3 Chapter

第3章　商业数据流概念漂移模型

商业数据流中的概念漂移现象越来越受到人们的关注，所谓数据流概念漂移是流数据集中一个变化的目标概念，其显著特征是随时间动态变化，由于发生改变的原因常常是隐式的、事先未知的，学习任务因这些隐式的变化而变得更加复杂。本章针对商业数据流的概念漂移现象，构建有效处理商业数据流概念漂移问题的模型，包括商业数据流概念漂移描述模型、概念漂移特征提取模型，以及概念漂移检测模型。

3.1　商业数据流概念漂移描述模型

3.1.1　商业数据流中的概念漂移概述

数据流中的概念漂移涉及的是流数据集中一个变化的目标概念[1]。数据流的显著特征是随时间动态变化，由于发生改变的原因常常是隐式的、事先未知的，学习任务因这些隐式的变化而变得更加复杂。数据流中的目标概念随着上下文的隐式改变而或多或少地发生根本性改变。数据特征的改变使得目标分类模型或目标概念随时间改变的现象被称为概念漂移(Concept Drift)。概念漂移在整体上分为两类，一类是突变，另一类是渐变。突变指的是数据流中的数据从一种分布率变为另外一种分布率，而渐变指的是数据流中的分布率以一个常数随时间不断地变化。

数据流的动态变化特征是人们感兴趣的概念，亦即目标概念常常发生根本性的改变[2]。例如，顾客购买偏好随购买时间、购买对象的范围、通货膨胀率、个体背景、环境情境等因素的改变而发生变化，天气预测规则随季节、地域等因素的改变而发生变化。造成商业数据流漂移特性的重要原因是各种商业信息的变化，包括商品季节特性、顾客兴趣、供货商供货能力、销售商商业利润等，它是商业数据流动态变化的重要成因，也是商业数据流有别于其他领域数据流的显著特点。由于商业数据流的漂移特性造成数据流的动态变化，使数据挖掘结果往往难以实时、准确地反映真实商业过程，这就要求数据挖掘具有更高的可靠性。

概念漂移挖掘的结果主要应用在监测控制、个人助手、决策支持和人工智能等方面[3]。

监测控制的应用包括监测 Web 网络上竞争对手的活动、计算机网络、通信和金融交易，个人信息助手的应用包括推荐系统、文本信息的分类和组织、市场营销中的客户分析等，决策支持的应用包括诊断、信用价值评估等，人工智能方面的应用则是那些不断地和变化的环境进行交互机器人、移动车辆和智能家电等。不同的应用在概念漂移学习速度、预测准确性、犯错误的成本等方面有不同程度的要求，这正是数据流中概念漂移挖掘的挑战所在。

3.1.2　基于粒计算的商业数据流概念模型

粒计算，作为一个新兴研究领域，为研究数据挖掘的许多问题提供概念框架[4-6]。Yao[7]提出了一个基于粒计算的规则挖掘框架，该框架基于研究形式化和数学建模方法提出数据挖掘的粒计算模型。模型中认为数据是被有限的对象集合和有限的属性集描述的，并且使用属性集合来定义内涵(intension)，论域对象的子集来定义外延(extension)。一个概念(concept)就是由外延和内涵组成的，其中概念的内涵表达挖掘规则，通过概念的外延来解释。Yao 所述的粒计算数据挖掘模型框架就是基于概念外延定义的粒度来构建的[7]。参考文献[8]中探讨了基于粒计算的数据、知识表示和处理，论述了数据挖掘的基础任务是对数据、知识表示寻求正确层次的粒度。但是参考文献[8]中论述的是粒计算对静态数据建模，不适用于动态的流数据。

粒计算中一种常见的概念分析方式是形式概念分析(Formal Concept Analysis，FCA)，FCA 已经被广泛应用到数据挖掘、信息抽取，以及程序聚类[9]中。参考文献[10]中，在分析问题时首先形成一个在形式概念分析中称之为形式化语境的二维表，并进一步转换为 Hasse 图。这种方法能建立概念与概念、概念与特征以及特征与特征之间的内在联系，形成概念格的信息表示结构。因而可在对数据进行粒度分析的基础上，采用概念形式化方法描述流数据，继而进行特征挖掘。

商业数据流的特征决定了数据中包含不完整的、不确定的或者模糊的概念，传统的数据挖掘方法不能获取理想的知识。使用 GrC 和 FCA 方法，即使不能得到精确解，至少可以较低的代价挖掘得到近似的知识。因此，本文采用 GrC 方法来粒化数据信息，采用 FCA 方法来分析概念间的关系，最终挖掘隐含概念漂移的数据流特征。

概念漂移问题中提取概念漂移特征和发生概念漂移的规则，是常见的数据流分类学习算法无法解决的。面对复杂的、难于准确把握的问题时，通常采用由粗到细、不断求精的多粒度分析法，避免复杂的计算，从而获得足够满意的解，使得原来看似非多项式的难解问题迎刃而解[11]。粒计算(Granular Computing，GrC)方法是人工智能领域中的一种新理念和新方法，它覆盖了所有和粒度相关的理论、方法和技术，主要用于对不确定、不精确、不完整信息的处理，对大规模海量数据的挖掘和对复杂问题的求解[4]。粒计算的实质是通过选择合适的粒度，来寻找一种较好的、近似的解决方案，从而降低问题求解的难度。

运用粒计算思想能够处理数据挖掘领域中的多类问题，面对日益复杂的海量、高维、动态数据，将复杂数据分层次或分批次地处理是数据挖掘的可行方法。Lin 等人[5]提出了利用信息颗粒的位表示来进行关联模式挖掘的思想，并用于关系数据库建模和关联规则挖掘中。从粒度的角度看，超级频繁模式可称为“粗粒”，子频繁模式可称为“细粒”，频繁模

式的父子关系相当于粒空间中的一种偏序关系。

从粒计算的角度，可以对动态数据环境下的数据挖掘方法进行这样的理解：原始的知识库可以构成一个原始的知识粒，该知识粒由多个知识子粒构成。当新数据到来时，寻找合适的知识子粒或子粒集合对新数据进行判断和处理，并更新知识子粒或知识子粒集合，在一定的时间粒度下进行粒子合并，可得到新知识粒。德国的 Wille 教授于 1982 年提出了形式概念分析理论，概念格是形式概念分析的基本数据结构，它的每个节点都是一个形式概念。形式概念由两部分组成：外延，即概念所覆盖的对象；内涵，即概念的描述，亦即概念所覆盖对象的共同属性。概念格通过 Hasse 图直观而间接地体现了这些概念之间的泛化和特化关系。当前大多数成果都是针对提高隐含概念漂移数据流分类准确率的，很少考虑概念漂移本身的特征，缺少隐含概念漂移的数据流形式化描述、概念特征提取以及发生概念漂移的原因等方面的研究，采用粒计算方法对粒化的概念进行形式化描述，分析概念之间的耦合关系，提取概念变化特征。数据流概念漂移的分类挖掘是对概念漂移数据所属类别的预测，极少有文献涉及概念漂移发生原因的挖掘，但漂移发生的规则对于挖掘变化的原因、商业决策具有重要意义，最近几年已经开始有文献涉及[3][12,13]。

通过数据流的粒化，对概念进行粒的表示、特征化、描述和解释。基于粒计算的概念分析主要基于以下步骤：概念分层，采用粒计算模型中的概念格、粒度划分；建立概念之间关系；描述概念的外延和内涵，对属性和对象进行描述，表明概念之间的泛化关系；通过对概念的外延耦合度、内涵耦合度和概念耦合度的分析，挖掘漂移特征。图 3-1 是概念形式化分析流程。

