

基于动态聚类分析的网络用户相似性特征搜索研究

马 杰

(南京审计大学 实验中心,江苏 南京 211815)

摘 要: 网络用户的相似性特征体现了用户的偏好,为了提高网络信息的自适应推荐能力,提出一种基于动态聚类分析的网络用户相似性特征搜索算法.采用非平稳随机序列分析方法构建网络用户信息传递模型,提取反映网络用户行为特征的关联规则特征量,采用关联信息标注方法进行网络用户相似性行为特征信息检索,对提取的网络用户相似性特征信息进行动态聚类分析,采用模糊聚类方法进行相似性行为特征信息的自适应归类识别,提高网络用户相似性特征的自适应搜索和信息处理能力.仿真结果表明,采用该方法进行网络用户相似性特征搜索的准确性较高,对网络用户的相似性行为特征信息的聚类识别能力较好,从而提高了网络信息的优化推荐能力.

关键词: 动态聚类;网络用户;特征搜索;特征提取;信息处理

中图分类号: TP393

文献标识码: A

文章编号: 1000—7180(2018)09—0113—05

Research on Similarity Feature Search of Network Users Based on Dynamic Clustering Analysis

MA Jie

(Experiment center, Nanjing Audit University, Nanjing, 211815 China)

Abstract: The similarity features of network users reflect the preferences of users, in order to improve the adaptive recommendation ability of network information, A similarity feature search algorithm for network users based on dynamic clustering analysis is proposed. Non—stationary random sequence analysis method is used to construct the network user information transfer model, and the association rule feature quantity which reflects the network user behavior feature is extracted, and the association information annotation method is used to retrieve the network user similarity behavior characteristic information. Based on the dynamic clustering analysis of the extracted network users' similarity feature information, the fuzzy clustering method is used for adaptive classification and recognition of the similarity behavior feature information, so as to improve the ability of adaptive searching and information processing of the network users' similarity features. The simulation results show that the proposed method is more accurate in searching network users' similarity features, and it has a better ability of clustering and recognition of network users' similarity behavior information, thus improving the ability of optimizing and recommending network information.

Key words: dynamic clustering; Network users; feature search; feature extraction; Information processing

1 引言

传统方法中,对网络用户相似性特征搜索方法主要有模糊 C 均值搜索方法、基于支持向量机的用户行为特征量提取和搜索方法、自适应分区搜索方

法以及决策树搜索方法等^[1,2],现 采用负载均衡控制和关联规则挖掘方法进行网络用户相似性特征检索和调度,结合模糊分区聚类方法进行网络用户的行为信息挖掘,提高用户相似性特征搜索能力,文献[3]中提出一种基于自适应均衡控制和关联规则

挖掘的网络用户行为特征搜索和资源混合推荐模型,提取网络用户相似性行为特征信息和语义信息,通过关键词索引方法实现网络用户相似性行为特征信息检索;文献[4]中提出一种基于语义本体模型构建的网络用户行为相似性特征挖掘方法,采用连贯性分析和阈值判断方法进行用户行为信息检索和自动识别,结合语义分析方法进行网络用户信息挖掘,提高信息搜索精度,

本文提出一种基于动态聚类分析的网络用户相似性特征搜索算法,首先构建的网络用户相似性特征信息模型,采用关联信息标注方法进行网络用户相似性行为特征信息检索,然后对提取的网络用户相似性特征信息进行动态聚类分析,实现网络用户相似性特征搜索和信息处理。

