DOI:10.13542/j.cnki.51-1747/tn.2021.01.007

基于用户访问树的分布式Web日志挖掘算法

陈宝国,宋　旸

(淮南师范学院　计算机学院,安徽　淮南　232000)

摘要:为了提高对分布式Web日志数据的准确挖掘能力,提出基于用户访问树的分布式Web日志挖掘算法。构建分布式Web日志的信息分布式检测模型,采用模糊信息粗糙集调度方法进行分布式Web日志信息的结构重组,提取分布式Web日志的统计特征量,采用用户访问树特征聚类方法进行分布式Web日志数据的空间分布式重组,结合粗糙集特征匹配方法进行分布式Web日志的离散融合处理,对多层分布式数据库中的主成分特征分量进行关联规则融合,结合信息融合结果进行分布式Web日志数据的特征参量聚集式调度,提取分布式Web日志的谱特征分量,采用空间信息聚类方法,实现分布式Web日志的用户访问树模型构造,结合决策树模型构建分布式Web日志挖掘的适应度参数,实现分布式Web日志挖掘。仿真结果表明,采用该方法进行分布式Web日志挖掘的准确性较高,抗干扰性较好,提高了分布式Web日志挖掘和用户信息访问能力。

关键词:用户访问树;分布式Web;日志挖掘
中图分类号:TP391　　**文献标志码**:A　　**文章编号**:2095-5383(2021)01-0026-04

Distributed Web Log Mining Algorithm based on User Access Tree

CHEN Baoguo, SONG Yang

(School of Computer Science, Huainan Normal University, Huainan 232000, China)

Abstract: In order to improve the accurate mining ability of distributed Web log data, a distributed Web log mining algorithm based on user access tree was proposed. An information distributed detection model of distributed Web logs was constructed, a fuzzy information rough set scheduling method was adopted to carry out structural reorganization of distributed Web log information, statistical feature quantities of distributed Web logs were extracted, a user access tree feature clustering method was adopted to carry out spatial distributed reorganization of distributed Web log data, a rough set feature matching method was combined to carry out discrete fusion processing of distributed Web logs, association rule fusion was carried out on the principal component characteristic components in the multi-layer distributed database, the characteristic parameter aggregation scheduling of the distributed Web log data was carried out in combination with the information fusion result, the spectral characteristic components of the distributed Web log were extracted, the spatial information clustering method was adopted to realize the construction of the user access tree model of the distributed Web log, and the fitness parameter of the distributed Web log mining was constructed in combination with the decision tree model to realize the distributed Web log mining. The simulation results show that the method has higher accuracy and better anti-interference performance in distributed Web log mining, and improves the ability of distributed Web log mining and user information access.

Keywords: user access tree; distributed Web; log mining

随着分布式Web日志数据信息处理技术的发展,相关的分布式Web日志挖掘和检测方法研究受到人们的极大重视[1]。对分布式Web日志的挖掘和检测方法研究是在对分布式Web日志的用户特征分布式融合的基础上,采用信息聚类和大数据挖掘方法进行的。为了提高数据的准确挖掘能力,本文提出了一种基于用户访问树的分布式Web日志挖掘算法。该算法通过构建分布式Web日志的信息分布式检测模型,结合信息融合结果进行分布式Web日志数据的特征参量聚集式调度,提取分布式Web日志的谱特征分量,采用空间信息聚类方法,实现了分布式Web日志的用户访问树模型构造。在此基础上结合决策树模型构建分布式Web日志挖掘的适应度参数,实现分布式Web日志挖掘优化。最后进行仿真测试分析,验证其有效性。

1　分布式Web日志数据结构分析和特征重组

1.1　分布式Web日志分布式体系结构

为了实现对分布式Web日志优化挖掘,结合深

收稿日期:2020-04-13
基金项目:安徽高校自然科学重点研究项目(KJ2018A0469);淮南师范学院科研项目(2019XJYB14)
作者简介:陈宝国(1978—),男,讲师,硕士,研究方向:数据挖掘、图形图像处理、算法设计。
宋旸(1979—),女,助教,硕士,研究方向:数据挖掘、图形图像处理、算法设计,电子邮箱:lllntdx@163.com。