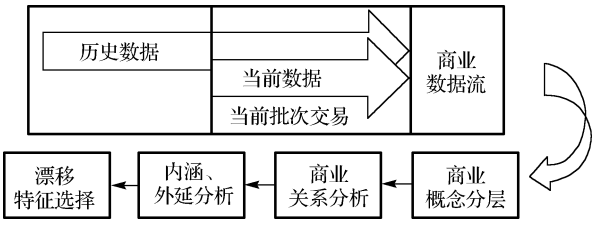

图 3-1　商业数据流概念形式化分析流程

模型是解决问题的基础，数据流模型可以采用数据流属性集合的方式来表示。数据流用多维元组的属性集合表示，这些属性包括时间、事务集合、限定的非空属性集等。

定义 3-1　数据流。$DS = \{TP, U, A\{O_a \mid a \in A\}, \{D_a \mid a \in A\}, \{I_a \mid a \in A\}, \{\sigma_a \mid a \in A\}, \{\lambda_{u,a} \mid a \in U, a \in A\}\}$。$TP$ 是一个时间段（TP 可以设置 1 秒、1 分钟或一个小时等，是时间层的粒度划分），可以表示为 $<\tau_b; \tau_e>$，其中 τ_b 为时间段的开始时间，τ_e 为时间段的结束时间，它决定了由具体数据流构成的信息集合的大小。

U 是有限的非空对象集，在一个具体的时间段内的交易数据流集合是有限的，它的基数（cardinality）就是在内的数据流长度 $DS.Length$。

A 是有限的非空属性集。

O_a 是概念的外延。

D_a 是概念的内涵。

I_a 是对象到属性的映射，亦称为隶属度函数。

σ 是概念的内涵(intension)对应于每个属性的平均隶属度(average membership),体现了这个概念具有各个属性的程度。

λ 表示概念内涵中各对象对于各个属性的隶属度偏离平均值的平均程度,它体现了这个概念的发散(divergent)程度。

定义 3-2　模糊形式背景[8]。一个模糊形式背景表示为 $K = (U,A,I)$,其中,U 为对象集,A 为属性集,映射 I 称为隶属度函数,它满足 $I: U \times A \rightarrow [0,1]$。

定义 3-3　窗口。对于模糊形式背景中的每个属性,选取两个阈值 θ_d 和 ψ_d,满足 $0 \leqslant \theta_d \leqslant \psi_d \leqslant 1$。$\theta_d$ 和 ψ_d 构成窗口,其上沿和下沿分别为 θ_d 和 ψ_d。该阈值既可以根据领域背景知识来确定,也可以由用户根据应用的目的而指定。

定义 3-4　映射 f 和 g。在模糊形式背景 $K = (U,A,I)$ 中,$O \in P(U)$,$D \in P(A)$(P 是幂集符号),在 $P(U)$ 和 $P(A)$ 间可定义两个映射 f 和 g:

$$f(O) = \{d \mid \forall o \in O, \theta_d \leqslant I(o,d) \leqslant \psi_d\};$$

$$g(D) = \{o \mid \forall d \in D, \theta_d \leqslant I(o,d) \leqslant \psi_d\}$$

定义 3-5　模糊参数 σ 和 λ。对于对象集 $O \in P(U)$ 和属性集 $D \in P(A)$,其中,$D = f(O)$,$o \in O$,$d \in D$,$|O|$ 和 $|D|$ 分别是集合 O 和集合 D 的基数,如果 $|O| \neq 0$、$|D| \neq 0$,则

$$\sigma_d = \left[\sum\nolimits_{o \in O} I(o,d)\right] / |O| \tag{3-1}$$

$$\sigma = \sum\nolimits_{d \in D} (\sigma_d / d) \tag{3-2}$$

$$\lambda_d = \sqrt{\left\{\sum\nolimits_{o \in O} [I(o,d) - \sigma_d]^2\right\} / |O|} \tag{3-3}$$

$$\lambda = \left(\sum\nolimits_{d \in D} \lambda_d\right) / |D| \tag{3-4}$$

式(3-2)中 Σ 是模糊集合中的符号,式(3-3)中 Σ 是通常的累加符号。当需要指明具体对象集和属性集 (O,D) 时,模糊参数和 σ 分别记作 $\sigma_d(O,D)$ 和 $\sigma(O,D)$,模糊参数和 λ 分别记作 $\lambda_d(O,D)$ 和 $\lambda(O,D)$。

定义 3-6　模糊概念[9]。如果对象集 $O \in P(U)$ 和属性集 $D \in P(A)$ 满足 $O = g(D)$ 和 $D = f(O)$,则 $C = (O,D,\sigma,\lambda)$ 是 K 的一个模糊概念。O 和 D 分别是 C 的外延和内涵,σ 和 λ 分别依据式(3-2)和式(3-4)计算。σ 和 λ 这两个参数对于基于模糊概念格的模糊规则提取、概念聚类等应用有重要的作用。

3.2　商业数据流概念漂移特征提取模型

3.2.1　商业数据流概念漂移特征发现模型

数据流中的概念漂移(Concept Drift)[1]是指由于数据流中上下文变化而导致所隐含的目标概念发生变化甚至是根本性改变的现象规律,其反映了实际领域的特征变化。由于目

标概念漂移的原因及上下文的变化常常是隐式的、事先不能预知的。因此，基于概念漂移特征发现的数据流挖掘与知识发现成为一项复杂的任务，备受国内外学者的关注并取得了一定的成果。目前常用的有三种概念漂移特征发现，分别是基于渐进式漂移特征发现、基于突变式漂移特征发现，以及同时基于渐进式和突变式两种漂移特征发现。

(1)基于渐进式的漂移特征发现。Hulten 等人于 2001 年提出基于决策树模型的分类概念漂移发现的算法 CVFDT(Concept-adapting Very-Fast Decision Tree learner)[14]；Haixun Wang 等人于 2003 年提出基于加权的整合决策树分类器处理概念漂移的数据流挖掘方法[15]；WeiFan 于 2004 年提出基于随机决策树处理数据流中概念漂移的方法[16][17]；王勇等人于 2006 年就如何发现模式中的变化和更新模型以反映这种变化的问题，提出一种基于累积和(CUSUM)控制图的变化发现方法[18]。

(2)基于突变式的漂移特征发现。W. Nick 等人于 2001 年提出基于多决策树的流整合算法 SEA[19]，2005 年提出进化数据流中的变化诊断算法[20]，2006 年提出基于需求处理进化数据流的分类框架[21]；HaiXun Wang 等人于 2006 年提出基于马尔可夫模型的整合分类器[22]；Yi Zhang 等人于 2006 年针对一些处理数据流的整合学习分类方法对概念漂移不敏感的问题提出了基于自动构建、组织机制的整合学习数据流 DCO(Dynamic Construction and Organization)方法。

(3)基于渐进式和突变式结合的漂移特征发现。吴信东等人于 2005 年针对数据流面临的挑战提出 Proactive-Reactive Prediction 模型[23]；Joao 等人于 2005 年提出利用超快森林树 UFFT(Ultra Fast Forest of Trees)模型处理动态数据流的方法[24][25]；Jiong Yang 等人于 2005 年提出一个有效的进化分类器 EvoClass(Evolutionary Classifier)算法[26]；王勇等人于 2006 年定义了一种相反分类器从错误中学习，提出了训练集合分类器对具有概念流动的数据流进行分类的算法 IWB[27]。

总体而言，研究人员利用各种方法来缓和、避免或者解决数据流漂移问题。主要有两大类：第一大类的方法并不去检测数据流是否发生了漂移，而是认为数据流一直在发生变化，然后利用数据流挖掘管理模型的自我更新方法来不断缓和漂移给系统带来的负面影响；第二大类的方法则是明确地检测出数据流漂移发生的位置，然后根据检测结果再对数据流进行建模。与第一类方法相比，这类方法多了一步检测漂移，虽然在计算量上有所增加，但是漂移问题也得到了真正的解决。

近年来也有学者采用融入粒计算思想处理数据流漂移特征发现。

3.2.2 商业数据流概念漂移特征提取模型

数据流的概念漂移特征提取属于异构数据流信息提取范畴，而信息提取已经成为自然语言处理的一个重要方向，理论研究不断得到发展。通常信息提取可分为显式信息提取和隐式信息提取。