2 网络用户行为特征的信息模型构建和预处理

2.1 网络用户相似性行为特征信息建模

为了实现对网络用户相似性特征的优化搜索,首先采用非线性时间序列分析和非平稳随机序列分析结合的方法进行网络用户行为特征信息模型构建^[5],假设待搜索的网络用户相似性特征信息流观测时间序列 $\{x_1, x_2, \dots, x_N\}$,令 $x(n)$ 为一组待检索的用户属性对象集,在 m 维嵌入空间中进行网络用户相似性特征信息的分布式重组结构式为:

$$X(n) = \{x(n), x(n+\tau), \dots, x(n+(m-1)\tau)\} \quad (1)$$

式中, τ 表示网络用户相似性特征信息在高维空间中的采样时间延迟。以 x_1, x_2, x_3 和 x_4 为4个信息检索通道,得到网络用户行为特征信息检索的信道结构模型表示为:

$$\begin{cases} x_1 = p_1 - m \\ x_2 = p_2 - m \\ x_3 = p_3 - m \\ x_4 = m \end{cases} \quad (2)$$

式中, m 为空间采样嵌入维数; p_i 表示网络用户个体。采用混合信息传导模型进行网络用户行为特征分解,得到一组反映网络用户行为特征量的时序为:

$$x(t) = (x_0(t), x_1(t), \dots, x_{k-1}(t))^T \quad (3)$$

采用一个 $1 \times N$ 的窗口进行网络用户相似性行为特征量的自适应加窗处理,确定窗口大小 N ,在横向和竖向两个梯度方向对网络用户相似性行为特征进行自适应分区调度,输入网络信息的检测函数

$x(t)$,采用自适应均衡调度方法进行网络用户特征信息的优化调度^[6,7],得到网络用户信息的空间采样节点分布距离为:

$$d_j = \sum_{i=0}^{k-1} (x_i(t) - \omega_j(t))^2, j = 0, 1, \dots, N-1 \quad (4)$$

式中, ω_j 为网络用户相似性行为特征分布的最大梯度差加权系数,表示为:

$$\omega_j = (\omega_{0j}, \omega_{1j}, \dots, \omega_{k-1,j})^T \quad (5)$$

进行网络用户相似性行为特征信息的语义指向性作为信息聚类中心矢量,得到用户信息特征量的分布阈值函数为 TL_X, TL_Y ,网络用户信息的特征聚类的判决式为:

$$TL_X(x, y) = \begin{cases} Text & , if (GD_X(x, y) > T_X \\ NonText & , Otherwise \end{cases} \quad (6)$$

根据上述聚类规则,采用相空间重构方法构建网络用户相似性信息重构向量模型为:

$$\mathbf{X} = [s_1, s_2, \dots, s_K] = \begin{bmatrix} x_1 & x_2 & \dots & x_K \\ x_{1+\tau} & x_{2+\tau} & \dots & x_{K+\tau} \\ \dots & \dots & \dots & \dots \\ x_{1+(m-1)\tau} & x_{2+(m-1)\tau} & \dots & x_{K+(m-1)\tau} \end{bmatrix} \quad (7)$$

式中 $K = N - (m-1)\tau$,表示的网络用户相似性特征信息搜索的空间分布维, τ 为信息采样时延,由此构建网络用户相似性行为特征信息模型,采用大数据挖掘和模糊聚类方法进行相似性特征搜索,提高网络用户的行为特征识别能力。

2.2 关联规则特征提取

在对网络用户相似性特征信息流进行分布式结构重组和信息建模的基础上,提取反映网络用户行为特征的关联规则特征量,采用关联信息标注方法进行网络用户相似性行为特征信息检索^[7],采用结构化数据重组方法,构建网络用户信息特征分布模型表达为:

$$\begin{cases} x = (x_1, x_2, \dots, x_n) \\ y = F(x) = (f_1(x), f_1(x), \dots, f_m(x))^T \end{cases} \quad (8)$$

式中, $x = (x_1, x_2, \dots, x_n)$ 为网络用户行为特征的相似度分布概念集; $y = F(x)$ 表示网络用户的行为特征信息的分配节点集;假设用户行为特征的语义向量集 n_i 的种类为 r_j ,采用关联规则挖掘方法,得到相似度评价集合 $P(n_i) = \{p_k \mid pr_{kj} = 1, k = 1, 2, \dots, m\}$,它表示网络用户相似度评价的优先级属性,利用语义本体特征重复性计数方法进行网络用户行

