现给用户。

用户的族群偏好经 ART 神经网络预处理产生族群信息,结合数据挖掘技术得到的历史交易数据,产生自动聚类族群特征及其关联规则。当用户请求服务时,系统将识别出他的偏好类型,并提取与之相匹配的关联规则来处理用户个人信息。利用这些合理的规则和信息,系统就可给出及时的在线推荐。

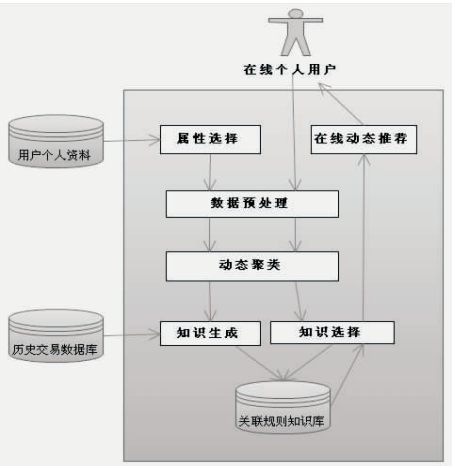

图 1　在线自动化推荐机制的框架

1　人工神经网络和自适应共振理论

1.1　人工神经网络

人工神经网络(Artificial Neural Network,即 ANN)[6]是一个十分有用的集群技术,是由生物学得到的灵感所开创的一种模仿生物神经网路的信息处理系统。它使用大量简单相连的人工神经元来模仿生物神经网路的

收稿日期:2011-03-21
基金项目:人工智能四川省重点实验室项目(2011RYY07)
作者简介:朱文忠(1971-),男,四川荣县人,副教授,硕士,主要从事计算机网络、嵌入式系统及应用方面的研究,(E-mail)zwz@suse.edu.cn

能力。人工神经元是生物神经元的简单模拟模式,它接收来自其他人工神经元或周围的刺激信息,透过大量简单的运算,最后输出其结果到外界环境或者其他人工神经元。

人工神经网络的结构由人工神经元(Neuron)和人工神经元间的连接权(Weight)组成,多数类型的神经网络将人工神经元以层(Layer)的形式组织在一起,其结构由输入层、输出层和至少一个隐含层,以及层间连接权组成,如图 2 所示。

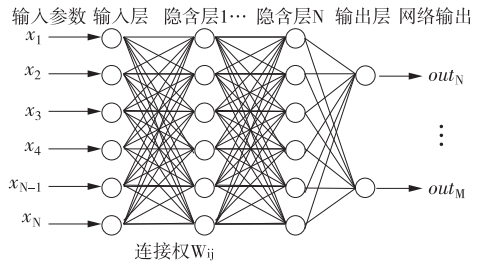

图 2 神经网络结构

人工神经网路的工作原理是将训练模式输入至输入层,并传至后面的隐含层,通过连接权向后传递,直至得到网络的输出。得到输出层的输出后,从这些输入向量及应该获得的输出值中,调整网路链结权重值,使用这些输入向量及调整权重值训练人工神经网路,一旦达到稳定状态,便是网路学习完毕。

1.2 自适应共振理论

自适应共振理论(Adaptive Resonance Theory,简称 ART)[7]是美国波士顿大学数学系自适应系统中心的 S. Grossberg 和 G. A. Carpenter 于 1987 年提出的一种神经网络模型。ART 是用生物神经细胞自兴奋与侧抑制的动力学原理指导学习,让输入模式通过网络双向连接权的识别和比较达到共振来完成自身的记忆,并以同样的方式实现联想。

ART 的原理是源于认知学习,也就是说,人类存储已知信息的记忆系统。当人类记忆新的信息(认知可塑性)时,要维护旧的记忆(记忆的稳定性)。然而,在这个时候,新的和旧的记忆可能发生混乱。应该有多少新的东西可以存储在内存中,有多少旧的记忆仍可维持,需要进行控制。