1. 显式数据流信息提取

商业数据流的显式信息提取要求显式提供各种数据，其中以用户数据为主，表现为在

线用户主动提供，即由用户主动填写、提供来获取用户的数据信息，主要包括：用户将自己感兴趣的信息或在线文档分类后提供给系统，系统从这些文档或信息中发现用户的兴趣；用户提供自己的研究方向和其他阅读爱好等信息，系统从这些信息中发现用户的兴趣；用户对系统检索到的信息结果进行评价打分，系统通过用户反馈信息来更新用户的兴趣数据描述。这种收集方式简单、直接，有助于加速学习算法的速度，但它要求用户确知其兴趣并花费相应的时间和精力积极参与，但多数情况下这两者都不成立。

2. 隐式数据流信息提取

采用隐式方式时，系统通过分析用户的上网数据，提取信息，主要包括服务器端挖掘和系统被动学习两种方法。服务器端挖掘可以通过以下两种方式来实现从服务器端获取用户的相关信息，即一般的访问模式挖掘和个性化的使用记录挖掘。一般的访问模式挖掘通过分析用户使用记录来了解用户的访问模式和倾向，而个性化的使用记录挖掘则倾向于分析单个用户的偏好，其目的是根据不同用户的访问模式，为每个用户提供定制的站点。系统被动学习，即监视用户的信息搜索与浏览过程等使用习惯来获取用户的兴趣信息。与显式用户信息获取方法相比，隐式用户兴趣信息获取并不需要用户显式地提供相关的兴趣信息，它在不打扰用户正常使用个性化服务系统的情况下，自动获取用户兴趣相关消息。

通常在进行信息数据提取的时候会结合显式和隐式两种方式。相应的信息提取模型有 3 种[28]，即基于词典的提取模型、基于规则的提取模型和基于隐马尔可夫模型(Hidden Markov Model，HMM)的提取模型。基于词典的文本信息提取模型需要首先构造提取模式词典，然后使用该模式词典从未标记文本中提取所需信息。基于规则的文本信息提取模型也需要先构造提取规则集。由于基于词典和基于规则的提取模型算法比较复杂，适应性较差，提取精度不高，因此学者们提出了基于 HMM 的提取模型。利用 HMM 进行信息提取是一种基于机器学习的信息提取方法，它具有易于建立、适应性强、提取精度高等优点。

隐马尔可夫模型(HMM)是一个二重马尔可夫随机过程，包括具有状态转移概率的马尔可夫链和输出观测值的随机过程，其状态只有通过观测序列的随机过程才能表现出来。隐马尔可夫模型提供了一种基于训练数据的概率自动构造识别系统。一个隐马尔可夫模型一般由以下五个元素组成。

(1)初始状态分布 $\pi = \{\pi_i\}$

其中，$\pi_i = P(q_1 = s_i)$，$\sum_{i=1}^{N} \pi_i = 1$，$\pi_i \geqslant 0$。

(2)状态集合 $S = \{s_1, s_2, \cdots, s_N\}$

其中，N 为状态个数。

(3)观测值集合 $V = \{v_1, v_2, \cdots, v_M\}$

其中，M 为观测值的个数。

(4)状态转移概率矩阵 $A = [a_{ij}]_{N \times M}$

其中，$a_{ij} = P(q_{t+1} = s_j \mid q_t = s_i)$，$1 \leqslant i, j \leqslant N$。

(5)观察值概率矩阵 $B=[b_j(k)]_{N\times M}$

其中，$b_j(k)=P(o_t=v_k \mid q_t=s_j)$，$1\leqslant j\leqslant N$，$1\leqslant k\leqslant M$

这样，一个隐马尔可夫模型可以由五元组(π, S, V, A, B)完整描述。π 是初始状态概率集合，π_i 是状态 s_i 作为初始状态的概率；S 是模型共有的 N 个状态的集合；V 是模型共有的 M 个可能输出的观测值的集合；A 是 $N\times N$ 的状态转移概率矩阵，a_{ij} 是从状态 s_i 转换到状态 s_j 的概率；B 是 $N\times M$ 的释放概率矩阵，$b_j(k)$ 是在状态 s_i 时释放观测值 v_k 的概率，由于 A，B 中包含了对 S，V 的说明，因此一个隐马尔可夫模型通常可简记为 $\lambda=(\pi,A,B)$。

实际应用时首先建立模型，从训练文本集中统计出来各个状态对应词表 m_i 的大小($1\leqslant i\leqslant N$，N 是状态的个数)和总词表 $m(m=m_1+m_2+\cdots+m_N)$的大小，并在文本分块的基础上，统计出来各个状态到其他 N 个状态的转移个数，以及作为初始状态的个数。依据上面的统计数据，根据下列公式计算隐马尔可夫模型 $\lambda=(\pi,A,B)$的参数值。

(1)初始状态概率分布：$\pi=\{\pi_1,\pi_2,\cdots,\pi_N\}$

$$\pi=\frac{\text{第 } i \text{ 个状态作为初始状态出现的个数}}{\text{训练文本中初始状态的总个数}},\ \sum_{i=1}^{N}\pi_i=1,\pi_i\geqslant 0 \tag{3-5}$$

(2)状态转移概率矩阵：$A=a_{ij}$

$$a_{ij}=\frac{\text{训练文本中状态 } i \text{ 转移到状态 } j \text{ 的次数}}{\text{训练文本中状态 } i \text{ 转移到所有状态的总次数}},\ \sum_{j=1}^{N}a_{ij}=1,\ a_{ij}\geqslant 0 \tag{3-6}$$

(3)从状态 j 观察到单词 k 的概率分布矩阵

$$b_j(k)=\frac{\text{单词 } k \text{ 在 } m_j \text{ 词表中出现的总个数}}{m_j \text{ 词表的大小}},\ \sum_{k=1}^{N}b_j(k)=1,\ b_j(k)\geqslant 0 \tag{3-7}$$

HMM$^{(2)}$相对 HMM$^{(1)}$来说，考虑了历史状态模型的关联性，对于提取的精度会有很大的提高。应用 HMM$^{(2)}$进行用户兴趣的特征提取，首先将已标记的样本进行预处理，并用 BW 算法对其进行训练，通过初始化构建其 HMM$^{(2)}$模型，然后将预处理后的待提取特征利用 Viterbi 算法代入模型进行解码，从而输出所需的标记序列，即提取的兴趣特征序列。其模型如图 3-2 所示。

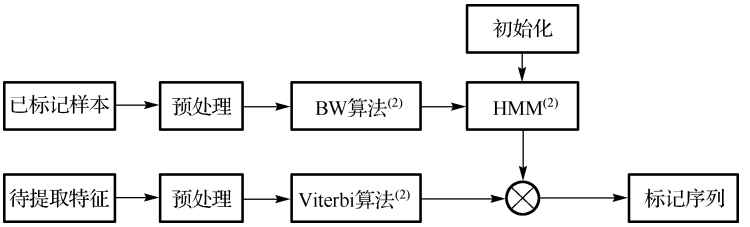

图 3-2　基于 HMM$^{(2)}$的兴趣特征提取模型

以用户兴趣获取为例，基于 HMM$^{(2)}$模型的特征提取分为以下两大步骤。

(1)训练部分

① 将用户兴趣的一些特征信息先后顺序进行预处理，形成文本文档，其中将搜索关键词(keywords)、用户兴趣页面主题(topics)、浏览页面的时间(times)、增加书签(book)、保

存页面(savepage)、拖动滚动条(scroll)、点击链接(links)七个特征作为隐含状态序列。

② 对已标记训练样本进行预处理，采集服务器和客户端的数据，经过初步处理后形成文本，对文本经过扫描后，利用分隔符、空格、换行、冒号等排版将已标记文本序列转换为标记的文本分块序列。

