doi：10.11830/ISSN.1000-5013.201606019

海量冗余数据干扰下数据库中数据优化检索方法

王晓英

（赤峰学院 数学与统计学院，内蒙古 赤峰 024000）

摘要： 针对传统方法对数据库中的数据进行检索的过程中，在海量冗余数据干扰时存在无法区分检索数据类别，降低数据检索的效率和精度的问题，提出一种基于特征模糊接近的海量冗余数据干扰下数据库中数据优化检索方法. 利用数据模糊集间的接近度表述海量冗余数据干扰下数据库中数据的一致度，结合数据融合技术，对类间数据实现分类处理. 利用模糊集算法准确查询分类数据，对分类数据实现二次聚类计算，细分其类边缘，通过加载辨别函数实现数据定位，完成数据检索. 实验结果表明：该方法进行数据检索时具有较高的检索效率和精度，且抗干扰能力较强.

关键词： 数据检索；冗余数据；特征模糊；模糊集算法；抗干扰

中图分类号： TP 311.5 **文献标志码**： A **文章编号**： 1000-5013(2016)06-0758-04

Optimization Method of Retrieving Data in the Database Under the Interference of Lage Redundant Data

WANG Xiaoying

(Institute of Mathematics and Statistics, Chifeng University, Chifeng 024000, China)

Abstract: In the process of using traditional method to retrieve data in the database, the interference of large redundant data is unable to distinguish when retrieving data category, which reduces the efficiency and accuracy of data retrieval. The paper puts forward an optimization method of retrieving data in the database under the interference of large redundant data based on the characteristics of fussy approaching mass. The method is to use the proximity in the fussy data regions to show the consistency of data in the database under the interference of large redundant data, combine the data fusion technology to classify the indirect data, use fussy set algorithm to query classified data accurately to realise secondary clustering calculation of classified data and segment the edge of class, position the data and complete the data retrieval by loading identification function. The experimental results show that the method for data retrieval has higher retrieval efficiency and accuracy, and strong anti-interference capability.

Keywords: data retrieval; redundant data; fuzzy feature; fuzzy set algorithm; anti-interference

在不同类型网络数据库的数据检索过程中，由于数据库信息资源的存储资源具有多源属性，对数据库进行信息检索过程中会产生海量干扰数据. 如何在海量数据的干扰下对数据库中的数据进行有效检索，提高数据库数据检索精度，是该领域亟待解决的问题，具有重要的应用价值[1-3]. 在传统的数据库中，

收稿日期： 2016-10-13

通信作者： 王晓英(1979-)，女，副教授，主要从事应用数学的研究. E-mail：527514533@qq.com.

基金项目： 国家自然科学基金资助项目(11402039)

数据优化检索方法有：基于虚拟数据加速分布重组的数据库索引技术[4]、多源数据相位谱补偿的数据库索引算法[5]、弱关联字符型数据的密文检索模型优化方法[6]. 然而，传统方法进行海量冗余数据干扰下数据库中数据检索时，存在无法区分检索数据类别，降低数据检索的效率和精度的问题. 本文提出一种基于特征模糊接近的海量冗余数据干扰下数据库中数据优化检索方法[7-8].

1　数据优化检索方法

1.1　模糊接近分类技术

模糊接近分类技术将存在类间集数据之间的关系进行连接、归类，实现数据检索[9-10]. 数据间的模糊接近分类具体实现过程如下.

计算数据集间特征的偏斜度，假设在 t 时域内，将第 i 个类间检索的数据用 $x_i(t)$，$i=1,2,\cdots,n$ 表示. 如果 $x_i(t)$，$x_j(t)$间差异性较大，则表明不同分类获取的数据一致性较低，偏斜度较大. 高度一致的数据可保障数据检索模型拥有较高的精准度，利用数据模糊集间近似度代表数据间的一致性，有

$$s_{i,j}(t) = \exp(-\lambda(x_i(t) - x_j(t)))^2. \tag{1}$$

将 t 时域偏斜度置信矩阵表示为

$$\boldsymbol{S}(t) = \begin{bmatrix} 1 & s_{1,2}(t) & \cdots & s_{1,n}(t) \\ s_{2,1}(t) & 1 & \cdots & s_{2,n}(t) \\ \vdots & \vdots & \vdots & \vdots \\ s_{n,1}(t) & s_{n,2}(t) & \cdots & 1 \end{bmatrix}. \tag{2}$$