为特征的关联路径检索,得到关联系数为:

$$RTT_s = (1 - \alpha) \times RTT_s + \alpha \times RTT \quad (9)$$

通过优先级控制方法进行特征属性的优先级排序,设 $Color(u) = Gray$ 是 u 维的用户行为特征分布集, u 表示为一矢量,使用偏置向量搜索方法,对于搜索空间 $Color(u) = Uncolor; Sch(u)$, 存在:

$$Sch(u) = NY \quad (10)$$

当冗余信息的强度 $Sch(u) = RY$, 采用关联语义特征相似度评价方法^[8], 得到网络用户相似性特征的关联规则分布序列 $Sch(u) = Done$ 的离散采样样本集为:

$$x_n = [x(0), x(1), \dots, x(N-1)]^T \quad (11)$$

网络用户的相似性特征的分布数据集为 $X = \{x_1, x_2, \dots, x_n\}$, n 是行为属性的分配节点 X 的数目, X 为模糊信息特征值. 网络用户相似性特征分布集的条件概率为 $|d_{n-\max} - d_{n-\min}| \cdot (1/K)$, 关联规则分布向量集满足: $A_1 \cup A_2 \cup \dots \cup A_k = A$, 采用梯度下降训练方法, 得到关联规则特征量满足 $A_i \cap A_j = \Omega$, 其中 $i, j = 1, \dots, m$ 且 $i \neq j$. 根据关联规则挖掘结果, 得到网络用户行为特征频繁项集为:

$$P = \{p_1, p_2, \dots, p_m\}, m \in N \quad (12)$$

此时待挖掘的网络用户行为特征信息流为:

$$flow_k = \{n_1, n_2, \dots, n_q\}, q \in N \quad (13)$$

上式中, q 表示个性化行为特征节点位置, n_q 表示第 q 个用户个性化行为特征节点的传递时延, N 表示用户特征搜索的任务总数. 由此提取反映网络用户行为特征的关联规则特征量,

3 网络用户相似性特征搜索算法优化

3.1 相似性特征信息动态聚类

在构建网络用户信息传递模型, 提取反映网络用户行为特征的关联规则特征量的基础上, 进行网络用户相似性特征信息搜索优化设计, 本文提出一种基于动态聚类分析的网络用户相似性特征搜索算法, 采用分段统计分析方法, 得到相似性特征信息的特征聚类尺度:

$$\begin{aligned} \min F(x) &= (\{f_1(x), f_2(x), \dots, f_m(x)\})^T \\ s. t. \quad &g_i \leq 0, i = 1, 2, \dots, q \end{aligned} \quad (14)$$

$$h_j = 0, j = 1, 2, \dots, p$$

上式中, 在统计时间段 T 内, 用户相似性特征量的有限数据集 r_k 的动态聚类中心为 $v_i = \{v_{i1}, v_{i2}, \dots, v_{ip}\}$, 模糊聚类的离散特征矩阵为:

$$E = [E_G, E_T, E_W, E_L] \quad (15)$$

式中, 每个聚类中心的动态搜索半径满足 $C \subset S$, 采

用动态聚类融合跟踪识别方法, 得到网络用户的相似性特征量的空间谱增益为:

$$S_p(u) = \{F^p[s(t)]\}(u) = \int_{-\infty}^{\infty} K_p(t, u) s(t) dt \quad (16)$$

在空闲时间片内, 网络用户相似性特征信息的负载量为 $x_1, x_2, \dots, x_{m+1}$, 对用户行为信息搜索的动态调节函数为:

$$x_1 + x_2 + \dots + x_{m+1} = T + t - m \times t \quad (17)$$

采用分段插值方法进行网络用户相似性特征信息的动态聚类处理, 得到聚类输出为:

$$H_i(x) = \sum_{k=1}^K p_k \ln \frac{1}{p_k} = - \sum_{k=1}^K p_k \ln p_k \quad (18)$$

式中, K 表示用户相似性行为特征的关联度; p_k 为用户相似性特征的概率分布, 由此实现对网络用户相似性特征的信息动态聚类处理.