(1) 如果新的信息的特点与一些旧的记忆的东西足够的相似,那么只对记忆中的这部分内容进行修改。因此,无论是新老项目的功能,可同时容纳以满足稳定性的要求。

(2) 如果信息的特点与记忆的信息完全不同,那么系统就会为新的信息新建一个完全的新记忆。在这种情况下,学习过程将需要更少的时间。这样也可以满足内存需求的适应性。

自适应共振理论(ART)是一个不需要监督的网络研究模型(如图 3 所示)。它从已知的或者真实的输入数据中直接获取研究实例。对时序信号可以进行实时学习、实时处理,能对已经学习过的对象快速响应和自动识别。

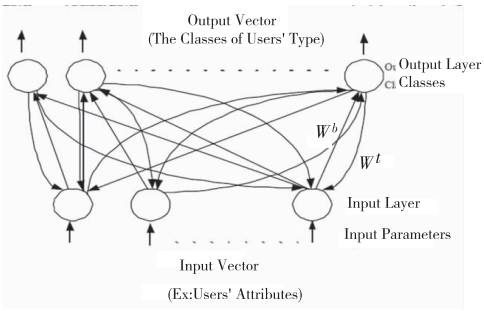

图 3 ART 网络的结构

2 关联规则挖掘(Association Rules Mining)

挖掘关联规则是数据挖掘中相当重要的一个议题,一般交易资料库中都储存着数量庞大的交易资料,而每一笔交易记录也都记载着相关的项目(item),包含使用者相关资料及交易的项目或时间等;而数据挖掘技术便可从这些大量的资料中,挖掘出各个项目之间的关联规则。例如:80% 的学生选修“C 语言”同时也会选修“数据结构”。

资料库中记录通常以 < TID, item > 的型式储存(TID 为每一笔纪录的唯一的识别码,Item 为记录中的项目)。设 $I = \{i_1, i_2, \ldots, i_m\}$ 代表所有元素或项目的集合,T 代表记录且 $T \subseteq I$, $D = \{T_1, T_2, \ldots, T_n\}$ 代表所有记录的集合,即一个记录数据库。

设 X 代表某些项目集合,称为项集,包含 k 个数据项的项集称为 k -项集。当且仅当 $X \subseteq T$ 时,称为记录 T 包含 X。

记录数据库 D 中项集 X 的支持度为: $support(X) = P(X) = \frac{|T_X|}{|D|}$,其中 $T_X = \{T \in D \mid X \subseteq T\}$。在用户给定的支持度阈值 minisupport 下,若 support $(X) \geqslant$ minisupport,则称 X 为频繁项集,否则 X 为非频繁项集。但为下文便于叙述,数据项集 X 的支持度是用 D 中包含 X 的事务数量来表示。

一个关联规则是“$X \Rightarrow Y$”形式的蕴含式,其中 $X \Rightarrow I, Y \Rightarrow I$ 且 $X \cap Y = \varphi$。如果 D 中包含事务 $X \cup Y$ 的百分比为 s,则称 s 为关联规则 $X \Rightarrow Y$ 的支持度,它是概率 $P(X \cup Y)$。如果 D 中包含 X 的事务同时也包含 Y 的百分比为 c,则称 c 为关联规则 $X \Rightarrow Y$ 的信任度,它是条

件概率 $P(Y|X)$)。即:

$$support(X \Rightarrow Y) = P(X \cup Y) = support(X \cup Y)$$

$$confidence(X \Rightarrow Y) = P(Y|X) = \frac{support(X \cup Y)}{support(X)}$$

挖掘关联规则的问题就是要生成所有满足 support ($X \Rightarrow Y$) ≥minisupport 和 confidence($X \Rightarrow Y$) ≥miniconfidence 的关联规则, 其中 minisupport 和 miniconfidence 分别为用户给定的最小支持度阈值和最小信任度阈值。同时满足这两个条件的关联规则称为强关联规则。

