

基于智能语义自相关特征的数据库实体识别查询

王海荣

(黄淮学院 信息工程学院, 河南 驻马店 463000)

摘 要: 为有效解决 Deep Web 数据库中数据纠错、消重和整合问题, 优化 Deep Web 数据库实体查询性能, 提出一种基于智能语义自相关特征的 Deep Web 数据库优化识别查询模型. 模型由文本匹配模型、语义自相关特征分析模型和分组统计模型构成, 设计语义自相关特征提取算法, 定义语义信息约束规则, 优化逼近数据库实体识别结果, 使用语义自相关特征建立表象关联知识库, 实现数据查询过程中的纠错、消重、整合. 最后用数学推导证明了算法的稳定性. 仿真模拟实验表明, 算法能综合考虑文本特征、语义自相关特征和约束规则, 数据库查询识别结果不断精化, 数据查询准确性和有效完备性提高显著.

关键词: 语义; 数据库; 自相关特征; 查询

中图分类号: TP393

文献标识码: A

文章编号: 1000-7180(2014)05-0159-04

Database Entity Recognition Intelligent Semantic Query Based on Autocorrelation Characteristics

WANG Hai-rong

(School of Information Engineering, Huanghuai University, Zhumadian 463000, China)

Abstract: In order to effectively solve the data correction, elimination and integration problems for Deep Web database, make the Deep Web database entity query performance with optimization, the database entity recognition intelligent semantic query method based on autocorrelation characteristics was proposed. The query model consisted of text matching model, semantic correlation characteristic analysis model and the statistical model, and semantic correlation feature extraction algorithm was proposed. The semantic information constraint rule was defined for optimal approximation database entity recognition results. The semantic correlation characteristics were used to establish the surface correlation knowledge base, the correction, elimination, integration was realized in the data query. The mathematical derivation proves that the stability of the algorithm, and simulation results show that algorithm can consider, text feature, semantic correlation feature and constraint rules, the accuracy and completeness of data query increased significantly and effectively.

Key words: semantic; database; autocorrelation; query

1 引言

Deep Web 数据库存储信息规模庞大, 动态更新迅速, 查询接口面向对象化和跨平台分布式查询设计, 因此, Deep Web 数据库的数据纠错、消重和整合问题突出^[1], 需要研究一种合理高效的数据库查询算法, 提高 Web 数据库的在线访问和查询效

率, 降低查询误差. 在对 Deep Web 数据库优化查询访问设计中, 文献[2]中提出一种基于网页上下文的 Deep Web 数据库分类查询方法. 通过利用 HTML 网页的内容文本特征对数据库知识进行分类标记, 实现 KNN 分类查询, 但网页上下文文本特征对 Deep Webs 数据库的深度信息不能合理表征, 导致分类精度和查全率不高. 文献[3]中采用属性相关度

收稿日期: 2013-07-19; 修回日期: 2013-09-03

基金项目: 河南省科技厅发展计划(132102210463)

估计的方法对 Web 数据库大小进行估计,方法能准确估算 Web 数据库大小,采用词频统计方法对数据库进行查询访问.文献[4]中提出一种 Web 数据的关键技术新格局,从应用维度入手,采用 noSQL 技术,实现大型关系数据库的准确查询.但传统方法都没能有效解决 Web 数据库中数据纠错、消重和整合等问题,导致 Deep Web 数据库实体查询性能不理想^[5].

针对上述问题,本文利用 Deep Web 数据库实体识别技术,对数据集中产生的重复记录进行重整检测,提出一种基于智能语义自相关特征的 Deep Web 数据库优化识别查询模型^[6].