度学习进行分布式 Web 日志存储节点结构重组,并对分布式 Web 日志输出信息进行融合,再结合用户的信息特征分布建立 Slope one 填充模型。采用 Web 日志源的多层次聚集方法进行分布式 Web 日志模糊调度,得到分布式 Web 日志信息聚类中心和 Pearson 相关系数。分布式 Web 日志 Pearson 相关系数用有向图表示为 $G=(V,E)$,对分布式 Web 日志图模型进行结构重组,其中:V 是分布式 Web 日志的相似度特征集;E 是分布式 Web 日志的模糊聚类中心。采用分布式 Web 日志的模糊信息聚类分析方法,得到分布式 Web 日志的有向边集合[2]。假设 $M_1,M_2,\cdots,M_N$ 为多分布式 Web 日志的信息存储节点,根据目标用户对目标项目的评分信息进行关联树结构分析,得到分布式 Web 日志的用户访问树结构模型如图 1 所示。

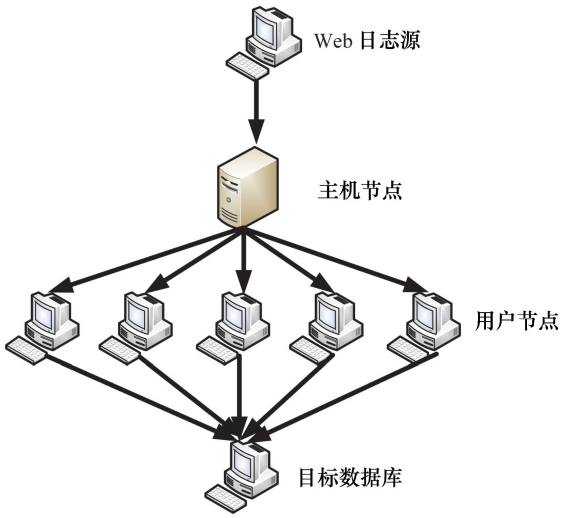

图 1　分布式 Web 日志的用户访问树结构模型

结合图 1,根据目标用户相关联的特征分析,计算分布式 Web 日志检测的测度信息,在分布式 Web 日志存储网络结构模型中,得到目标用户相关联系数为 $W=\{u,w_1,w_2,\cdots,w_k\}$,在分布式 Web 日志的信息覆盖区域,假设目标用户相关联特征量为 M,对目标用户相关联分布集,采用模糊信息聚类方法,得到目标用户相关联特征分布形式为 $x(k-1),\cdots,x(k-M)$,以用户对项目类型的平均评分,得到用户访问的差分量化集为 $x_s=[x(\eta_1),\cdots,x(\eta_N)]^{\mathrm{T}}$,特征估计值:

$$\hat{x}_s=W_s^{\mathrm{T}}y \tag{1}$$

基于优先级划分方法,构建分布式 Web 日志的共同评分的项目用集合为:

$$r(t)=\sum_i\sum_{j=0}^{N_f-1}\sum_{l=0}^{L-1}b_i\alpha_l p(t-iT_s-jT_f-c_jT_c-\tau_l)+$$
$$\omega(t)=\sum_i\sum_{j=0}^{N_f-1}b_i p_h(t-iT_s-jT_f-c_jT_c-\tau_0)+\omega(t) \tag{2}$$

其中:$p_h(t)=\sum_{l=0}^{L-1}\alpha_l p(t-\tau_{l,0})$。

另外,$\omega(t)$ 为目标项目的评分信息,$p_h(t)$ 为分布式 Web 日志信息分布集。根据上述分析,构建分布式 Web 日志分布式体系结构模型,结合共同评分的项目用集合的特征分布集,进行分布式 Web 日志分布式体系融合设计[3]。

1.2　分布式 Web 日志特征提取

将用户近邻与项目近邻评分信息融合,进行分布式 Web 日志的特征重构,提取分布式 Web 日志的统计特征量[4],对 Web 日志数据中各自已评分项目进行特征融合,得到分布式 Web 日志的时隙特征分布值为:

$$\mathrm{ITrust}_{a\to c}=\frac{\sum_{b\in adj(a,c)}\mathrm{DTrust}_{a\to b}\times(\mathrm{DTrust}_{b\to c}\times\beta_d)}{\sum_{b\in adj(a,c)}\mathrm{DTrust}_{a\to b}} \tag{3}$$

根据特征提取结果,进行分布式 Web 日志的表面信息重构[5],计算用户间的相似度,得到分布式 Web 日志源的频谱 Z 服从参数为 β_d 的高斯分布:

$$\beta_d=(\mathrm{MPDist}-d+1)/\mathrm{MPDist},d\in[2,\mathrm{MPDist}] \tag{4}$$