挖掘关联规则主要包含两个步骤(如图 4 所示),找出事务数据库 D 中所有的频繁项集。

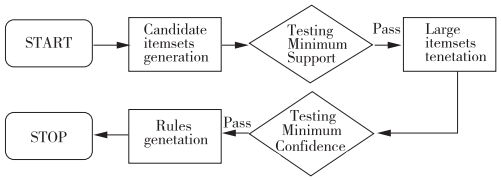

图 4　关联规则生成过程

根据获得的频繁项集产生强关联规则:假设找出的频繁项集为 $\{X,Y\}$, 其中可能产生的法则是 $X \Rightarrow Y$, 依此我们可以计算当 X 发生时,也发生 Y 的信任度,若是到达设定的最低信任度(Minimum Confidence Level),则此强关联规则就成立。

事实上,挖掘关联规则的整个执行过程中第一个子问题是核心。当找到所有的频繁项集后,相应的关联规则采用 ART 算法很容易生成。

3　ART 聚类算法描述

3.1　Apriori 算法

ART 算法中,Apriori 算法[8-9]是一种最有影响的挖掘布尔关联规则频繁项集的算法,步骤如下:

(1) 设 $I_1, I_2, \ldots, I_n$ 表示用户属性,与用户属性 I_n 对应的输入向量为

$$I_n = [X_1^n, X_2^n, \cdots, X_m^n]$$

t_n 表示第 n 个用户属性的网络输入点的个数,且 $X_{\in}^n$ $\{0,1\}$ 因此, 用于 ART 网络的输入向量可表示为:

$$[X_1^1, X_2^1, \ldots, X_{t1}^1, X_1^2, X_2^2, \ldots, X_{t2}^2, \ldots, X_1^n, X_2^n, \ldots, X_{tn}^n]$$

其中 $t1 + t2 + \cdots + tn = P$ 且 $t1, t2, \cdots, tn \geqslant 1$。$P$ 为输入点总个数。

(2)初始化输出点的个数 Nout = 1(Nout ≥ 1)。即,设 ART 网络开始只有一个初始输出点。

(3)初始化权重矩阵 W 为

$$W^t[i][1] = 1, W^b[i][1] = \frac{1}{1+p}$$

其中 P 是输入点的个数。

$W^t[i][1]$ 是连接从输出层到输入层的自适应权重,可称之为外权向量(假定开始时只有一个输出点),$0 \leqslant W^t \leqslant 1$。

$W^b[i][1]$ 是连接输入层到输出层的自适应权重,可称之为内权向量, $0 \leqslant W^b \leqslant 1$。

(4)在 ART 网络的输入层输入一个实验向量。

(5)输入向量与第 j 个输出集的匹配度表示为

$$net[j] = \sum_{i=1}^{p} W^b[i][j] \cdot X\{i\}, 1 \leqq j \leqq Nout$$

其中 $X[i]$ 为输入向量,其值为二进制, $1 \leqslant i \leqslant P$。

(6)找到与输入向量有最大匹配度的第 j^* 个输出集

$$net[j^*] = max\ net[j^*], 1 \leqq j \leqq Nout$$

其中 $j* \in j$。

(7)计算输入向量与第 j^* 个输出集的相似度 V_j^*,表示为:

$$\|X\| = \sum_{i=1}^{p} X[i] \tag{1}$$

$$\|W_j^{t*} \cdot X\| = \sum_{i=1}^{p} W^t[i][j^*] \cdot X[i] \tag{2}$$

$$V_j^* = \frac{\|W_j^{t*} \cdot X\|}{\|X\|} \tag{3}$$

(8)测试是否存在另一个相似的输出集。设 Icount 表示为已被使用的输出集个数,初始值为 0,如果 Icount < Nout,计算出输入向量与剩余输出集的下一个最大匹配度 net [j],令 Icount = Icount + 1, net[j^*] =0, 转入步骤(6)。否则,转入步骤(4)(输入一个新的向量 X)。

(9)如果 $V_{j*} > r$, 即,境界测试通过, 那么

$$W^t[i][j^*] = W^t[j^*] \cdot X[i], and$$

$$W^b[i][j^*] = \frac{W^t[i][j^*] \cdot X[i]}{0.5 + \sum_{i=1}^{p} W^t[i][j^*] \cdot X[i][j^*]}$$