③ 用 HMM[(2)] 模型需对其计算以下模型参数，其参数的确定算法如公式(3-8)~(3-12)。

- 初始概率分布矢量

$$\pi_i = \frac{\mathrm{Init}(i)}{\sum_{j=1}^{N} \mathrm{Init}(j)}, 1 \leqslant i \leqslant N_s \tag{3-8}$$

其中，$\mathrm{Init}(i)$ 是指已标记的整个训练样本中，以状态 S_i 为开始状态序列的个数，$\sum_{j=1}^{N} \mathrm{Init}(j)$ 则是指以所有状态为开始状态序列的个数总和。

- 初始状态转移概率

$$a_{ij} = \frac{C_{ij}}{\sum_{k=1}^{N} C_{ik}},\ 1 \leqslant i,j \leqslant N \tag{3-9}$$

$$a_{ijk} = \frac{C_{ijk}}{\sum_{u=1}^{N} C_{iju}},\ 1 \leqslant i,j,k \leqslant N \tag{3-10}$$

其中，C_{ij} 和 C_{ijk} 分别表示从状态 S_i 到 S_j 的转移次数，以及 $t-1$ 时刻的状态 S_i，t 时刻的状态 S_j，转移到 $t+1$ 时刻状态为 S_k 的次数。$\sum_{k=1}^{N} C_{ik}$ 和 $\sum_{u=1}^{N} C_{iju}$ 分别表示从状态 S_i 到所有状态的转移次数之和，以及 $t-1$ 时刻的状态 S_i，t 时刻的状态 S_j，转移到所有状态的次数之和。

- 观察值释放概率

$$b_j(O_k) = \frac{E_j(O_k)}{\sum_{i=1}^{M} E_j(O_i)},\ 1 \leqslant j \leqslant N \tag{3-11}$$

$$b_{ij}(O_k) = \frac{E_{ij}(O_k)}{\sum_{i=1}^{M} E_{ij}(O_u)},\ 1 \leqslant i,j \leqslant N,\ 1 \leqslant k \leqslant M \tag{3-12}$$

其中，$E_j(O_k)$ 和 $E_{ij}(O_k)$ 分别表示状态 S_j 时释放观察值 O_k 的次数，以及 $t-1$ 时刻的状态 S_i，t 时刻状态 S_j，释放观察值 O_k 的次数。$\sum_{i=1}^{M} E_j(O_i)$ 和 $\sum_{i=1}^{M} E_{ij}(O_u)$ 分别表示状态 S_j 时释放所有观察值的次数之和，以及 $t-1$ 时刻的状态 S_i，t 时刻的状态 S_j，释放所有观察值的次数之和。

(2)提取部分

① 对待提取特征的文本进行预处理，对文本经过扫描后，利用分隔符、空格、换行、冒号等排版将已标记文本序列转换为标记的文本分块序列。

② 结合训练部分输出的 HMM[(2)] 模型，利用 Viterbi 算法进行计算。应用已建立好的

HMM[2]模型进行用户兴趣特征提取。将处理得到后的状态输出观察值 $O = O_1O_2\cdots O_T$ 作为模型输入，从中找出状态标签序列中概率最大的，用户特征提取的内容就是被标记为目标状态标签的观察文本。

3.3 商业数据流概念漂移检测模型

3.3.1 基于概念格的数据流漂移检测模型

概念包括内涵和外延，用二元组 $<O_a; D_a>$ 表示，其中 O_a 是 *DS* 的外延，D_a 是 *DS* 的内涵。概念之间的关系主要是基于概念的外延的[7]。随着时间变化，数据流不断更新，一个具体时间段内的 *DS* 集合中的元素也发生变化：如果集合的元素被添加或削减，那么一个概念的外延也可能会扩大、减少或保持；如果在同一时间段内同时由元素添加和削减，那么概念的外延也可能会扩大、减少或保持。总之，在不同的时间段内外延和内涵都可能发生变化，假设一个时间段 $<\tau_b; \tau_c>$ 中的概念为 $<O_a; D_a>$，时间来到 $<\tau'_b; \tau'_c>$，概念变化为 $<O'_a; D'_a>$。$<\tau_b; \tau_c>$ 时间段内概念集合所构成的概念格设为 CL_1，其后续时间段 $<\tau'_b; \tau'_c>$ 内的概念集合构成的概念格设为 CL_2。

定义 3-7 外延耦合度。c_1，c_2 分别是数据流 *DS* 中包含的概念，且分别由 *DS* 的外延 O_a 和 *DS* 的内涵 D_a 组成；c_1，c_2 的外延耦合度 *coe* 定义为 $coe = 2 * |A(O_{a_1}, O_{a_2})| / (|O_{a_1}| + |O_{a_2}|)$，其中 O_{a_1} 是 c_1 的外延，O_{a_2} 是 c_2 的外延，$A(O_{a_1}, O_{a_2})$ 是 c_1、c_2 外延集合的交集，$|A(O_{a_1}, O_{a_2})|$、$|O_{a_1}|$ 和 $|O_{a_2}|$ 分别表示各自外延的基数。

定义 3-8 内涵耦合度。c_1，c_2 的内涵耦合度 *coi* 定义为 $coi = 2 * \sum_{a \in A_1 \cap A_2} |1 - |\sigma_a inc_1 - \sigma_a inc_2|| / (|D_{a_1}| + |D_{a_2}|)$，其中 $|\sigma_a inc_1 - \sigma_a inc_2|$ 表示内涵集合的交集的一个属性在各自概念中的 σ 偏差值，D_{a_1} 和 D_{a_2} 分别是 c_1、c_2 的内涵。

定义 3-9 概念耦合度。$Coincidence(c_1, c_2) = \dfrac{coe * a + coi * (1 - a)}{(1 + b)^{|l_1 - l_2|}}$，其中 l_1、l_2 是 c_1、c_2 在概念格 CL_1、CL_2 中的层次；b 是为了体现概念深度对概念耦合度的影响而进行的修正。

一般来说，数据流中的概念 $<O_a; D_a>$ 是会随着时间的流逝而发生漂移的。而数据流中的概念漂移实际就是主要取决于概念外延 O_a 的变化[13]，也就是说，这个概念的含义集合是具有漂移性的。概念格中，层次越深的概念拥有更多的内涵和更少的外延，通过分析对象、属性集合在概念格中的变化能够检测漂移的发生。

定义 3-10 概念格对的概念松弛匹配耦合度 Pride。基于松弛匹配格对耦合度[9,29]，得出两个概念格两者之间的耦合程度。概念格对的概念松弛匹配耦合度算法(算法 3-1)和概念格对相似概念查询[29]的算法(算法 3-2)表示如下。

算法 3-1　概念格对的概念松弛匹配耦合度算法

输入：概念栈 S，概念格 CL_1 与 CL_2，相似概念集合 Set1，相异概念集合 Set2，相似概念子格集合 Set3，$L_{cur}=1$，Set1 = Set2 = Set3，C_{cur} = NULL，C^* = NULL，CL_{cur} = NULL。

输出：概念松弛耦合度 Pride，相似概念集合 Set1，相异概念集合 Set2，相似概念子格对集 Set3。

Step 1：**if** L_{cur}层中有概念未查询过 **then** $C_{cur}=C(O,D,\sigma,\lambda)$，$C$ 为未查询过的概念

else $L_{cur}=L_{cur}+1$

if $L_{cur}>L_{max}$ **then goto** Step 4

Step 2：C^* = SimilarConceptSearch(C_{cur}, CL_1, CL_2)

if C^* = NULL **then** Set2 = Set2 ∪ C_{cur}, **goto** Step 3

else Set1 = Set1 ∪ $\{(C_{cur}, C^*)\}$, $CL_{cur}=CL_{cur}\cup\{(C_{cur}, C^*)\}$, $C^*=NULL$

if child(C_{cur}) = NULL **then goto** Step 3

else push(C_{cur},S), C_{cur} = child(C_{cur}), **repeat** Step 2

Step 3：**while**(top(S)！= NULL)

if 有 child(top(S)) 未查询 **then** C_{cur} = child(top(S))), **goto** Step 2

else pop(S)

if CL_{cur}！= NULL **then** Set3 = Set3 ∪ CL_{cur}, CL_{cur} = NULL

goto Step 1

Step 4：令 n 为相似概念子格 Set3 的基数，每一个相似概念子格包含一个或者多个概念节点，m 表示所有相似概念子格中的相似概念的个数之和，概念格对的概念松弛匹配耦合度，得到：$Pride(CL_1,CL_2)=[|Set1|/(2*|CL_1\cap CL_2|)]*[1+m/(|Set1|+|Set2|)]^n$。