其中:$adj(a,c)$ 表示用户 u 和用户 v 之间的相似度特征分布集。

对分布式 Web 日志源进行本体结构重组,在分布式 Web 日志进行稀疏性重组,得到分布式 Web 日志的相似性权重[6],得到分布式 Web 日志的映射关系表示为 $A\to B,B\to C$,推出回归分析模型为:

$$\mathrm{MSD}_{a\to b}=1-\frac{\sum_{i=1}^{|I_{a,b}|}\sqrt{(d_{a,i}-\bar{d}_a)^2+(d_{b,i}-\bar{d}_b)^2}}{|I_{a,b}|\times\sum_{i=1}^{|I_{a,b}|}\left[\sqrt{(d_{a,i}-\bar{d}_a)^2}+\sqrt{(d_{b,i}-\bar{d}_b)^2}\right]} \tag{5}$$

引入了用户对项目类型的偏好,进行分布式 Web 日志的日志信息挖掘,输出分布式 Web 日志的用户评分互信息参数为:

$$I(Q,S)=H(Q)-H(Q\mid S) \tag{6}$$

其中:

$$H(Q \mid s_i) = -\sum_j \left[p_{sq}(s_i,q_j)/p_S(s_i)\right] \cdot \log_2\left[p_{sq}(s_i,q_j)/p_S(s_i)\right] \quad (7)$$

设定 Iu 为用户 u 所有已评分的项目集合,采用核心数据融合聚类分析方法,提分布式 Web 日志的惯性参数,结合用户的偏好进行特征融合和信息聚类[7]。

2 分布式 Web 日志的挖掘优化

2.1 分布式 Web 日志的特征提取

在上述采用相空间重构方法进行分布式 Web 日志的特征重构的基础上,结合粗糙集特征匹配方法进行分布式 Web 日志的离散融合处理,对多层分布式数据库中的主成分特征分量进行关联规则融合,根据分布式 Web 日志的相关性信息分布[8],得到分布式 Web 日志挖掘的特征提取判决准则满足:

准则(1):

$$\sqrt{\frac{R^2_{(m+1)n}-R^2_{mn}}{R^2_{(m+1)n}}} = \frac{|x_{\eta(n)+m\tau}-x_{n+m\tau}|}{R_{(m+1)n}} \geqslant R_{tol} \quad (8)$$

准则(2):

$$\frac{R_{(m+1)n}}{\sqrt{\frac{1}{N}\sum_{k=1}^{N}\left[x_k-\frac{1}{N}\sum_{k=1}^{N}x_k\right]^2}} > A_{tol} \quad (9)$$

根据寻找数据对象的邻域半径[9],结合判决准则,进行主成分分析。在数据的特征分布属性集中,设$\{u_1,u_2,\cdots,u_N\}$为用户访问分布集合,$\{v_1,v_2,\cdots,v_M\}$为待预测评分的项目集合,$R=[R_{u,v}]_{N\times M}$为属性规则集,结合用户访问特征,构建用户访问规则树模型,得到模糊递推公式为:

$$p_i^* = \frac{1}{\sum_{j=i}^{N}\frac{2m_j}{\sum_{k=j+1}^{N+1}L_k p_k - \sum_{k=j}^{N}E_k}} - 1, i=1,2,\cdots,N+1 \quad (10)$$

分布式 Web 日志数据集有 n 个数据对象,则其粗糙概念集为:

$$X(n) = \{x(n),x(n+\tau),\cdots,x(n+(m-1)\tau)\}$$
$$n=1,2,\cdots,N \quad (11)$$

其中:τ 表示分布式 Web 日志的调度延迟[10]。

2.2 分布式 Web 日志

结合粗糙集特征匹配方法进行分布式 Web 日志的离散融合处理,对多层分布式数据库中的主成分特征分量进行关联规则融合,分布式 Web 日志特征构成的用户访问树模型为:

$$p(y \mid \alpha,\theta) = \sum_{k=1}^{K}\alpha_k p_k(y \mid \mu_k, \sum{}_k) \quad (12)$$

挖掘分布式 Web 日志的数据对象邻域,得到核心对象中 Web 日志用户访问树结构模型:

$$\max_{x_{a,b,d,p}} \sum_{a\in A}\sum_{b\in B}\sum_{d\in D}\sum_{p\in P} x_{a,b,d,p}\, V_p \quad (13)$$

$$\text{s.t.} \sum_{a\in A}\sum_{d\in D}\sum_{p\in P} x_{a,b,d,p}\, R_p^{bw} \leqslant K_b^{bw}(S), b\in B \quad (14)$$