3.2 网络用户相似性特征的自适应搜索

采用关联信息标注方法进行网络用户相似性行为特征信息检索, 得到用户行为特征的空间采样数据集 $X = \{x_1, x_2, \dots, x_n\}$, 采用二维阈值判断方法进行用户相似性特征搜索^[9,10], 得到判断阈值为:

$$AT(v) = \max_{u \in FI(v)} [AT(u) + delay(v)] \quad (19)$$

$$RT(v) = \min_{w \in FO(v)} [RT(w) - delay(v)] \quad (20)$$

采用模糊聚类方法进行相似性行为特征信息的自适应归类识别, 得到用户相似性特征检索的边值为:

$$bmr_\beta(X) = R_\beta X - R_\beta X_1 \quad (21)$$

根据动态聚类结果进行属性判别, 得到用户相似性统计特征量为:

$$AVG_X = \frac{1}{m \times n} \sum_{x=1}^n \sum_{y=1}^m |G_X(x, y)| \quad (22)$$

式中, m, n 分别概念集和属性集的大小, 网络用户相似性特征搜索梯度特征差异分量 R_m 为 X_n 与 $X_{q(n)}$ 的距离, 信息搜索的边界函数为:

$$\begin{aligned} R_m &= \|X_{q(n)} - X_n\| (m)_2 = \min_{j=N_0, \dots, N, j \neq n} \|X_n - X_j\|_2 \\ &= \left(\sum_{l=0}^{m-1} (x_{q(n)+l} - x_{n+l})^2 \right)^{\frac{1}{2}} \end{aligned} \quad (23)$$

将相空间的嵌入维数从 m 增加到 $m+1$ 时, 根据特征模糊聚类结果, 进行用户相似性特征搜索的迭代, 迭代式为:

$$\begin{aligned} R_{(m+1)n} &= \|X_{q(n)} - X_n\| (m+1)_2 \\ &= \left(\sum_{l=0}^m (x_{q(n)+l} - x_{n+l})^2 \right)^{\frac{1}{2}} \end{aligned}$$

$$= (\lceil \|X_{\eta(n)} - X_n\| (m)_2 \rceil^2 + \lceil x_{\eta(n)+m\tau} - x_{n+m\tau} \rceil^2)^{\frac{1}{2}} \tag{24}$$

若 $R_{(m+1)n}$ 比 R_{mn} 大很多,网络用户相似性特征搜搜满足收敛条件,由此可见,本文方法进行网络用户行为相似性特征搜索过程是稳健收敛的,由此实现网络用户相似性特征搜索和信息动态聚类. 改进算法的实现流程如图 1 所示.

