第33卷 第13期 计 算 机 工 程 2007年7月
Vol.33 No.13 Computer Engineering July 2007
·软件技术与数据库· 文章编号：1000—3428(2007)13—0056—03 文献标识码：A 中图分类号：TP 311.13

基于“层次分析法”的数据库预处理方法研究

夏骄雄，徐 俊，高 珏
(上海大学计算机工程与科学学院，上海 200072)

摘 要：层次分析法作为运筹学方法，把复杂的决策系统层次化，通过逐层比较各种关联因素组建有效模型，为分析和决策提供定量的依据。该文提出的基于“层次分析法”的数据库预处理方法在数据仓库构建的数据清理阶段，对每个准备导入数据仓库的数据库进行3种数据指标(完整性、平滑性和一致性)的评估，以“层次分析法”的准则选择合适的数据库进行数据清理，提高数据预处理的效率。

关键词：层次分析法；数据库预处理；完整性；平滑性；一致性

Database Preprocessing Approach Based on Analytic Hierarchy Process

XIA Jiaoxiong, XU Jun, GAO Jue
(School of Computer Engineering and Science, Shanghai University, Shanghai 200072)

【Abstract】 Analytic hierarchy process (AHP) is put forward suits to multiple attribute decision problem of complicated hierarchy structure. With the AHP, complicated decision system is structured, weightiness of the connection factor is compared, validity model is formed, and quantitative basis for decision and analysis is provided. The database preprocessing approach on AHP is to evaluate every database for data warehousing building. And the approach selects the appropriate database to preprocess so as to raise the efficiency of data preprocessing. The model is an application to data cleaning in data preprocessing, so three data indexes, such as completeness, smoothness and consistency, are defined differently to incomplete, noisy, and inconsistent in data cleaning. The approach regards the three data indexes as a rule, insures the valuation's stability, flexibility and contrast, and increases the effects on data cleaning.

【Key words】 analytic hierarchy process; database preprocessing; completeness; smoothness; consistency

提高数据分析的效率，围绕数据资源建立高质量的结构平台是关键。数据仓库作为一种满足结构平台需求的表现形式[1]，它的内容来自相关主题、不同来源的数据库。数据库本身的数据是“脏”的、不完整的和不一致的，这些数据在载入数据仓库时必须进行清理，并确保数据质量[2]。

文献[3,4]所论述的数据清理及其解决方法，围绕数据元对数据进行分析和转换，确保高质量数据的产生。这类数据清理模式，视每个候选数据库具有同等重要的地位，未对其各自本身存在的价值进行利用。本文借鉴运筹学中“层次分析法”的基本理念，构建独立的、针对每个需要有数据载入数据仓库的候选数据库进行评估的过程模型，选择合适的数据库进行数据清理，由此提高数据预处理的效率。

1 层次分析法

层次分析法(analytic hierarchy process，AHP)是Thomas L. Saaty针对复杂的非结构化问题提出的运筹学方法，它将半定性、半定量的问题转化为定量计算的问题，解决多目标、多准则、多时段等各种类型问题的决策分析过程，具有广泛的实用性[5]。层次分析法通过比较若干因素对于同一目标的影响，从而确定它们在目标中所占的具体权重。由于诸多因素的存在受到主观因素的影响，而且众多因素的聚集将引发更多、更复杂的不确定性问题，因此，层次分析法通过引入“成对比较法”构造判断矩阵，来提高诸多因素比较的准确度问题。

心理学界普遍认为：在同时比较若干对象时，区别差异的心理学极限为7±2个对象。因此，构造的判断矩阵A中a_{ij}的确定需要引入Thomas L. Saaty所提出的数字1~9及其倒数作为标度的具体标准(如表1所示)，以便保持事物之间差别性估计的连贯性和应用性[6]。

表1 判断矩阵标度及其含义

标度	含义
1	表示两个因素相比，具有同样的重要性
3	表示两个因素相比，一个因素比另一个因素稍微重要
5	表示两个因素相比，一个因素比另一个因素明显重要
7	表示两个因素相比，一个因素比另一个因素强烈重要
9	表示两个因素相比，一个因素比另一个因素极端重要
2,4,6,8	上述相应两相邻判断的中值
倒数	相邻两因素交换次序比较的重要性