2 Deep Web 数据库语义自相关特征

通常进行 Deep Web 数据库访问,数据纠错、消重和整合等问题的解决是实现优化查询的关键,基于语义统计分析的自相关特征能有效反映 Deep Web 数据库深度特征信息.故本文将采用智能语义自相关特征,构建语义自相关特征建立表象关联知识库,实现数据查询过程中的纠错、消重、整合.语义自相关特征提取算法描述如下:

采用有向图形式表示语义信息,首先寻找两个有向图之间的交集.假设 G_c 是数据库两个语义特征信息转换而成的有向图 G_1 和 G_2 的交集,那么说明 G_c 中的语义结点同时存在于 G_1 和 G_2 在上述语义特征目标函数提出的基础上,以目标函数为目标收敛终止条件,采用语义自相关函数成分分析的方法,提取语义自相关特征,首先使用自相关成分独立随机变量分析的相关知识,对于两个标量时间序列 y_1 和 y_2 ,其联合概率密函数为 $f(y_1, y_2)$,时间序列具有自相关特征,采用积分变量单独成分分析的思路,得到数据库语义特征概率分布:

$$f_1(y_1) = \int f(y_1, y_2) dy_2 \quad (1)$$

两个向量特征在相互邻域内,期望均值具有独立性,故其包含的信息成分对盲源节点具有相对动态协调分布能力,可实现数据信息盲分离,从而有效对信号进行自相关成分的独立快速分析与分离,这时: $f(y_1, y_2) = f_1(y_1) f_2(y_2)$ 可以扩展到 n 维变量中,可以对多用户信息进行有效检测,基于自相关函数的独立的变量在盲分离时主成分具有频移特性,对于连续变量 $x(t)$,其自相关函数 $C(\tau)$ 定义为

$$C(\tau) = \lim_{T \rightarrow \infty} \frac{1}{T} \int_{-\frac{T}{2}}^{\frac{T}{2}} x(t) x(t + \tau) d\tau \quad (2)$$

Deep Web 数据库信息流序列 $\{x_i\}$ 时间跨度为 $j\tau$ 的自相关函数为

$$C_{xx}(j\tau) = \frac{1}{N} \sum_{i=0}^{N-1} x_i x_{i+j\tau} \quad (3)$$

得到信号特征变量的期望值相互独立,有

$$E\{h_1(y_1) h_2(y_2)\} = E\{h_1(y_1)\} E\{h_2(y_2)\} \quad (4)$$

此时有

$$\begin{aligned} & E\{h_1(y_1) h_2(y_2)\} \\ &= \iint h_1(y_1) h_2(y_2) f(y_1, y_2) dy_1 dy_2 \\ &= \int \int h_1(y_1) f_1(y_1) h_2(y_2) f_2(y_2) dy_1 dy_2 \\ &= \int h_2(y_2) f_2(y_2) dy_2 \int h_1(y_1) f_1(y_1) dy_1 \\ &= E\{h_1(y_1)\} E\{h_2(y_2)\} \end{aligned} \quad (5)$$

定义 $E\{[X - E(X)][Y - E(Y)]\}$ 为随机变量 X 与 Y 的协方差,记为 $\text{Cov}(X, Y)$,此时假设随机变量 X 与 Y 分别是数据库信息流变量和提取的语义特征分量,其自相关协方差为

$$\text{Cov}(X, Y) = E\{[X - E(X)][Y - E(Y)]\} \quad (6)$$

此时,得到两组信息流的自相关系数,表示为

$$\rho_{XY} = \frac{\text{Cov}(X, Y)}{\sqrt{D(X)} \sqrt{D(Y)}} \quad (7)$$

式中, ρ_{xy} 是一个无量纲的量.在上述证明和分析的基础上,需要对各个 Deep Web 源数据进行语义关联,这里为各个用户终端采集的信息流信号,设定先验概率知识,根据干扰项的高斯密度函数,利用数理统计的中心极限定理,对多随机信号进行语义自相关特征信息提取.通过上述算法描述提取语义自相关特征,为构建数据库实体查询模型奠定算法基础.