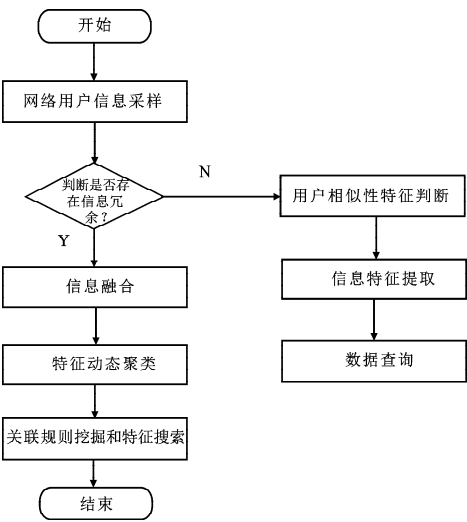


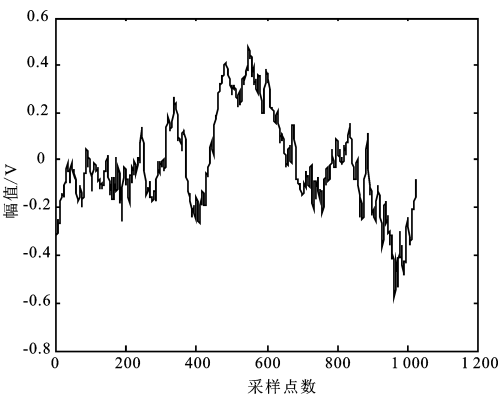
图 1 改进算法实现流程

4 仿真实验与结果分析

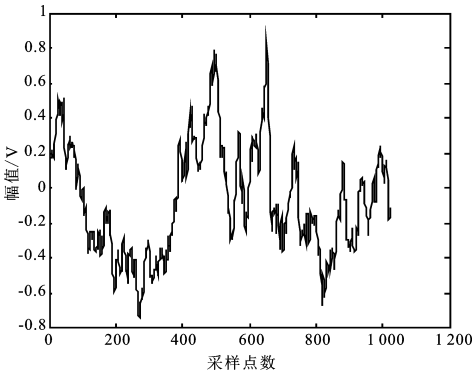
为了测试本文方法进行网络用户相似性特征搜索中的应用性能,进行仿真实验,实验采用 Matlab 7 实验设计,网络用户相似性特征信息集来自于云平台 Hadoop,对网络用户行为特征采集的时间间隔为 2.25s,测试集的规模为 12Gbit,特征采集的样本长度为 1024,动态聚类中心为 $[x,y,z]=[-1,0,1]$,迭代步长为 $h=0.01$,关联规则分布的维度为 4,数据采样的延迟为 $\tau=12$,根据上述仿真环境和参数设定,进行网络用户相似性特征搜索,得到用户行为特征的原始信息采集结果如图 2 所示.

以图 2 的网络信息为样本进行网络用户相似性特征搜索,采用关联信息标注方法进行网络用户相似性行为特征信息检索,实现用户相似性特征信息动态聚类,得到聚类结果如图 3 所示.

分析图 3 得知,采用本文方法进行网络用户相似性特征搜索的属性聚类性较好,提高了对网络用户行为特征的识别能力. 为了对比性能,采用不同的网络用户相似性特征搜索方法进行网络信息推荐,得到网络用户相似性特征搜索的准确率对比结果如图 4 所示.



(a)测试集



(b)训练集

图 2 用户行为特征的原始信息采集

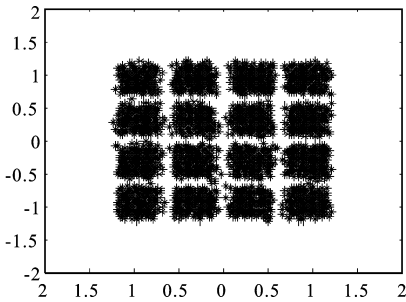


图 3 网络用户相似性特征信息动态聚类结果

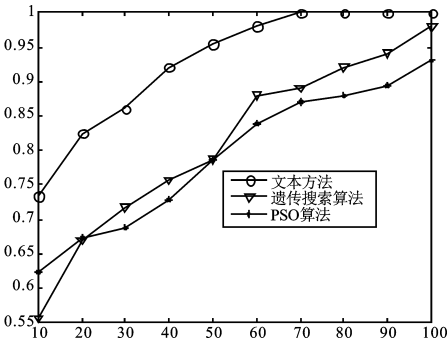


图 4 网络用户相似性特征搜索的准确率对比

分析图 4 得知,随着搜索迭代步数的增大,对网

络用户特征搜索的准确率不断体改,本文方法在搜索 70 步后对网络用户相似性特征搜索的准确率达到 100%,特征挖掘能力比传统方法平均提升 13.86%,计算开销缩短 21.75%,算法的全局收敛性较好,有效提高了网络信息的优化推荐能力。