由于数据集间特征偏斜度的置信矩阵拥有空间及时间两个维度的数据置信性，在 t 时域内检索相同数据的偏斜度表示为

$$p_i(t) = \frac{1}{n}\sum_{i=1}^{n} s_{i,j}(t). \tag{3}$$

利用式(3)的偏斜度对数据进行模糊分类，可得到最高一致性的置信数据. 为提高这种类间偏斜度的分类性能，利用反向传输(BP)神经网络对偏斜度计算的方法进行优化. 为保障神经网络实际输出的偏斜度与期望输出的偏斜度误差及均方差均为最小，在神经网络中代入最小二乘法，保障偏斜度运算的精准度[11-12].

设定海量数据干扰下数据库中的类间数据样本为：(X_k, Y_k)，其中，$k=1,2,\cdots,m$，输入的样本为 X_k，且有 $X_k^T=(x_{1,k}, x_{2,k}, \cdots, x_{n,k})$，表述数据样本的维度，描述神经网络数据模型，如图 1 所示.

若数据库神经的隐层数为 L，代入偏斜度计算，则第 i 个输出为

$$\hat{y} = \sigma_0(\bar{y}_{i,k}), \tag{4}$$

$$\bar{y}_{i,k} = W_{i,k}^{(0)T}\hat{H}_k^{(l)} = \sum_{j=1}^{m,l}\omega_{i,j}^{(0)}h_{j,k}^{(l)}. \tag{5}$$

利用式(5)的偏斜度基准对偏斜度权值矩阵第 p 行实现微分转换，有

$$\frac{\partial \boldsymbol{E}_k}{\partial W_{p,k}^{(0)}} = \frac{\partial \boldsymbol{E}_k}{\partial e_{p,k}}\frac{\partial e_{p,k}}{\partial \hat{y}_{p,k}}\frac{\partial \hat{y}_{p,k}}{\partial y_{p,k}}\frac{\partial \bar{y}_{p,k}}{\partial \hat{y}_{p,k}^{(0)}} = -e_{p,k}\sigma'_0(\bar{y}_{p,k})\hat{H}_{p,k}^{(1)}, \tag{6}$$

$$\Delta W_{p,k}^{(0)} = W_{p,k}^{(0)} - W_{p,k-1}^{(0)} = -\alpha\frac{\partial \boldsymbol{E}_k}{\partial W_{p,k}^{(0)}}. \tag{7}$$

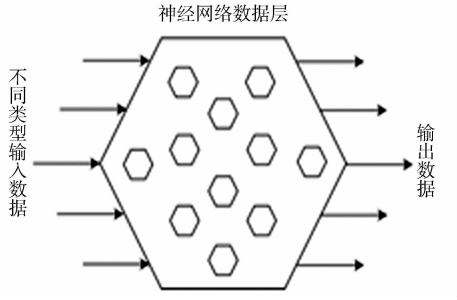

图 1　神经网络数据模型

Fig. 1　Neural network data model

对 BP 网络以输出层为起始方向，向输入层实现反向递推，对第 r 层的偏斜度权值进行修正，提高偏斜度计算的精度，实现模糊接近分类的优化，有

$$\Delta W_{p,k}^{(r)} = W_{p,k}^{(r)} - W_{p,k-1}^{(r)} = \alpha\varepsilon_{p,k}^{(r)}\hat{H}_{p,k}^{(r+1)}, \tag{8}$$

$$\varepsilon_{p,k}^{(r)} = \sigma'_r(\bar{h}_{p,k}^{(r)})^{n-1}\sum_{i=1}^{r}\varepsilon_{i,k}^{(r-1)}\omega_{i,p}^{(r-1)}. \tag{9}$$

数据模糊分类技术可避免数据间的近似性干扰，为高效、准确地进行数据检索奠定基础.

1.2 海量数据干扰下数据库中数据优化检索

引入三角模糊集算法:将设定论域 F 内的某一个模糊集表示为对任意 $x\in q$,均有一个数 $\mu(x)\in[0,1]$与之相互对应.将 x 对 q 的隶属度表示为 $\mu(x)$,μ 为隶属函数,设定 q 为模糊数目的上限,s 为模糊数的下限,可能性最大的值为 m,$T=\{t_1,\cdots,t_n\}$代表检索目标数据的组合,组合序号为 j 的记录为 t_j,$I=\{i_1,\cdots,i_{m+1}\}$代表数据集 T 的特征集,其中,数特征为 $i_1,\cdots,i_m$,类特征为 i_{m+1},利用模糊 C 算法使 $i_1,\cdots,i_m$ 划分为不同的三角模糊集,详细实现过程如下.