对于构造的判断矩阵必须检验其是否具有传递性质[5]。如果不具有传递性质，则相应的特征向量**w**不能真实反映$y=(y_1,y_2,\dots,y_n)$在目标Z中所占的重要程度。

定义1 一致性指标(consistency index) n 阶的判断矩阵$\boldsymbol{A}=(a_{ij})_{n\times n}$，其不一致性的程度定义为一致性指标$CI=\frac{\lambda_1-n}{n-1}$。

定理 一致性的判断矩阵的充分必要条件 n 阶的判断矩阵$\boldsymbol{A}=(a_{ij})_{n\times n}$满足传递性质，当且仅当其一致性指标 CI 等于零。

由于$\sum_{i=1}^{n}\lambda_i=n$，一致性指标$CI$相当于$n$-1个特征根$\lambda_2,\dots,\lambda_n$(最大特征根$\lambda_1$除外)的平均值[5]。

基金项目：上海市高校科学技术发展基金资助项目(04AB29)
作者简介：夏骄雄(1973－)，男，博士、副教授，主研方向：数据挖掘，决策支持系统；徐 俊，硕士、助教；高 珏，硕士、副教授
收稿日期：2006-08-04 **E-mail：**jshardrom@staff.shu.edu.cn

定义 2 平均随机一致性指标(random index) 针对固定值 n 随机构造判断矩阵 $\boldsymbol{A}'$，其中 a_{ij}' 是 1,2,…,9,1/2,1/3,…,1/9 中随机抽取的值，同时选取充分多的样本信息得到判断矩阵 $\boldsymbol{A}'$ 最大特征根的平均值 λ_1'，则其不一致性的程度定义为平均随机一致性指标 $RI=\frac{\lambda_1'-1}{n-1}$。

对于 1~10 阶的判断矩阵，参考文献[5]给出具体的 *RI* 值(如表 2 所示)。

表 2　平均随机一致性指标 *RI*

N	1	2	3	4	5	6	7	8	9	10
RI	0	0	0.52	0.89	1.11	1.25	1.35	1.40	1.45	1.49

定义 3　一致性比率(Consistency Ratio) 对于 n 阶判断矩阵 $\boldsymbol{A}=(a_{ij})_{n\times n}$，其一致性比率 *CR* 定义为该矩阵的一致性指标与平均随机一致性指标之比，即 $CR=\frac{CI}{RI}$。

文献[7]指出，当 *CR*<0.1 时，判断矩阵具有满意的一致性比率，否则需要调整。

基于以上定义表述若干因素对于同一目标的影响在目标中所占的具体权重问题，以及诸多因素比较的准确度问题，层次分析方法的具体步骤大致分为 4 大过程：

(1)分析系统中各因素之间的关系，建立系统的递阶层次结构。

通常，递阶层次结构分为：最高层——分析问题的预定目标或者理想结果；中间层——包括为实现目标所涉及的因素、策略和准则等；最低层——为实现目标而提供选择的各种措施、决策与方案。

(2)构造两两成对比较的判断矩阵 ***A***。

判断矩阵 ***A*** 的元素值反映出外界对因素关于目标的相对重要性认识，在相邻的两个层次中，高层次为目标，低层次为因素。

(3)层次单排序及其一致性比率检验。

确定判断矩阵 ***A*** 的特征根 $Aw=\lambda_1 w$，并将 w 归一化，确定诸多因素对于目标的相对重要性排序数值，计算出 *CI* 值、*RI* 值和 *CR* 值，并根据 *CR* 的值确定层次单排序的结果是否具有满意的一致性比率。