计算数据集中数据对象两两间的欧式距离,得到第 i 个分布式 Web 日志的散乱点集为 $P_i=(p_{i1},p_{i2},\cdots,p_{iD})$其中:

$$j\in N_i(k), N_i(k)=\{\parallel x_j(k)-x_i(k)\parallel < r_d(k)\} \quad (15)$$

在核心对象中选取密度最大的 Web 日志的,调整分布式 Web 日志的关联规则项,构建分布式 Web 日志的初始簇心:

$$\begin{cases} a(H_{ac}) = 1-\dfrac{H_{ac}}{\max(H_{ac})+l} \\ \max(H_{ac}) = \log_2 k \end{cases} \quad (16)$$

分布式 Web 日志挖掘的边界函数为:

$$w_{j\vec{i}}(k+1) = w_{j\vec{i}}(k) - \alpha\frac{\partial F}{\partial w_{j\vec{i}}} \quad (17)$$

$$z_{kj}(k+1) = z_{kj}(k) - \alpha\frac{\partial F}{\partial z_{kj}} \quad (18)$$

在初始簇密度很高的情况下,采用用户访问树结构重组方法,进行分布式 Web 日志的多维空间信息挖掘,提取分布式 Web 日志的谱特征分量,采用空间信息聚类方法,实现分布式 Web 日志的用户访问树模型构造,得到分布式 Web 日志挖掘输出为:

$$\boldsymbol{X}=[s_1,s_2,\cdots,s_K] = \begin{bmatrix} x_1 & x_2 & \cdots & x_K \\ x_{1+\tau} & x_{2+\tau} & \cdots & x_{K+\tau} \\ \vdots & \vdots & \vdots & \vdots \\ x_{1+(m-1)\tau} & x_{2+(m-1)\tau} & \cdots & x_{K+(m-1)\tau} \end{bmatrix} \quad (19)$$

其中:$K=N-(m-1)\tau$,表示分布式 Web 日志数据的空间嵌入维数;τ 为时延;m 为各个初始中心的用户访问节点数;$s_i=(x_i,x_{i+\tau},\cdots,x_{i+(m-1)\tau})^{\mathrm{T}}$ 是分布式 Web 日志的空间嵌入特征量。综上分析,采用用户访问树模型结构,实现分布式 Web 日志挖掘。

3 仿真实验与结果分析

为验证本文方法的有效性,进行仿真实验。采用 100 KB 的测试数据集,对 Web 日志数据的聚类簇

数设定为 7,对分布式 Web 日志数据信息融合的关联系数设定为 0.67,分布式 Web 日志填充的回归系数为 0.34,分布式 Web 日志信息聚类中心设定为 $m=3$,用户访问的时延 $\tau=8$,给出分布式 Web 日志的原始信息分布如图 2 所示。