3 Deep Web 数据库优化识别查询模型

3.1 模型设计与算法实现

在采用语义自相关特征基础上,设计 Deep Web 数据库优化识别查询模型,模型由文本匹配模型、语义自相关特征分析模型和分组统计模型构成,通过语义信息约束规则,优化逼近数据库实体识别结果,以适应 Deep Web 数据库动态关联完备性要求,Deep Web 数据库优化识别查询模型如图 1 所示.

通过数据库表象关联知识指导数据资源的选择、消重和数据整合.现在用数学语言来描述自相关成分分析问题,数据库实体集为 $\mathbf{x} = (x_1, x_2, \dots, x_m)^T$,对 n 个未知的零均值自相关随机

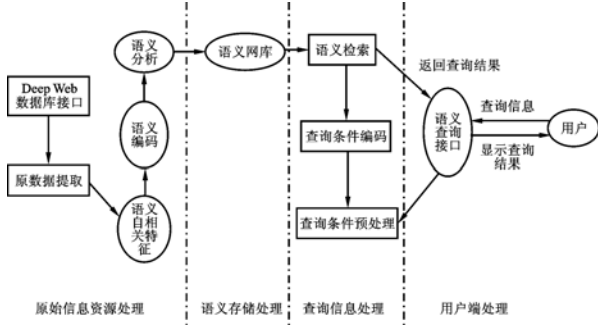


图1 Deep Web 数据库优化识别查询模型

信号源 $s = (s_1, s_2, \dots, s_n)^T$ 线性混合组成的, 这种线性混合, 也就是代表了同一实体的表象集, 形成自相关特征与静态分析文本相似集的线性混合, 定义语义信息约束规则, 优化逼近数据库实体识别结果, 在语义自相关成分分析基础上, 得到语义信息约束规则为:

$$x = Hs = \sum_{j=1}^n h_j s_j, j = 1, 2, \dots, n \quad (8)$$

式中, $H = [h_1, h_2, \dots, h_n]$ 为 $m \times n$ 阶满秩源信号混合矩阵; 系统文本匹配模型输出模式可以改写为矩阵形式为

$$\begin{bmatrix} x_1(t) \\ \vdots \\ x_m(t) \end{bmatrix} = \begin{bmatrix} h_{11} & \dots & h_{1n} \\ \vdots & \ddots & \vdots \\ h_{m1} & \dots & h_{mn} \end{bmatrix} \begin{bmatrix} s_1(t) \\ \vdots \\ s_n(t) \end{bmatrix} \quad (9)$$

在相同实体表象属性特征范围内, 应用文本匹配模型, 编辑距离函数衡量数据库语义表象属性之间的相似度, 通过系统重建, 采用属性选择算法, 当 $s = Ax$, 求解矩阵的逆矩阵, 得到组合属性 s 与属性级的相似度, 如前文所述, 通过建构的分离矩阵 $W = (w_{ij})_{n \times n}$ 得以实现数据集的初始文本相似集计算, 得到辅助信息库。源数据集信息中语义特征 x 经过分离矩阵 W 逆变换后, 产生输出特征为

$$y = [y_1, y_2, \dots, y_n]^T \quad (10)$$

综上所述, 采用语义自相关特征成分分析的语义分析模式可表示为

$$y(t) = Wx(t) = WHs(t) = Gs(t) \quad (11)$$

式中, G 为全局传输矩阵(或全局系统矩阵)。采用语义信息约束规则, 优化逼近数据库实体识别结果, 通过自适应调整 $G = I$ (I 为 $n \times n$ 阶单位矩阵), 得到 $y(t) = s(t)$, 最终达到快速分离源信号的目的, 以适应 Deep Web 数据库动态关联完备性要求, 使用语义自相关特征建立表象关联知识库, 实现数据查询过程中的纠错、消重、整合。语义自相关特征数据库实体识别匹配原理图如图2所示。