(4)层次总排序。

层次总排序是指计算同一层次所有因素对于最高层(总目标)相对重要性的排序权值。若上一层次P包含m个因素 $P_1,P_2,\dots,P_m$，其层次总排序的权值分别为 $P_1,P_2,\dots,P_m$；下一层次Q包含n个因素 $Q_1,Q_2,\dots,Q_n$，它们对于因素P_j的层次单排序权值分别为 $q_{1j},q_{2j},\dots,q_{nj}$；当 Q_k 与P_j无联系时，取 $q_{kj}=0$。这一过程从高层到底层逐层进行。如果Q层次某些因素对于P_j单排序的一致性指标为 CI_j，相应的平均随机一致性指标为 RI_j，则Q层次总排序的一致性比率为

$$CR=\frac{\sum_{j=1}^{m}p_j CI_j}{\sum_{j=1}^{m}p_j RI_j}$$

同时根据 *CR* 的值确定层次总排序的结果是否具有满意的一致性比率。

2　基于“层次分析”的数据库预处理

针对数据清理过程处理的空缺值、噪声数据和不一致数据，结合数据库本身特点，引入 3 个定义：完整性(completeness)，平滑性(smoothness)和一致性(consistency)。

定义 4 数据库完整性 完整性是指数据库中数据属性缺少属性值、数据记录缺少记录数据的程度。它标识着数据库清理过程所需要填充空缺值的效率。

本文所定义的数据库完整性与数据库系统中的数据完整性(data integrity)并不相同。数据完整性是指数据库中数据的精确度、正确度或合法性，强调数据的精确和正确；而数据库完整性则注重属性值、记录值是否空缺，强调属性值、记录值必须非空缺，并不确保其准确。数据库完整性是对数据库中的数据对象进行整体描述的过程，具有特征描述的平面特性。

对于数据库中任意给定的数据对象，若其本身没有任何冗余的属性和记录，且为其加入任何一个属性或者记录，都将造成属性冗余、记录冗余或者所添加的属性、记录在该数据对象周围 ε 半径范围内的领域(ε-领域)之外，则称该数据对象所在的数据库具有最佳完整性(optimum completeness)。而对于数据库中任意给定的数据对象，其 ε-领域内包含所有的属性和记录，则称该数据对象所在的数据库具有最全完整性(entire completeness)。显然，最佳完整性是最全完整性的一种特殊状态，不存在任何属性冗余的情况[8]。

定义 5 数据库平滑性 平滑性是指数据库围绕特定主题的相应数据对象所含有噪声数据的程度。平滑性是数据库本身对于特定主题贡献度的集中体现，标识着数据库清理过程中围绕特定主题开展有效数据对象提取的效率和支撑程度。

对于数据库中任意给定的数据对象而言，围绕特定主题进行噪声数据的剔除时，如果没有涉及该数据对象 ε-领域和以其为中心，上下总高度为 σ 的周围空间领域(σ-层面)内的内容，则称该数据对象所在的数据库具有最佳平滑性(optimum smoothness)。

数据库平滑性也是一种对数据库中的数据对象进行整体描述的过程，但它具有特征描述的立面特性。围绕数据库清理过程中的特定主题，当对一个数据对象进行平滑操作时，噪声数据的剥离将对数据对象本身产生必然的影响。如果剔除噪声数据的操作必须涉及到数据对象 ε-领域和 σ-层面范围内的内容时，则相应的数据库将不具备最佳平滑性。

定义 6 数据库一致性 一致性是指数据库针对特定主题，所对应的数据对象之间具有的一致性程度。一致性是数据库之间对于特定主题贡献度的集中体现，标识着数据库清理过程中围绕特定主题开展多个数据库操作，进行有效数据对象提取的效率和支撑程度。

数据库一致性是一种对数据库之间相关联的数据对象进行整体描述的过程，具有特征描述的立体空间特性。针对数据库清理过程的特定主题，数据库一致性是多个数据库中数据对象信息或者已经处理的数据对象相关属性的集中反映，它意味着形成信息项的所有数据对象都是基于相同假定、定义、时期以及其他因素。通常，数据库一致性体现在由所有具有数据库完整性和平滑性的数据库集合所组成的立体空间集合上，这个空间的中心所围绕的关键内容是数据库清理的特定主题。