算法 3-2　相似概念查询算法(当前概念 C,概念格 CL_1,概念格 CL_2)

输入：概念格 CL_1 与 CL_2，当前概念 C

输出：相似概念 similar concepts

Step 1：以 CL_1 为基础概念格，计算概念格 CL_2 第|A|层中所有的概念与概念 C 的概念耦合度，取耦合度最大的一个概念，并记录为 C'。

Step 2：计算所有 C'的父子概念与概念 C 的耦合度，取最大的概念记录为 C^*。

Step 3：若 C^* 与 C 有 Coincidence(C, C^*) ≥ γ (γ 为概念的耦合度阈值)，则 C 在 CL_2 有相似的 C^*，查询成功；否则查询失败。

当数据流不断随时间变化，概念格不断被更新，调用算法 3-1 之后得到新的相似概念集合、相异概念集合、相似概念子格对集和概念格对的概念松弛匹配耦合度，与历史的信息作对比，即可侦测漂移的发生以及导致漂移发生的关键对象、属性。漂移特征检测算法(算法 3-3)表示如下。

算法 3-3 漂移特征检测算法

Step 1：数据流 DS 在一时间段 $<\tau_b;\ \tau_e>$ 的概念格 CL_1，$<\tau'_b;\ \tau'_e>$ 的概念格 CL_2，对两个时间段的数据流信息集合所构成的概念格进行分析，得到这两个概念格(一个概念格对)的相似概念集合 Set1、相异概念集合 Set2、相似概念子格对集 Set3 和概念格对的概念松弛匹配耦合度 Pride。

Step 2：分析确定关键概念格子块即为特征(对象属性)集合：关键概念格子块和相似概念集合 Set1、相似概念子格对集 Set3 有关，它是在多个关键概念子格对序列中且在相似概念集合序列中的概念耦合度大于耦合度阈值的概念集合。

Step 3：分析概念格对的概念松弛匹配耦合度来侦测是否有可能的概念漂移发生：在多个连续时间序列下利用概念格对的概念松弛匹配耦合度算法，即在不同时间窗口下重复 Step 1、Step 2，得到不同时间粒度的关键概念格子块和相异概念集合。

Step 4：由关键概念子格的演化和相异概念集合中各个概念的外延变化，得到漂移特征。

漂移特征检测算法的模型如图 3-3 所示。

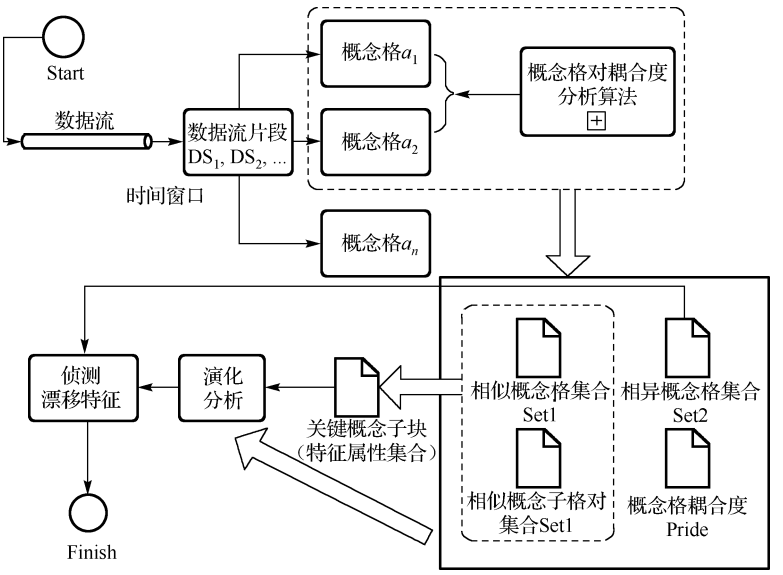

图 3-3 漂移特征检测算法的模型

数据流的概念漂移实际上是由概念外延的变化决定的，也就是概念的含义集合发生漂移[8]，在粒计算中可描述成概念所能描述对象(D_a)的范围发生了变化。在满足同一个 D_a 的不同的数据可归入到同一时间段 $<\tau_b;\ \tau_e>$ 中，亦即满足同一个内涵 D_a 的形式化描述决定一个等价类，在这一段时间中这个等价类不发生改变。而数据的更新，该等价类可能发生改变，又因为概念格的完备性，粗糙集中的每一个等价类都对应在概念格中的一个概念。可用如下两个属性特征来描述概念变化。

(1)当且仅当概念的外延 O_a 扩展时，该概念的外延 O_a 变得更强健，即外延的耦合度增大。

(2)当且仅当概念的外延 O_a 缩减时，该概念的外延 O_a 变得更衰弱，即外延的耦合度减小。

上述(1)和(2)都导致概念漂移的发生。隶属度、模糊参数 σ、模糊参数 λ、外延 O_a 变化与概念漂移都存在相关关系。根据相异概念集合 Set2 分析得出外延变化的概念集合，这些外延变化的概念就是造成概念发生漂移的关键。造成外延变化可描述为对象所包括的属性数量发生了变化，或者属性在对象中的分布发生变化。

3.3.2　基于 HSMM 的用户兴趣漂移检测模型

网上用户在浏览中的购物行为过程是受浏览目的、文化背景、兴趣爱好等多种个体因素影响的复杂过程，本节研究用户的兴趣漂移问题，但不仅仅从方法上着手，同时还考虑用户个体的背景因素，通过对背景因素、用户行为以及兴趣内容来综合考虑用户的兴趣，并建立隐半马尔可夫模型(HSMM)来检测用户兴趣是否发生漂移。

1. 用户兴趣行为与模型

在用户的网页访问浏览中，需要获得的用户兴趣是隐含的信息内容，不是通过观察就能显而易见得到的信息，而用户在网站中如点击链接、保存网页等行为是可以被观察到的，当发觉用户的行为发生了变化，我们由此可以推测用户的隐含兴趣也可能发生了变化，甚至漂移。对于用户兴趣漂移的检测，可以通过用户个体情境因素计算用户的综合兴趣度，采用一定的漂移算法检测用户是否发生了漂移。

针对用户的兴趣漂移，将建立两个 HSMM 模型，一个用于描述一个或一群用户不变兴趣的行为轮廓，另一个用于描述用户兴趣发生转移的行为轮廓。按照用户访问行为的路径序列对其进行分类，并根据用户不变兴趣行为的训练数据(历史兴趣不变行为)确定每个状态对应的观测值集合。将用户访问网页的路径序列作为漂移行为模式分类的依据，把路径序列相似的兴趣行为模式化为同一种类。

一个有 N 个状态的 HSMM[30] 是由六元组参数 $\lambda=(N, M, \pi, A, B, P)$ 组成，其中：N 表示状态的数目；M 表示观察值的数目；$\pi=\{\pi_i\}$；$A=\{a_{ij}\}$；$B=\{b_j(k)\}$；$P=\{p_i(d)\}$。第 t 个观察向量用 o_t 表示，在 o_t 中存在时间间隔 τ_t，表示第 t 个请求在对象 r_t 与 r_{t-1} 之间，即 $o_t=(r_t,\tau_t)$；o_1^T 表示整个长度为 T 的观察向量序列，o_b^a 则表示的观察向量序列是从第 a 个到第 b 个的；在 t 时刻的状态用 s_t 表示；对于当前的状态即将输出的观察值数用 ε_t 表示，$1 \leqslant t \leqslant T$。