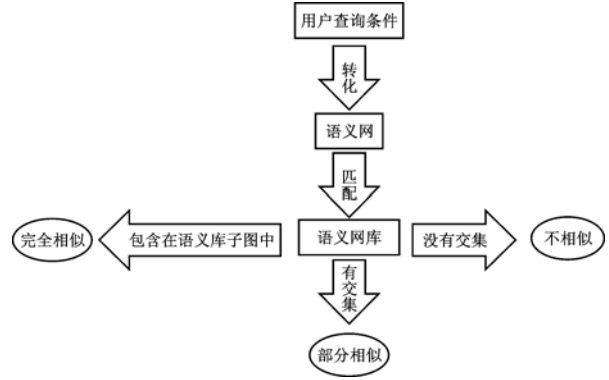


图2 语义自相关特征数据库实体识别匹配原理图

3.2 数据库实体识别查询稳定性分析

在算法实现的基础上, 为从数学原理上证明查询识别算法的稳定性, 采用语义信息约束规则, 优化逼近数据库实体识别结果, 以适应 Deep Web 数据库动态关联完备性要求, 并进行稳定性分析, 采用数学证明的方式进行研究验证。对于适当选择的语义信息约束函数 $J_i^*(\cdot)$ 的临界状态值, 使得数据库实体识别查询目标函数递减, 需成立:

引理1 Deep Web 数据库系统实体识别预测序列的时间 t_0 时刻存在最优特征解的集合, 那么存在正常数:

$$\xi = 2\rho_{\max} \lambda_{\max}(Q_i) N \delta^2 K T \quad (12)$$

数据库优化逼近控制函数满足条件:

$$J^*(x(t_{k+1})) - J^*(x(t_k)) \leq - \sum_{i=1}^N \int_{t_k}^{t_{k+1}} L_i(x_i^*(s; t_{k+1}), u_i^*(s; t_{k+1})) ds + \xi \delta^2 \quad (13)$$

式中:

$$\rho_{\max} \geq \|x(s; t_{k+1}) - x_i^c\| \quad (14)$$

上式表示为 Deep Web 数据库系统矢量函数的状态偏移点出的最大临界值, $\lambda_{\max}(Q_i)$ 为状态系数耦合矩阵 Q_i 的最大特征值。通过上述引理可知, 通过选择合适的状态耦合系数, 得到数据库语义自相关特征的最优值函数 $J^*(\cdot)$ 为有界函数。通过采样时间 δ 趋于最小值时, 可得 $J^*(\cdot)$ 是单调递减函数, 此时有一个合理的 δ 值存在使得 $\forall$ 时刻 $\tau \in [t_k, t_k + \delta]$, 即

$$J^*(x(\tau)) - J^*(x(t_k)) \leq - \sum_{i=1}^N \int_{t_k}^{t_{k+1}} \|x_i(t_k) - x_i^c\|_{Q_i}^2 ds \quad (15)$$

由于 $J^*(\cdot)$ 是数据库实体识别查询系统稳定性的前提条件, 式(14)的成立表明, 该查询系统是渐进稳定的。在无穷时域空间积分存在且具有有界

特征,对 $\forall x(t_0)$,在有效查询空间内使得 $x(T) \in W_p$,保证了数据库查询识别系统的稳定性.

4 仿真实验与结果分析

为了测试本文提出的数据库查询算法的可行性,进行实际 Deep Web 数据库查询索引仿真实验,算法采用 Windows 7 硬件环境,并以 SQL 和 MATLAB 语言混合编程,实现算法仿真.通过随机生成了 Web 数据集数据长度为 100 200 的预测数据时间序列,采样得到 4 次采样,构建 Deep Web 数据库级联系统 i .采用查全率和查准率进行检索质量指标的测评.查准率是检索到的关联信息数目与检索到的所有信息数目之间的比值.查全率表示检索到的相关信息数与数据库中检索需求相关的所有信息数之间的比值.本文选用了搜索引擎 Lucene 为实验对象,提供一定数量的样本,在不断扩大检索样本的覆盖面中,提取语义自相关特征,采用本文方法和传统方法进行 Deep Web 数据库实体识别查询实验,得到两种方法下的数据库文本匹配性能分布图如图 3 所示,图中可见采用本文算法,匹配性能分布较传统方法平滑,没有出现折叠重合,有效实现了数据库查询过程中纠错、消重和整合效果.