5 结语

为了提高网络用户的 QoS 个性化服务能力,对网络用户的个性化相似性特征进行优化挖掘和自动搜索,推荐能够反映用户感兴趣的网络信息,提高网络信息的处理和利用能力,本文提出一种基于动态聚类分析的网络用户相似性特征搜索算法。采用关联信息标注方法进行网络用户相似性行为特征信息检索和动态聚类分析,实现用户相似信息特征优化搜索。采用本文方法进行网络用户相似性特征搜索的准确性较高,对全局收敛性较好,对用户的偏好信息准确挖掘能力较强,在网络信息的优化推荐中具有很好的应用价值。

参考文献:

- [1] 庞俊,于戈,许嘉,等.基于 MapReduce 框架的海量数据相似性连接研究进展[J]. 计算机科学,2015,42(1):1-5.
- [2] 张景祥,王士同,邓赵红,等.融合异构特征的子空间迁移学习算法[J]. 自动化学报,2014,40(2):236-246.
- [3] GOFERMAN S, ZELNIKMANOR L, TAL A. Context-aware saliency detection[J]. IEEE Transactions

on Pattern Analysis and Machine Intelligence, 2012, 34(10):1915-1926.

- [4] CHENG M M, MITRA N J, HUANG X, et al. Global contrast based salient region detection [J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2015, 37(3):569-582.
- [5] 毕安琪,王士同.基于 Kullback-Leiber 距离的迁移仿射聚类算法[J]. 电子与信息学报,2016,38(8):2076-2084.
- [6] SUN L and GUO C H. Incremental affinity propagation clustering based on message passing [J]. IEEE Transactions on Knowledge and Data Engineering, 2014, 26(11):2731-2744.
- [7] 毕安琪,董爱美,王士同.基于概率和代表点的数据流动态聚类算法[J]. 计算机研究与发展,2016,53(5):1029-1042.
- [8] 郭磊,马军,陈竹敏,等.一种结合推荐对象间关联关系的社会化推荐算法[J]. 计算机学报,2014,37(1):219-228.
- [9] 张卫,曹先彬,尹洪章.基于多特征融合的聊天室社会网络挖掘方法[J]. 中国科学技术大学学报,2009,39(5):540-546.
- [10] 孟祥武,刘树栋,张玉洁,等.社会化推荐系统研究[J]. 软件学报,2015,26(6):1356-1372.

作者简介:

马杰男,(1977-),硕士,实验师.研究方向为计算机技术应用、网络与信息处理、信息化教学设计.

E-mail:majie@nau.edu.cn.

(上接第 112 页)

- [7] Wang Gang, Li Wei, Li Ping, et al. A novel antifuse structure based on amorphous bismuth zinc niobate thin films [J]. Journal of Semiconductor, 2012, 33(8):44-47.
- [8] 罗刚,曾献君.超大规模集成电路的时钟偏斜及其补偿技术[J]. 制造业自动化,2011,33(3):115-117.
- [9] 李芝燕,严晓浪.高速多级时钟网布线[J]. 半导体学报,2000,21(3):290-297.
- [10] 童琼,张晓林,苏琳琳,等.双频双系统导航芯片的时钟树分析和设计[J]. 微电子学,2011,41(2):246-250.
- [11] 赵萌,蔡懿慈,洪先龙,等.合理偏差驱动的时钟线网构造及优化[J]. 半导体学报,2003,24(4):438-444.
- [12] 马游春,曾青林,李锦明.基于 ACTEL 反熔丝 FPGA

的分频器优化设计[J]. 计算机测量与控制,2012,20(9):229-231.

作者简介:

杜涛男,(1976-),博士,讲师.研究方向为大规模集成电路设计与验证. E-mail:dutao@ussc.edu.cn.

许百川男,(1991-),硕士研究生.研究方向为大规模集成电路设计.

李威男,(1959-),教授,博士生导师.研究方向为可编程逻辑器件、集成电路可靠性技术.

晁醒男,(1993-),硕士研究生.研究方向为大规模集成电路设计.

吴方明男,(1991-),硕士研究生.研究方向为大规模集成电路设计.