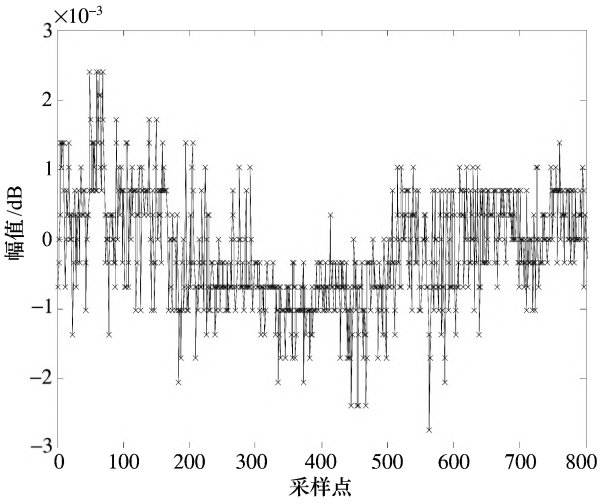

图 2　分布式 Web 日志的原始信息分布

以上述数据为研究样本,结合信息融合结果进行分布式 Web 日志数据的特征参量聚集式调度,提取分布式 Web 日志的谱特征分量,得到挖掘结果如图 3 所示。

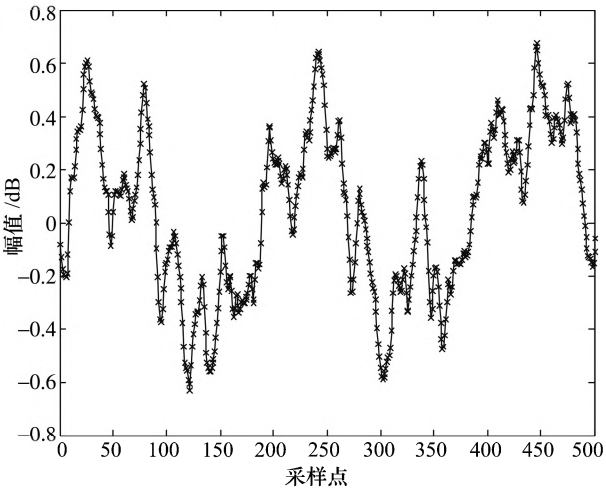

图 3　分布式 Web 日志挖掘结果

由图 3 可知,采用本文方法能有效实现对分布式 Web 日志挖掘。图 4 为优化结果,由图 4 可知,本文方法的挖掘精度较高,迭代次数较小,说明挖掘的收敛性和抗干扰性较好。

4　结语

网络的发展带动了 Web 的发展,人们对 Web 站点的设计和功能提出了更高的要求,要求其具有智能性,能快速、准确地找到用户所需信息,能为不同用户提供不同的服务以及产品营销策略信息等等。由于 Web 服务器的日志记录了用户访问站点时所访问的页面、时间、用户等信息,因此对分布式 Web 日志进行挖掘,从而为管理者优化网站结构以及提升网站性能提供决策具有重要意义,所以本文提出基于用户访问树的分布式 Web 日志挖掘算法。实验测试结果表明,采用本文方法进行分布式 Web 日志挖掘的准确性较高,优化性能较好,可以在实际中得到进一步推广。

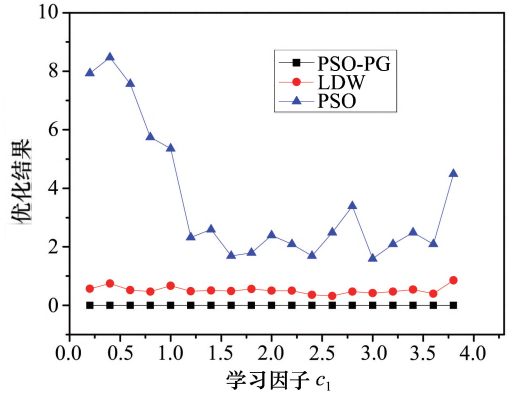

图 4　分布式 Web 日志挖掘的优化性能对比

参考文献:

[1]张晶,陈诚,郑焕科. 面向软件漏洞检测的 Fuzzing 样本优化方法[J]. 山东大学学报(理学版),2019,54(9):1-8,35.

[2]汤俊伟,刘家帧,李瑞轩,等. Android 应用软件漏洞静态挖掘技术[J]. 华中科技大学学报(自然科学版),2016,44(S1):20-24.

[3]蔡军,邹鹏,杨尚飞,等. 软件漏洞分析中的脆弱点定位方法[J]. 国防科技大学学报,2015,37(5):141-148.

[4]余传明,冯博琳,田鑫,等. 基于深度表示学习的多语言文本情感分析[J]. 山东大学学报(理学版),2018,53(3):13-23.

[5]吴磊,原鹏,丁维龙. 智能家居网关与云服务器数据同步协议的研究[J]. 计算机技术与发展,2018,28(9):151-155.

[6]李佳,范巍. 基于改进 D-S 证据理论的网络入侵检测[J]. 控制工程,2017,24(11):2362-2367.

[7]王晓雷,陈云杰,王琛,等. 基于 Q-learning 的虚拟网络功能调度方法[J]. 计算机工程, 2019, 45(2): 64-69.

[8]肖云鹏,孙华超,戴天骥,等. 一种基于云模型的社交网络推荐系统评分预测方法[J]. 电子学报,2018,46(7):1762-1767.

[9]王刚,郭雪梅. 社交网络环境下基于用户行为分析的个性化推荐服务研究[J]. 情报理论与实践,2018,41(8):102-107.

[10]SUN J W, WU Y Y, CUI G Z, et al. Finite-time real combination synchronization of three complex-variable chaotic systems with unknown parameters via sliding mode control [J]. Nonlinear Dynamics,2017,88(3): 1677-1690.