数据库完整性、平滑性、一致性是评估数据库的重要因素，是选择数据库进行数据库清理的基本准则。根据以上定义和“层次分析法”的基本步骤，基于“层次分析法”的数据库预处理方法包括：

(1)构建数据库预处理的递阶层次结构。

对候选数据库进行预处理，构建三层模式的递阶层次结构：目标层(Z)，选择适合清理的数据库；准则层(C)，主要运

用所定义的 3 个数据指标：*CI* 为完整性，*C*2 为平滑性，*C*3 为一致性；措施层(*P*)，用于导入数据仓库的候选数据库。

(2)构造数据库预处理的判断矩阵。

基于递阶层次结构，准则层各数据指标的权重表示各指标对于选择合适数据库的重要性。而措施层的权重表示对各个数据库完整性、平滑性和一致性的评估。确定各因素的权重，通过两两比较建立判断矩阵。判断矩阵的数值根据数据资料、专家意见和分析者的认识，加以平衡后给出[5]。

(3)确定判断矩阵的最大特征根及特征向量。

层次分析法计算的根本问题是计算判断矩阵的最大特征根及其对应的特征向量。精确算法(幂法)是两种最主要的求解方法之一，包括：

1) 初始化。设 $k=0$，任取初始正向量 $\boldsymbol{x}^{(0)}=(x_1^{(0)},x_2^{(0)},\ldots,x_n^{(0)})^{\mathrm{T}}$，计算 $m_0=\left\|x^{(0)}\right\|_\infty=\max_i\{x_i^{(0)}\}$ 和 $y^{(0)}=\dfrac{x^{(0)}}{m_0}$。

2)累计归纳计算。对于 $k=0,1,2,\ldots$，计算 $x^{(k+1)}=Ay^{(k)}$、$m_{k+1}=\left\|x^{(k+1)}\right\|_\infty$ 和 $y^{(k+1)}=\dfrac{x^{(k+1)}}{m_{k+1}}$。

3)判断条件。判断具体条件，当 $|m_{k+1}-m_k|<\varepsilon$ 时，转至第(4)步；否则令 $k=k+1$，转至第(2)步。

4)结论计算。计算特征根 $\lambda_1=m_{k+1}$，其对应的特征向量 $\boldsymbol{u}=\dfrac{y^{(k+1)}}{\sum_{i=1}^{n}y_i^{k+1}}$。

(4)层次总排序的一致性比率检验与计算组合权重。

从高层到底层逐层进行层次总排序的权值求解，计算出每一个层次的 *CI* 值、*RI* 值和 *CR* 值，从而获得每一层次对于最高层(总目标)的相对重要性排序权值，确定层次总排序的一致性比率，并以此计算候选数据库的各自权重。

3 应用示例

利用基于“层次分析法”的数据库预处理方法，通过综合分析上海大学计算机工程和科学学院2001级学生3个学年(2001~2002 学年、2002~2003 学年、2003~2004 学年)的基本信息数据库，围绕“培养优秀本科毕业生攻读硕士学位”这一用户主题信息，分别对这 3 个数据库进行数据库预处理，确定其各自对于特定用户主题要求所提供的贡献度，分析与归纳出相应的基本特点。同时，对比具有最优先选择权的数据库与最后构建的数据仓库之间针对同样用户主题的处理结果，表明其在预测领域的覆盖特性。

根据基于“层次分析法”的数据库预处理方法的描述，利用具体目标、准则和措施，构建如图 1 所示的数据库选择递阶层次结构。

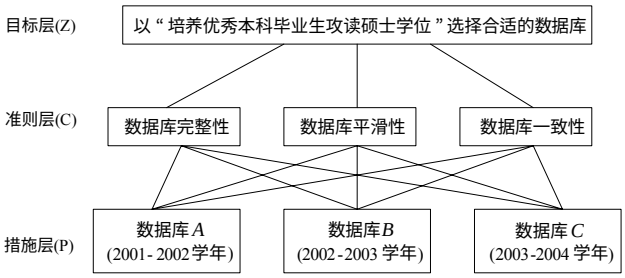

图 1　数据库选择的递阶层次结构

根据学生基本信息数据库的特性，确定数据库完整性的数据对象 ε-领域定义为聚合 10 个数据对象的最小范围，明确数据库平滑性的数据对象 σ-层面定义为学业成绩数据对象覆盖区域，各个层次指标的权重如表 3 所示。通常，同一层次各因素针对上一层次特定因素的权重之和为 1。