1) 假设循环次数表示为 s,建立 $\boldsymbol{F}^{(0)}\in\boldsymbol{M}_{f,c}$初始化矩阵,即

$$\boldsymbol{M}_c=\{F\in\mathbf{R}^n\mid 0\leqslant u_{i,j}\leqslant 1,\ u_{1,j}+u_{2,j}+\cdots+u_{c,j}=1\}. \tag{10}$$

2) 计算初始化矩阵中的向量 $\boldsymbol{v}_i$,有

$$\boldsymbol{v}_i=\sum_{j=1}^{n}(u_{i,j})^m x_j/\sum_{j=1}^{n}(u_{i,j})^m. \tag{11}$$

3) 执行 $S+1$ 次循环对初始化矩阵进行更新,对于任意 $\boldsymbol{v}_i$,当满足 $1\leqslant v_i\leqslant n$,且 $v_i=d(x_i,v_i)$大于零时,转至步骤 4);

4) 设置结束参数 λ,满足 $\|\boldsymbol{F}^{(s+1)}-\boldsymbol{F}^{(s)}\|\leqslant\lambda$ 时,停止分割循环;否则,返回步骤 2),继续循环.

在循环过程中,对 F 实现反复计算,此时,$\boldsymbol{v}_i$ 已被分类成 c 类,与模糊集隶属度函数 f 相融合,设定因子满足 $A=s-\dfrac{f(s)(v-s)}{1-f(s)}$,$B=q+\dfrac{f(q)(q-s)}{1-f(q)}$时,则有

$$f(x)=\begin{cases}\dfrac{x-A}{v-A}, & a\leqslant x\leqslant v,\\[2ex] \dfrac{B-x}{B-v}, & v\leqslant x\leqslant b.\end{cases} \tag{12}$$

通过以上步骤可以将 $i_1,\cdots,i_m$ 分解成 $l_1,\cdots,l_m$ 个模糊集,完成分类数据实现二次聚类计算,细分其类边缘.对于模糊数据属性集 H,设惩罚参数,当出现海量数据干扰时,利用该惩罚参数对海量数据进行消除,有

$$\sup(H)=\sum_{j=1}^{n}\prod_{n=1}^{p}l_j/n. \tag{13}$$

针对划分为不同类型 $l_1,\cdots,l_m$ 个数据模糊集,建立数据辨别函数,通过加载辨别函数对待检测数据定位进行定位,完成数据库中的数据检索优化,有

$$I(x,y)=-\sum_{x\in X}P(x,y)\times\lg\frac{p_{x,y}(x,y)}{(x,y)}. \tag{14}$$

2 实验结果与分析

为证明文中提出基于特征模糊接近算法的海量数据干扰下数据优化检索方法的有效性,对文中算法与传统算法进行对比实验.实验平台为 Windows 7 操作系统.

实验 1 采用文中算法对数据库中的不同数据集进行检索实验.文中算法的测试结果,如表 1 所示.表 1 中:ST 为检索时间;RA 为检索精准度;RE 为检索误差率实验结果与目标数据的距离平方根.由表 1 可知:文中算法在处理海量数据干扰下数据库的子类数据集检索时间较短,且检索精准度较高,具有可执行性.

不同算法数据检索结果与目标距离,如表 2 所示.由表 2 可知:文中算法聚类检索整体优越性远高于其他两种算法.这主要因为文中方法先利用数模糊集间的接近度表述海量冗余数据干扰下数据库中数据的一致度,利用模糊集算法准确查询目标数据,对目标数据实现二次聚类计算,细化其类边缘,保障

表 1 文中算法的测试结果

Tab.1 Test results of algorithm in the paper

数据集	ST/s	RA/%	RE/%
A1	9.521	98.2	1.8
A2	10.251	97.9	2.1
A3	10.296	96.7	3.3
A4	8.964	97.9	2.1

表 2 算法数据检索结果与目标距离

Tab.2 Algorithms data retrieval result and target distance

数据量/T	传统方法	文献[8]方法	文中方法
1	90.25	69.54	16.54
2	198.20	201.30	42.39
3	412.30	506.30	163.20
4	598.20	649.40	298.20

了数据优化检索的高效、精准性.