用户在浏览网页的过程中会触发很多 HTTP 请求，如用户的点击页链接、保存页面等行为，当请求传达到服务器后，一些用户的行为信息就会被记录在服务器日志文件里面，所以想要获取用户的行为信息可以从服务器日志文件分析获取，即能够反映用户浏览行为的“轨迹”是从 log 文件中的 HTTP 请求记录中体现的。

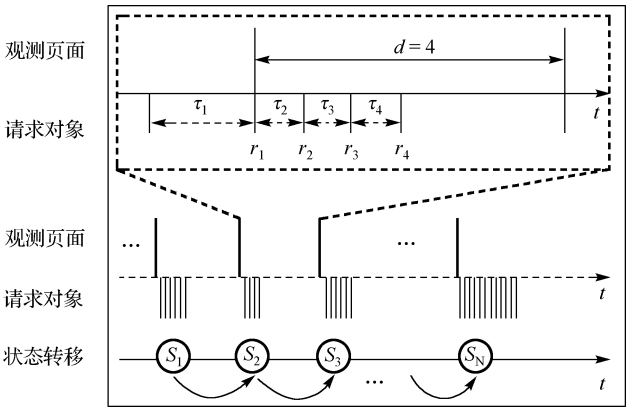

图 3-4　隐半马尔可夫模型

假设用户在浏览网页的过程中，其浏览行为符合马尔可夫模型(如图 3-4 所示)，本节选取两个观察值来描述用户的浏览行为，即用户访问网页的浏览路径序列和从一个网页到达另一个网页的时间间隔。所有状态集合表示为 $S = \{S_1, S_2, \cdots, S_N\}$，相对应的观察值集合表示为 $V = \{v_1, v_2, \cdots, v_N\}$，时间间隔表示为集合 $I = \{1, 2, \cdots\}$。对于用户的某一浏览行为，其浏览路径链接的个数是一个随机变量，在给定状态下输出的观察值的个数可将该浏览行为表示成集合 $\{1, \cdots, D\}$。把用户浏览路径序列即二维观察值序列表示成 $O = \{(r_1, \tau_1), \cdots, (r_T, \tau_T)\}$，其中，$r \in V$ 表示用户浏览网页内容的对象，$\tau_1 \in I$ 表示用户从一个页面跳转到另一个页面 r_t 与 r_{t-1} 之间的时间间隔。模型的输出概率矩阵用 $B = \{b_i(v,q)\}$ 表示，对于给定状态 $i \in S$，$b_i(v,q)$ 表示用户在一个页面 $r_t = v \in V$ 且与前一个页面的时间间隔为 $\tau_t = q \in I$ 的概率，且满足 $\sum_{v,q} b_i(v,q) = 1$。用 $P = \{p_i(d)\}$ 表示在给定状态 i 下输出观察值个数为 $d \in \{1, \cdots, D\}$ 的概率，是隐半马尔可夫模型中状态驻留时间的概率矩阵，且满足 $\sum_d p_i(d) = 1$。状态转移概率矩阵通过 $A = \{a_{ij}\}$ 进行表示，a_{ij} 表示从 $i \in S$ 向 $j \in S$ 转移的概率。初始概率向量用 $\pi = \{\pi_i\}$ 表示，π_i 表示初始状态在 $i \in S$ 时的概率。

将用户的一条重要的兴趣行为记录定义为：$U_{interest} = \{\text{user}, \text{background}, \text{history}, \text{behavior}, \text{timestamp}, \text{content}\}$。其中：user 表示用户，如 ID；background 表示用户具体背景因素；history 表示用户的历史购买记录；behavior 表示具体兴趣行为操作结果；timestamp 表示用户行为的执行时间；content 表示兴趣主题内容。

在用户访问事务中，任意两个行为操作之间存在着访问转移概率 $P(q_i \rightarrow q_j)$，得兴趣权重如下：

$$P(q_i \rightarrow q_j) = P(q_j \mid q_i) = \frac{P(q_i q_j)}{P(q_i)} = \begin{cases} \dfrac{\theta_1 W_B(q_i,q_j) + \theta_2 W_{HI}(q_i,q_j) + \theta_3 W_{IB}(q_i,q_j) + \theta_4 W_L(q_i,q_j)}{\theta_1 W_B(q_i) + \theta_2 W_{HI}(q_i) + \theta_3 W_{IB}(q_i) + \theta_4 W_L(q_i)}, & i \neq j \\ 0, & i = j \end{cases} \tag{3-13}$$

对于每个 q_j 及其相对应的概念 σ_z^j，都存在一个观察值概率分布 $P(\sigma_z^j \mid q_i)$，即 u 对 q_j 的

所有访问中，对 σ_z^j 的兴趣概率。假设 at_i 所包含被访问节点的集合为 $Q_i = \{q'_1,\cdots,q'_f \mid q' \in IC\}$，则 $Q_{i,j}$ 表示 at_i 中在 q_j 之后的所有被访问节点的集合，Q_{i,j,σ_z} 表示 $Q_{i,j}$ 中含有 σ_z^j 节点的集合：

$$Q_{i,j} = \begin{cases} \{q'_{k+1} \mid q'_k = q_j,\ l = 0,\cdots,(f-k)\}, q_j \in Q_i \\ Null \end{cases} \tag{3-14}$$

则将 u 在 q_j 上观察值概率分布 $P(\sigma_z^j \mid q_i)$ 定义为：

$$P(\sigma_z^j \mid q_j) = \frac{\text{所有访问事务在 } q_j \text{ 之后对概念 } \sigma_z^j \text{ 相关访问节点集合总数}}{\text{问事务在 } q_j \text{ 之后对所有概念访问节点集合总数}} = \frac{\sum_{i=1}^{|AT_u|} \mid Q_{i,j,\sigma_z} \mid}{\sum_{i=1,\sigma'_z \in \Sigma}^{|AT_u|} \mid Q_{i,j,\sigma'_z} \mid} \tag{3-15}$$

然后在用户 u 根据 σ_z^k 的所有可能访问序列中寻找一个状态序列，建立用户兴趣行为的隐半马尔可夫模型，使其具有最大的访问概率：

$$P_{\max}(\sigma_z^k) = \arg\max \Pi P(q_k \rightarrow q_{k+1}) P(\sigma_z^k \mid q_k) \tag{3-16}$$

2. 用户兴趣行为漂移检测步骤

在对用户兴趣漂移进行检测的过程中，首先需要采集 HSMM 模型中的观察序列，本文主要是将用户的浏览行为数据用作观察值序列，并且在模型进行训练之前对数据进行预处理，确定模型参数后，通过调用 HSMM 算法，得到用户兴趣不变的概率值，其概率值用平均对数或然概率进行计算。当用户的兴趣值处在正常范围内，则将用户数据加入到训练数据集中，以更新隐半马尔可夫模型的参数；否则，该用户将被认为是兴趣漂移。漂移检测的实现方法如图 3-5 所示。

采用 HSMM 模型分为训练和检测两个步骤。

(1)训练阶段。将观察获取到的用户兴趣行为数据进行预处理，形成观察序列 $O^l = \{(r_1^{(l)},\tau_1^{(l)})\},\cdots,(r_{T_l}^{(l)},\tau_{T_l}^{(l)})\},l = 1,\cdots,L$。然后根据 HSMM 中的学习算法，迭代计算模型的最优参数值，直至模型或然概率对数和 $\sum_{l=1}^{L} \ln(\Pr[o_1^{T_l} \mid \lambda])$ 在一定的区间收敛。