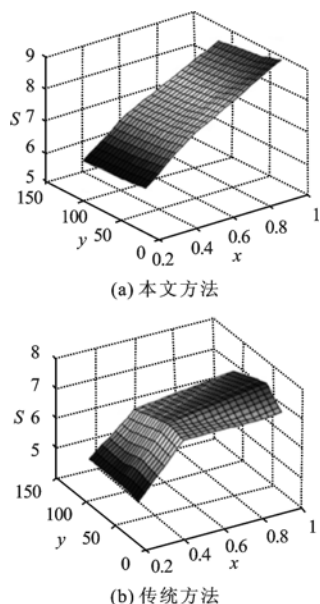


图 3 两种方法下的数据库文本匹配纠错、消重和整合性能

通过计算得到查准率关系曲线对比如图 4 所示.可见,随着查全率增大,关键词检索查准率比本文提出的语义自相关特征检索查准率下降得快,说明了关键词检索存在许多不符合检索要求的信息,主要原因是关键词检索只是对简单的字、词进行匹

配,然而对词语间的关系并没有进行分析理解,可见本文算法得到数据库查询的查准率远高于传统算法.

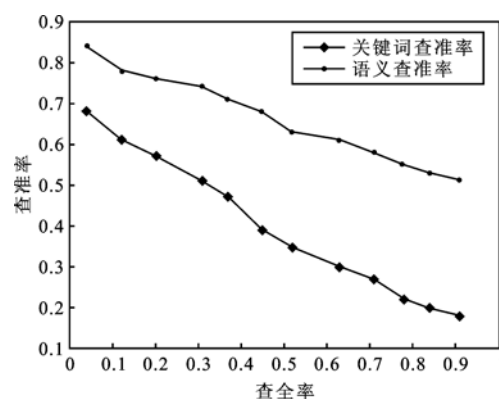


图 4 查准率关系曲线对比

5 结束语

本文着力于解决 Deep Web 数据库识别查询中数据纠错、消重和整合问题,优化 Deep Web 数据库实体查询性能.给出了智能语义自相关特征求解算法,提出一种基于智能语义自相关特征的 Deep Web 数据库优化识别查询模型.分别设计了文本匹配模型、语义自相关特征分析模型和分组统计模型构成,通过语义信息约束规则,优化逼近数据库实体识别结果,以适应 Deep Web 数据库动态关联完备性要求,使用语义自相关特征建立表象关联知识库,最终通过仿真实验实现数据查询过程中的纠错、消重、整合.通过数学推导证明了算法的稳定性.算法能综合考虑文本特征、语义自相关特征和约束规则,使得数据库查询识别结果不断精化,查准率和查全率提高显著,比传统的关键词检索方法更具应用前景.

参考文献:

- [1] 罗利民,周震. 基于 IPV6 的网络安全入侵检测技术研究[J]. 科技通报, 2012, 28(4):114-116.
- [2] 马军,宋玲,韩晓晖,等. 基于网页上下文的 Deep Web 数据库分类[J]. 软件学报, 2008,19(2):267-274.
- [3] 凌妍妍,孟小峰,刘伟. 基于属性相关度的 Web 数据库大小估算方法[J]. 软件学报, 2008, 19(2):224-236.
- [4] 覃雄派,王会举,李芙蓉,等. 数据管理技术的新格局[J]. 软件学报, 2013, 24(2):175-197.
- [5] 刘伟,孟小峰,孟卫一. Deep Web 数据集成研究综述[J]. 计算机学报, 2007, 30(9):1475-1489.