实验 2　多次实验求取平均值，传统算法与文中算法搜索最优查询方法的搜索代价消耗比变化，如图 2 所示. 执行最优查询方案查询代价消耗比变化，如图 3 所示. 图 3 中：n 为连接数.

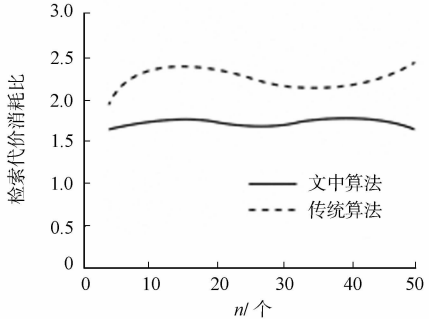

图 2　不同算法的检索代价消耗比

Fig. 2　Retrieval cost consumption ratio of different algorithms

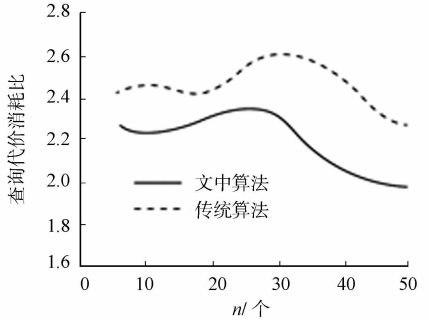

图 3　不同算法的数据查询代价消耗比

Fig. 3　Data query cost consumption ratio of different algorithms

由图 2，3 可知：文中提出的基于特征模糊接近的海量冗余数据干扰下数据库数据优化检索方案算法，降低了数据库查询执行的时间与代价. 这主要因为文中算法将数据模糊集间的接近度表述数据的一致度，利用模糊集算法准确查询目标数据，通过加载辨别函数实现数据定位，实现海量冗余数据干扰下数据的最优检索.

3　结束语

针对传统方法检索效率低且精度差等缺点，基于特征模糊接近提出海量冗余数据干扰下数据优化检索方法. 首先，运用数据模糊集间的接近度描述数据的一致度，结合数据融合技术，实现对类间数据的分类处理. 其次，利用模糊集算法准确查询分类数据，对其进行二次聚类计算，细分其类边缘，加载辨别函数以定位数据，实现数据优化检索. 结果表明：用文中方法检索数据具有较高的检索精度和效率，且具有较强的抗干扰能力.

参考文献：

[1]　祝钢. 数据库中密文检索优化模型仿真与研究[J]. 计算机仿真，2014，31(11)：336-339.

[2]　刘兴明. 采用频域波束分级聚焦的多源数据库幂级检索[J]. 科技通报，2015，31(10)：202-204.

[3]　冯祥斌，陈永红. 应用 P-Fibonacci 加密的模糊自适应水印算法[J]. 华侨大学学报(自然科学版)，2014，35(3)：287-292.

[4]　潘晓萌，王维哲. 基于虚拟数据加速分布重组的数据库索引技术[J]. 科技通报，2015，31(8)：135-137.

[5]　王小琼，王艳淑. 引入多源数据相位谱补偿的数据库索引算法[J]. 科技通报，2015，31(12)：173-175.

[6]　王小英，白灵，孙晓玲，等. 弱关联字符型数据的密文检索模型优化仿真[J]. 计算机仿真，2014，31(2)：432-435.

[7]　刘静. 数据挖掘技术在教务管理实践中的应用研究[J]. 电子设计工程，2014，22(24)：1-3.

[8]　徐新爱. 无人机海量飞行数据快速检索方法研究[J]. 计算机测量与控制，2014，22(12)：4181-4183.

[9]　王艳，刘继华. 基于多维索引树编码的数据库分层访问技术研究[J]. 软件导刊，2016，15(5)：173-175.

[10]　孙皓. 基于神经网络的上海光源光束故障预警的方法研究[D]. 上海：中国科学院研究生院(上海应用物理研究所)，2016：15-20.

[11]　彭良睿，李学明. 一种基于树型结构的 P2P 系统高维数据检索方法[J]. 计算机应用研究，2015，32(3)：842-845.

[12]　张兴忠，王运生，曾智，等. 一种高效过滤提纯音频大数据检索方法[J]. 计算机研究与发展，2015，52(9)：2025-2032.

(责任编辑：钱筠　　英文审校：吴逢铁)