(2)实时计算各个路径序列对于给定模型的或然概率，从而对用户兴趣行为进行漂移检测。本文定义平均对数或然概率 $\log \Pr[o_1^{T_l} \mid \lambda]/T_l$ 作为第 l 个观察序列与模型符合程度的度量。具体做法：当用户的第 t 个 Web 请求到达时，记录下到达时间和请求对象，并计算出当前请求与前一请求的时间差；然后计算该用户在 $[1,t]$ 区间内对应于该 HSMM 模型的平均对数或然概率。

此处将对数或然概率的均值 lkh 作为正常行为的参考点，该均值是训练数据集内所有序列的平均值。本文中先定义一个阈值 T_0 用于表示观察序列长度，l 用户浏览过程中，当路径达到 T_0 时，然后就计算相应的平均对数或然概率 $lkh^{(l)}$。用户模型的偏离度就可以经过 $lkh^{(l)}$ 和 lkh 的比较得到。绝对值差值越小，偏离越低，兴趣漂移度越小。

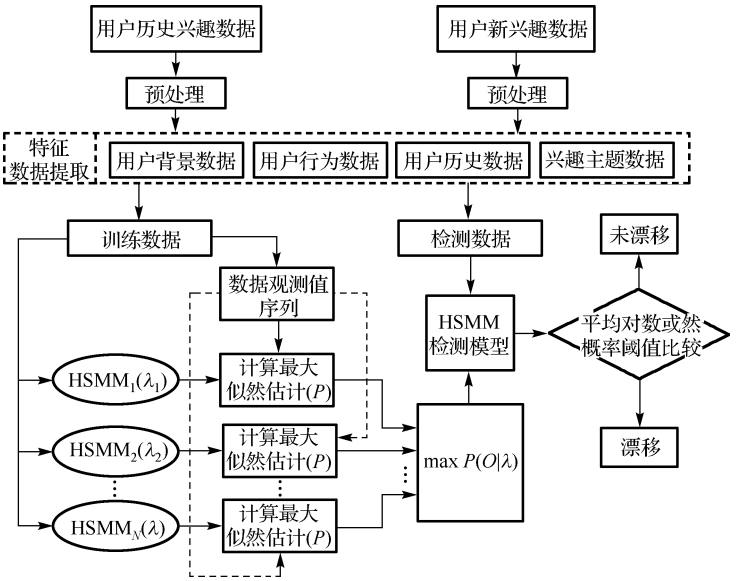

图 3-5　兴趣漂移检测流程

3.3.3　融入簇强度的数据流漂移检测模型

1. 模式信息的存储结构

由于基于聚类结果的数据流漂移检测是针对“群”进行跟踪分析，因此本文在设计时就为了便于跟踪和追溯变化情况，因此在聚类时就已经为簇进行编号，即 UCF 中的 CID。本文设计类似于邻接表的结构来存储数据流的簇信息，利用簇编号 CID 对簇进行链接。对于 CID 的编号，本文做如下规定：Rule1：两个簇合并后的簇编号以最小编号为准；Rule2：新的簇编号设置为已有 CID 的最大值加 1。

图 3-6 为从 t 到 $t+\Delta t$ 时刻的聚类信息表。图中实线为根据 CID 号而将同标号的簇链接在一起，如 UCF_1、UCF_2等为 CID 相同的簇。虚线为根据漂移检测分析后得出的演变路径，例如，虚线①代表从 t 到 $t+\Delta t$ 时刻，UCF_2与 UCF_1发生合并，合并必然代表着要删除之前的某一簇信息，所以在 $t+\Delta t$ 时刻，UCF_2被删除了；虚线②代表从 t 到 $t+\Delta t$ 时刻，UCF_4发生了分裂，分裂为 UCF_4和 UCF_{i+2}两个簇，UCF_{i+1}为新形成的簇。通过簇编号进行链接，可方便追踪跟踪、锁定变化情况。

2. 数据流漂移检测规则

数据流的漂移检测模型主要是发现和跟踪数据流中的模式发生的一系列变化，如移动、分裂、新增和消失等。不确定数据流则需要对不确定也要进行考虑描述，因此为反映簇存在强度的变化情况，本文做如下定义。

定义 3-11　在时间段 $t+\Delta t$ 内，随着新的元组不断流入，则簇 C 的变化一定是下面形式之一：

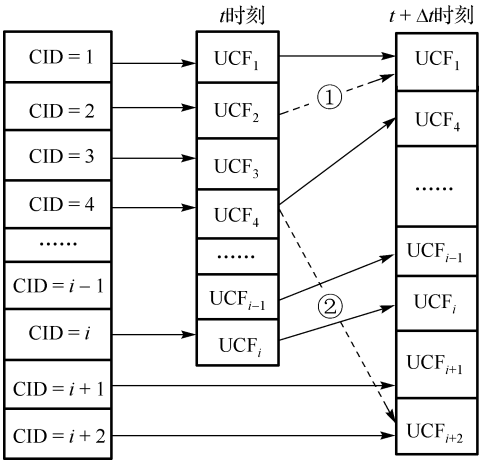

图 3-6　模式信息的存储结构

① C 在 t 时刻是弱簇，在 $t + \Delta t$ 时刻依旧是弱簇；

② C 在 t 时刻是弱簇，在 $t + \Delta t$ 时刻是过渡簇；

③ C 在 t 时刻是弱簇，在 $t + \Delta t$ 时刻是强簇；

④ C 在 t 时刻是过渡簇，在 $t + \Delta t$ 时刻是弱簇；

⑤ C 在 t 时刻是过渡簇，在 $t + \Delta t$ 时刻依旧是过渡簇；

⑥ C 在 t 时刻是过渡簇，在 $t + \Delta t$ 时刻是强簇；

⑦ C 在 t 时刻是强簇，在 $t + \Delta t$ 时刻是弱簇；

⑧ C 在 t 时刻是强簇，在 $t + \Delta t$ 时刻是过渡簇；

⑨ C 在 t 时刻是强簇，在 $t + \Delta t$ 时刻依旧是强簇。

其中，②、③、⑥称为簇 C 在时段 $[t, t + \Delta t]$ 内发生正向变化，④、⑦、⑧称为簇 C 在时段 $[t, t + \Delta t]$ 内发生负向变化，①、⑤、⑨称为簇 C 在时段 $[t, t + \Delta t]$ 内发生常变化。

定义 3-12　簇的变化率：在簇 C 接收元组数 Δn_i 中，正向变化的次数(PCN)所占的比例，称为簇的正向变化率 PCR_i (Positive change Ratio)。

$$\mathrm{PCR}_i = \frac{\mathrm{PCN}}{\Delta n_i} \tag{3-17}$$

簇的负向变化率 NCR_i(Negative changes Ratio)：在簇 C 接收元组数 Δn_i 中，负向变化的次数(NCN)所占的比例。

$$\mathrm{NCR}_i = \frac{\mathrm{NCN}}{\Delta n_i} \tag{3-18}$$

不确定数据流的漂移检测分析是在同一区内的不同聚类模型下的微簇快照集上进行的，假设 $P(C_1, C_2, \cdots, C_m)$，$P'(C'_1, C'_2, \cdots, C'_n)$ 分别为 t 时刻和 $t + \Delta t$ 时刻挖掘得到的数据流的聚类结果集合，其中 $n \geqslant m$，则检测规则如下。

(1) Rule 1 新簇产生。由于本文对簇进行编号，所以检测在 $t + \Delta t$ 时刻的新簇时，只需检查该时刻簇编号大于 t 时刻的最大簇编号的簇即为新簇。

(2) Rule 2 簇模式的变化。簇模式的变化是针对具有相同编号的簇的变化情况，变化情况具体包括以下几种类型。

① 簇中心漂移。比较簇中心的漂移距离 ΔO_i，值越大表明漂移距离较大，簇 C_i 接收的元组拉偏了中心的位置。

② 簇的膨胀。比较 Δn_i 的值，值越大说明在 Δt 时间内，微簇 C_i 接收的元组越多，在该阶段该簇比较活跃。

③ 簇半径。比较 ΔR_i 值，当 $\Delta R_i > 0$ 时，表明簇 C_i 在 Δt 时间内接收的元组距离中心点的平均距离较大，扩大了其半径。相反，当 $\Delta R_i < 0$ 时，表明簇 C_i 在 Δt 时间内接收的元组距离中心点的平均距离较小，缩小了其半径。