(下转第 166 页)

在负载电流为 0、10 μA 、500 μA 条件下分别为 89.2°、75.9°、49.8°, 在负载电容为 60 pF、200 pF、500 pF 条件下分别为 71.4°、66.5°、48.5°, 保证了系统的稳定。

瞬态测试结果如图 6 所示, 在低于 2.5 V 的情况下, 输出保持和输入电压相同。在高于 2.5 V 的情况下, 输出电压则稳定在 2.5 V, 电压波动会在 2 μs 内稳定, 过冲电压的最大变化为 3.4%, 达到了输出稳定电压的效果。

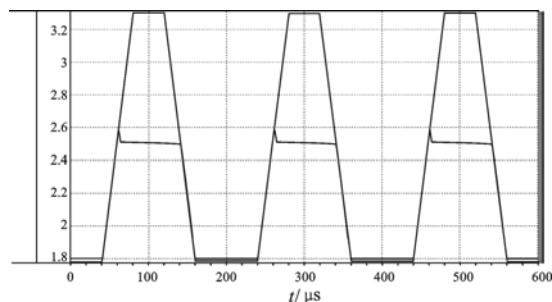


图 6 LDO 的瞬态响应

5 结束语

文中设计了一种低功耗, 高电源抑制比的 LDO, 在 TSMC 0.18 μm CMOS 工艺下, 静态电流仅有 10 μA , 非常低的功耗适合于在芯片上集成。同时, 整体结构的 PSR 也得到了改进, 低频下达到 -45 dB, 最差的情况也达到了 -20 dB 左右, 很好地抑制了输入电源电压的纹波噪声, 更适合于对噪声敏感的电路。

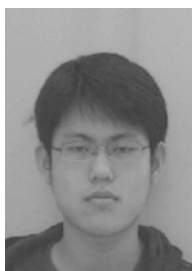
参考文献:

- [1] Tsz Yin Man, Philip K T Mok, Mansun Chan. Design of area-efficient, low-quiescent-current LDOs for chip-

level power management[C]//Integrated Circuits, Singapore: IEEE, 2007, 26-28.

- [2] Huei-Sheng Jhuang, Jia-Hui Wang, Zi-Yu Zeng, et al. A low dropout linear regulator with high power supply rejection[C]// ISIC 09. Proceedings of the 2009 12th International Symposium on Integrated Circuits, 2009: 41-44.
- [3] Wenguan Li, Ruohe Yao, Lifang Guo. A CMOS low-dropout regulator with high power supply rejection[C]// IEEE International Conference of Electron Devices and Solid-State Circuits. Xi'an: IEEE, 2009: 384-387.
- [4] Gupta V, Rincon-Mora G A. A low dropout, CMOS regulator with high PSR over wideband frequencies[C]// IEEE Int Symp Circuits Syst. [s. l.]: IEEE, 2005: 4245-4248.

作者简介:



金禹铮 男, (1988—), 硕士。研究方向为集成电路工程。



付宇卓 男, 教授, 博士生导师。研究方向为系统级芯片设计。

(上接第 162 页)

- [6] 闫源江, 胡光波. 关联维估计在舰船辐射噪声特征提取中的应用[J]. 科学技术与工程, 2010, 10(27): 6756-6759.
- [7] 方露. 浅析物联网技术在科技馆中的应用[J]. 科技通报, 2012, 28(3): 165-168.
- [8] 蔡勇智. 一种改进的神经网络算法车牌识别算法研究[J]. 科技通报, 2012, 28(10): 128-130.
- [9] 周迪斌. 基于多 GPU 的千万级高维空间实时检索

[J]. 科技通报, 2013, 29(1): 118-123.

- [10] 柳虹, 周根贵, 傅培华. 分层供应链复杂网络局部演化模型[J]. 计算机科学, 2013, 40(2): 270-273.

作者简介:

王海荣 女, (1981—), 硕士, 讲师, 软件设计师。研究方向为计算机应用、数据挖掘、图形图像处理。