(10)如果没有新的聚类生成且正好完成一个学习循环,则输出聚类结果并终止执行。否则转入步骤(4)输入新向量 X。

3.2　ART 算法改进(MART)

ART 的相似度是比较最大匹配度节点 j^* 的外权向量 $W_{j^*}^t$ 与输入向量 X 对应位子中“1”的个数,其表达式为

$$V_{j^*} = \frac{\|W_{j^*}^t X\|}{\|X\|} = \frac{\sum_{i=1}^{p} W^t[i][j^*]x[i]}{\sum_{i=1}^{p} x[i]}, i = 1, \ldots, p$$

网络第一次输入向量 X_1 时 $X = X_1$, 与第一个节点匹配相似,外权向量的权值进行调整,计算出第一个节

点调整的内外权向量。第二次输入向量 X_2 时 $X = X_2$，计算 X_2 与输出层各个节点的匹配相似度，可知在第一个节点处最大(大于其它未使用的节点)，按公式(3)，它的相似度公式得到的值 V_1 为 1.000，这说明向量 X_2 与 X_1 是完全相同的，然而我们可以清楚地看到 X_2 与 X_1 是有差异的。而将向量 X_1 和 X_2 的输入次序调换一下，根据公式(3)得到的相似度 V_1 为 17/18 =0.944。就 X_1 与 X_2 而言，在不同的输入次序下却得到不同的相似度。这是因为在比较相似度的时候，只要 $W^t[i][j^*]$ 为0，在其对应的位置上 $x[i]$ 也为0，此时不论 $W^t[i][j^*]$ 为1的对应位置上，输入 $x[i]$ 为1还是为0，所得相似度一定为1。

4 MART 算法实验与分析

在仿真实验中，为了演示新算法 MART，在一个实例中选择四个用户属性，它们分别是职业，教育背景，性别和专业。分别用四位来代表这些属性，每一位的具体情况见表1。选取 10 个随机生成的用户(表2)，每个用户具有不同的属性。最后使用 MART 算法得到的聚类结果和 ART 算法得到的结果进行分析。

表 1 用户属性映射表

位号	1		2		3		4			
含义	职业		学历		性别		专业			
每一位的表示规则	学生	职员	本科以下	本科及以上	男	女	计算机应用	软件工程	应用化学	艺术设计
	0	1	0	1	0	1	00	01	10	11

表 2 用二进制输入向量表示的用户属性表

用户 ID	用户属性
U_{01}	00100
U_{02}	00101
U_{03}	01101
U_{04}	01010
U_{05}	00000
U_{06}	10000
U_{07}	11101
U_{08}	11111
U_{09}	10111
U_{10}	10010

基于不同的权重属性，MART 算法将进行三次实验。在实验 1 中，每个属性具有相同的权重值 $M^t[i]$，这就意味着每个属性的重要性是一样的。结果显示用户被分成四个聚类。实验结果见表3。

在实验 2 中，跟其它位相比，第一位和第二位设置的值较高，这就意味着职业和教育背景这两个属性被视作较为重要的考虑因素。实验结果显示具有相同教育背景和职业的用户将被划分到相同的聚类。实验结果见表4。

表 3 每个属性具有同等的重要性(r =0.800)

由 W^t 划分得到的聚类结果	$M^t[i]$ =[1,1,1,1,1] 输入模式		
00000	00100(U_{01})	00101(U_{02})	00000(U_{05})
	10000(U_{06})	10010(U_{10})	
01010	01010(U_{04})		
01101	01101(U_{03})	11101(U_{07})	
10111	11111(U_{08})	10111(U_{09})	

表 4 前两位较重要(r =0.667)

由 W^t 划分得到的聚类结果	$M^t[i]$ =[3,3,1,1,1] 输入模式		
00000	00100(U_{01})	00101(U_{02})	00000(U_{05})
01000	01101(U_{03})	01010(U_{04})	
10000	10000(U_{06})	10111(U_{09})	10010(U_{10})
11010	11101(U_{07})	11111(U_{08})	