④ 簇强度稳定性。比较 $PCR_i + NCR_i$ 值，其值越大，表明簇 C_i 在 Δt 时间内，其簇变化率较大，簇的存在强度比较不稳定。相反，其值越小，则簇变化率较小，簇的存在强度保持得越稳定。

比较簇模式间的相似度，当具有相同编号的簇相似度低于模式匹配阈值 PMT 时，说明簇模式没发生较大变化，即 $[t, t + \Delta t]$ 时间内，簇模式较为稳定，反之则簇性质发生了概念漂移。当不同编号的簇相似度高于 PMT 时，则说明两个簇可能会发生合并。

上述的变化往往会随着数据流的到达而联动变化，即上述的四种变化会同时发生。

(3) Rule 3 簇模式的消失。比较在 t 时刻的簇编号没出现在 $t + \Delta t$ 时刻中的编号对应的簇即为消失或被合并的簇。

3. 用户行为模式漂移检测模型

用户行为变化是指用户在一定时间内，在购物时对产品种类、搜索点击、消费金额、购物频度等方面的变化，这种变化通常受客户的情境因素影响，用户的行为变化可以用规则的形式进行描述。用户的个体特征在用户聚类时已经考虑到，所以本节所指的用户情境因素是除用户个体特征的其他社会环境和时间等，具体可细分为 CPI、季节、温度、购物体验和知觉感受评价等。通过对融入用户情境的挖掘可以从多维度分析客户行为变化，识别客户购买行为的变化，为个性化推荐系统提供支持。

用户行为模式的漂移检测模型如图 3-7 所示。该模型是提出的 Web 用户分析模型中第四层的细化。用户行为模式的漂移检测模型共分为四层，首先：第一层是对融入不确定性和情境因素的用户进行关联规则挖掘得到不同时刻的行为模式；第二层计算行为模式间的相似度和相异度，根据规则匹配阈值 RMT 和期望支持度 expSup(X) 对变化程度进行度量；第三层根据变化程度进行分析，获得行为模式变化规则集合；第四层，根据用户行为变化集识别用户行为的变化情况，然后进行推荐策略的更新，为个性化推荐系统提供有力支持。

行为变化可以用规则的形式进行描述，关联规则的表述由条件属性和结果属性组成。现假设如下：

D_i^t，$D_j^{t+\Delta t}$ 分别为 t 时刻和 $t + \Delta t$ 时刻的数据集；

P_i^t，$P_j^{t+\Delta t}$ 分别为 t 时刻和 $t + \Delta t$ 时刻挖掘得到的融入情境属性的用户行为规则集合；

r_i^t，$r_j^{t+\Delta t}$ 分别为行为集合 P_i^t 和 $P_j^{t+\Delta t}$ 对应的行为规则，其中 $1 \leqslant i \leqslant |P_i^t|$，$1 \leqslant j \leqslant |P_j^{t+\Delta t}|$。

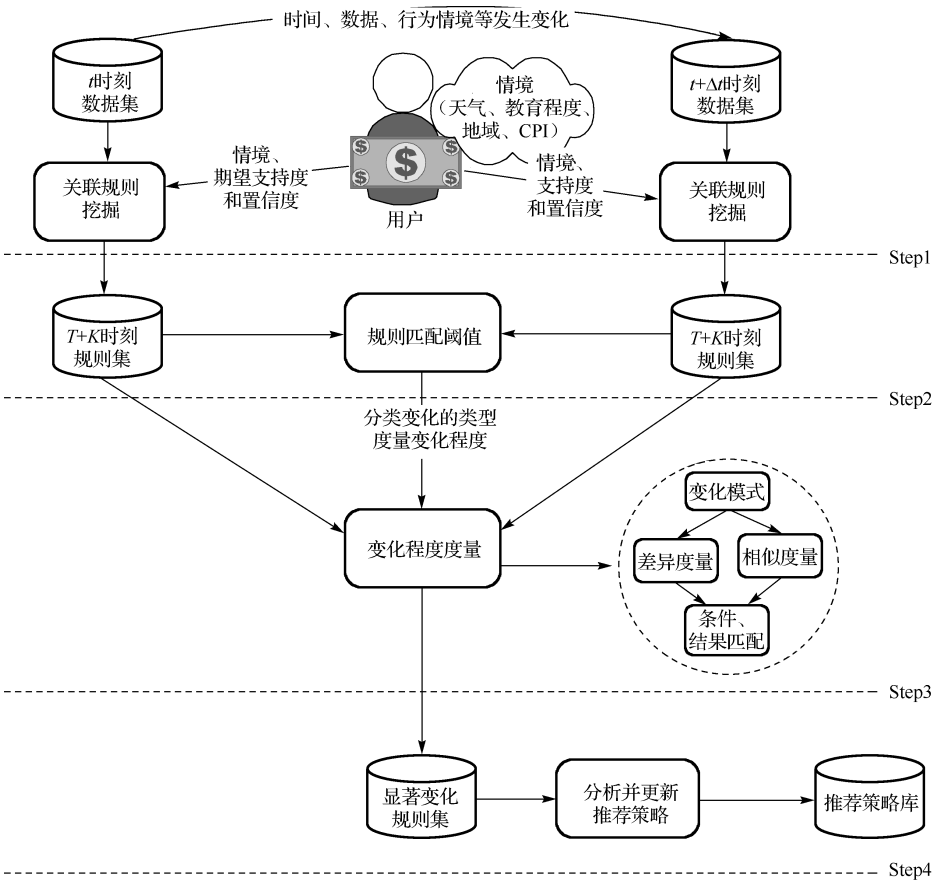

图 3-7　用户行为模式漂移检测与推荐策略更新模型

用户行为模式的变化主要是发现和跟踪行为模式的变化情况，例如，随着季节的变化，用户购买产品的组合也将会发生变化。本文主要从五个方面描述用户行为的变化。

(1)规则的稳定保持。当 r_i^t，$r_j^{t+\Delta t}$ 两个规则的条件及结果完全相同，并且规则的期望支持度没有发生较大变化时，则称在 $[t,t+\Delta t]$ 时间内，用户行为处于稳定保持阶段。

(2)模式异常变化。模式的异常变化主要是指用户的具有同样购物特征的用户行为变化。异常变化是两个模式的条件部分并没有随着时间变化而变化，但是结果部分却有着明显的变化。例如，研究生学历、女性、已婚在夏季对丝质产品感兴趣，而到了秋季，具有同样背景知识的用户群体对棉麻类产品感兴趣，这样的变化便是行为模式的条件属性没发生变化，但是其结果属性发生了变化，即其行为模式发生了变化。

(3)新生模式。新生模式是指行为规则的条件和结果都不在原有的行为模式集合中，全部都是新生的，即如果 $r_j^{t+\Delta t}$ 的条件和结论与 P_i^t 中所有的 r_i^t 的条件和结果有明显差异，则 $r_j^{t+\Delta t}$ 即为 $t+\Delta t$ 时刻的新生模式。

(4)消亡模式。原有模式的条件和结果都消失在新的模式集合中，例如，r_i^t 中的条件及结果和 $t+\Delta t$ 时刻的模式集合 $P_j^{t+\Delta t}$ 中的 $r_j^{t+\Delta t}$ 具有显著性差异，那么 r_i^t 为消亡的模式规则。由于本文将不确定性引入到用户的聚类和行为规则的挖掘中，因此不确定性也是考虑消亡模式的一个重要因素，上述的消亡是真正意义的消失，但是对于不确定数据的行为规则而