表 3　各层次指标的权重

项目	完整性(C1)	平滑性(C2)	一致性(C3)
准则层指标的权重(C)	0.18	0.09	0.73
措施层指标的权重(P)	-	-	-
数据库 A(P1)	0.595	0.128	0.630
数据库 B(P2)	0.128	0.595	0.219
数据库 C(P3)	0.277	0.277	0.151

基于各个层次的权重，构造两两成对比较的判断矩阵，然后进行层次单排序以及一致性比率检验。在本例中，判断矩阵的数值由两个因素对目标的重要性(即权重)之比，所得结果为四舍五入取整得到。

同时，计算判断矩阵的最大特征根及其对应的特征向量基于“层次分析法”的数据库聚类预处理方法描述中的精确算法(幂法)实施的，具体过程如表 4 所示。

表 4　幂法求解判断矩阵的最大特征根及其对应的特征向量

Z-C 判断矩阵	***Z***	***C1***	***C2***	***C3***	***W***
	C1	1	2	1/4	0.18
	C2	1/2	1	1/8	0.09
	C3	4	8	1	0.73
结论	λ_1=3，CI=0，RI=0.58，CR=0<0.1				
***C1-P* 判断矩阵**	C1	***P1***	***P2***	***P3***	***W***
	P1	1	5	2	0.595
	P2	1/5	1	1/2	0.128
	P3	1/2	2	1	0.277
结论	λ_1=3.005535，CI=0.002768，RI=0.58，CR=0.004772<0.1				
***C2-P* 判断矩阵**	***C2***	***P1***	***P2***	***P3***	***W***
	P1	1	1/5	1/2	0.128
	P2	5	1	2	0.595
	P3	2	1/2	1	0.277
结论	λ_1=3.005535，CI=0.002768，RI=0.58，CR=0.004772<0.1				
***C3-P* 判断矩阵**	***C3***	P1	P2	P3	W
	P1	1	3	4	0.630
	P2	1/3	1	1	0.219
	P3	1/4	1	1	0.151
结论	λ_1=3.009203，CI=0.004601，RI=0.58，CR=0.007933<0.1				

最后进行层次总排序的一致性比率检验和计算组合权重：$CR=\dfrac{\sum_{j=1}^{3}a_jCI_j}{\sum_{j=1}^{3}a_jRI_j}$=0.007080<0.1，组合权重满足一致性比率检验的要求，从而获得各个数据库的具体权重：

$$\begin{bmatrix}0.595 & 0.128 & 0.63\\ 0.128 & 0.595 & 0.219\\ 0.277 & 0.277 & 0.151\end{bmatrix}\begin{bmatrix}0.18\\ 0.09\\ 0.73\end{bmatrix}=\begin{bmatrix}0.578520\\ 0.236460\\ 0.185020\end{bmatrix}$$

分析计算结果，数据库 A 所占的权重为 57.8520%，远远高于数据库 B 所占的 23.6460%和数据库 C 所占的 18.5020%，说明在选择数据库进行基于“培养优秀本科毕业生攻读硕士学位”的数据仓库构建过程中，数据库 *A* 应该被优先选择。

通过对数据库 *A* 直接进行用户主题信息挖掘，以及对数据库 *A*、数据库 *B*、数据库 *C* 构建的数据仓库 $A-B-C$ 进行用户主题信息挖掘，挖掘结果显示：基于“层次分析法”的数据库预处理方法直接对数据库 *A* 进行挖掘的目标准确率达到 50%，时间复杂度为 $O(|A|)$；通过对数据仓库 $A-B-C$ 进行挖掘的目标准确率亦在 50%，而时间复杂度达到 $O(|A|\times|B|\times|C|)$。

（下转第 61 页）