在实验 3 中，第二位、第四位和第五位被设置为较高的权重值。同时，第四位和第五位的权重值比第二位的权重值高。这就意味着专业是最重要的考虑因素，然而，学历成为较重要的考虑因素。实验结果显示具有相同专业的用户较为容易被分到同一个聚类中去。实验结果见表5。

表 5 2^{nd}4 th 和 5 th 位较重要 (r =0.667)

由 W^t 划分得到的聚类结果	$M^t[i]$ =[1,3,1,5,5] 输入模式		
00000	00100(U_{01})	00000(U_{05})	10000(U_{06})
00101	00101(U_{02})	01101(U_{03})	11101(U_{07})
00010	01010(U_{04})	10010(U_{10})	
10111	10111(U_{09})	11111(U_{05})	

这样具有相同专业的用户被划分到了相同的聚类。其中，U_{02} 和 U_{01} 并不是同一个专业却被划分到了同一个聚类。显然通过 MART 算法的相似值计算公式，推荐系统可以根据用户属性的重要性来设置每一个属性节点的权重。这样，与传统的 ART 算法相比，可以得到更为合理和灵活的输出结果。

5 结束语

结合了人工神经网络和数据挖掘技术，展示了一个新的自动化推荐机制，即一个以用户为导向的推荐机制。该推荐机制首先对经典的 ART 算法进行优化，以优化后的 ART 算法产生用户聚类，由此得出每个用户的类型，然后通过实例和仿真验证了算法的有效性。

参 考 文 献:

[1] Sarwar B,Karypis G,Konstan J,et al.Analysis of Recommendation Algorithms for E-commerce[C].ACM Conference on Electronic Commerce,2000,158-167.

[2] Yu P S.Data Mining and Personalization Technologies

[C]. The 6th International Conference on Database Systems for Advanced Applications,1999,6-13.

[3] Hill W C, Stead L, Rosenstein M, et al. Recommending and evaluating choices in a virtual community of use [C]. Proceedings of CHI'95,1995,194-201.

[4] Konstan J,Miller B,Maltz D,et al. GroupLens Applying Collaborative Filtering to Usenet News[J]. Communications of ACM,Vol.40,No.3,1997,40(3):77-87.

[5] Shardanand U, Maes P. Social Information Filtering: Algorithms for Automating 'Word of Mouth' [C]. Proceedings of the Computer-Human Interaction Conference(CHI'95),1995.

[6] Patterson D W, Neural Network Learning: Theory and Application[M]. New York:Prentice Hall,1996.

[7] Carpenter G A,Grossberg S.A Massively Parallel Architecture for a Self -Organizing Neural Pattern Recognition Machine[J]. Trans. IEEE on Computer Vision, Graphics and Image Processing,1987,37(1):54-115.

[8] 陆丽娜,陈亚萍.挖掘关联规则中 Apriori 算法的研究[J].小型微型计算机系统,2000,21(9):940-94

[9] 罗 可,吴 杰.一种基于 Apriori 的改进算法[J].计算机工程与应用,2001,5(22):20-22.

Algorithm of Automatic Recommender System Based on Data Mining

ZHU Wen-zhong

(School of Computer, Sichuan University of Science & Engineering, Zigong 643000, China)

Abstract: A typical online recommendation system is described. By using ART neural network and data mining technology, it can automatically cluster population characteristics and can dig out the associated characteristics. Aim at the online recommendation system applied on network, how effectively use data mining techniques to mine the complete knowledge from a large number of databases is discussed, then the appropriate information is recommended to users to help them to find really needed and useful documents or information in the vast flow of information. A pattern is put forward that combines ART neural network and data mining technology. Aim at the characteristics of recommendation system, a new modified ART algorithm(MART algorithm) is proposed. The result shows that the proposed algorithm is effective.

Key words: automatic recommender system; neural network; ART; data mining