优先关联的 Web 日志数据逼真生成算法①

丘志鹏, 肖如良, 张 锐

(福建师范大学 软件学院, 福州 350117)
(福建省公共服务大数据挖掘与应用工程研究中心, 福州 350117)

摘 要: 字段关联的构建方法是 Web 数据逼真生成中的困难问题. 提出一种基于 MIC 的字段优先关联的 Web 数据逼真生成算法. 该算法与现有的方法完全不同: 首先, 提取真实 Web 日志数据集中相应字段间的 MIC 系数; 然后, 结合字段的重尾特性, 采用 SE 分布对字段的重尾性进行建模; 最后, 建立字段关联模型, 模拟出真实数据集中的字段间依赖性, 从而逼真生成目标数据集. 实验表明, 生成的数据集能够保持合理的字段间的均衡性以及节点间的相似性.

关键词: 字段关联; 数据生成; MIC 系数; 重尾

Simulate Generating Web Log Algorithm Using Fields' Priority Relevance

QIU Zhi-Peng, XIAO Ru-Liang, ZHANG Rui

(Faculty of Software, Fujian Normal University, Fuzhou 350117, China)
(Fujian Provincial Engineering Research Center of Public Service Big Data Analysis and Application, Fuzhou 350117, China)

Abstract: The construction method of field relevance is a difficult problem in the Web data generation. A new algorithm for fields' priority relevance based on maximal information coefficient is proposed. The algorithm is completely different from the existing method. Firstly, the maximal information coefficient between the appropriate fields needs to be extracted from real Web log data. Then, combined with the field of heavy tailed characteristics, the field is modeled by stretched exponential distribution. Finally, real data's field dependence is simulated by the fields' relevance model, so as to generate a realistic target data set. The experiments show that the generated data sets can maintain a reasonable balance between the fields and the similarity between the nodes.

Key words: fields' relevance; data generation; maximal information coefficient; heavy tail

合理分析 Web 日志数据的字段内容, 有助于对其领域系统的构建及测试, 然而 Web 日志数据通常达到 TB 甚至 PB 级别, 极其耗费网络资源, 并且数据中用户行为及相关物品属性等相关字段内容涉及隐私信息, 因此, 企业及政府等机构极少愿意分享其数据供研究人员使用. 随着互联网规模的不断扩大, Web 日志数据中重尾现象也越发普遍, 各个字段间的关联变得愈加复杂, 生成具有真实数据特性的数据集极具难度. 因此, 构建一个可模拟出真实字段间关联关系的数据生成算法成为众多科研工作中模拟数据来源的基础, 也是本文研究的重点.

现有的数据生成算法的研究主要分为时间字段相关性质的研究与非时间字段相关性质的研究两个方面. 前者主要应用于网络流量预测、时序分析等方面, 现已较为成熟, 有相应的商用与科研软件供研究人员使用, 如 OPNET; 而后者主要在于对字段分布特性的数学建模及字段间关联研究, 主要应用于特定的研究项目中, 需要根据不同业务场景进行逼真生成, 复杂度高, 主要代表性工作有加拿大萨斯喀彻温大学 Busari 提出的 proWGen[1]数据生成器, 通过分析 Web 用户行为字段值分布情况, 用 Zipf-like 分布刻画字段重尾性[2]进行数据生成, 采用多参数的机制, 使得该生成器具

① 基金项目:福建省科技计划重大项目(2016H6007)
收稿时间:2016-07-04;收到修改稿时间:2016-08-08 [doi:10.15888/j.cnki.csa.005662]

有良好的扩展性, 能应用于 Web 服务器的压力测试及缓存性能研究. 缺点在于: proWGen 对字段关联仅采用简单的正/负相关的方式实现, 难以逼真生成实际中复杂多样的数据.

随着互联网数据量的爆炸式增加, Zipf-like 已经不再适用于描述具有重尾特性的 Web 数据分布, 文献[3]指出采用 SE 分布描述 Web 数据的重尾性更加合理. 若采用 Zipf-like 进行数据生成, 对于生成数据所应用的系统而言, 其测试性能评估上会存在高估的结果, 与真实数据情况对比有较大的误差, 意味着生成了不可靠的数据. 所以目前非时间字段相关性质研究仍处于成长阶段. 本文主要以非时间字段相关性质研究作为主要工作.

针对以上问题, 本文提出了一种基于 MIC 的字段优先关联的 Web 日志数据逼真生成(Simulate Generating Web Log algorithm using fields' priority relevance based on maximal information coefficient, SGWL)算法. 该算法与现有的方法完全不同, 通过利用 Web 日志数据特征进行参数提取, 采用 SE 分布代替 Zipf-like 分布对字段重尾性进行刻画. 然后对数据字段关联提出一种全新的模型(基于 MIC 的字段优先关联模型)代替传统的正/负相关模型, 进行指导关联. 通过该算法生成的数据, 不仅在整体上能拟合一个逼真的分布趋势, 在局部上也能够准确刻画字段重尾性并保持合理的字段间的均衡性以及节点间的相似性, 可应用于 Web 数据驱动的软件过程.

1 相关工作

目前, 国内外已有大量的仿真数据生成研究. 按其是否与时间因素相关可分为二个类别: 其一, 与非时间字段性质相关的研究; 其二, 与时间字段性质相关的研究.

(1) 非时间字段相关性类型的研究主要涉及非时间相关字段的建模及字段间关联研究, 例如字段值出现次数分布建模、重尾性刻画等. 通过已有 Web 数据作为驱动, 对其进行数学建模, 从而来模拟生成新的数据. 中科院计算所詹剑锋研发的可扩展大数据生成器 BDGS[4]在生成量、速率、多样性、真实性这四个角度进行仿真数据, 能够自定义生成结构化、半结构化、非结构化数据并能保持数据的重尾性, 但是其缺陷在于: 字段关联模型单一、缺乏物理意义; 新加坡国立大学 Tay[5]通过研究照片评论数据, 指出字段关联关系在数据生成中的重要性. 在 Tay 的研究工作中定义了五种数据类型, 实际上这种做法存在一定的局限性, 五种类型不足以囊括复杂多样的真实数据; proWGen 数据生成器采用多参数可调机制, 运用数学模型建模可生成具有 Web 访问特征的数据. 该方法具有较大的灵活性, 可以较为逼真的模拟单列 Web 数据字段, 但是在多字段数据生成中仅仅采用正/负相关的方式进行仿真, 不足以描述 Web 数据中不同字段属性的复杂关系; 加拿大多伦多大学 Rabl 设计的 PDGF[6]数据生成器, 目前是 TPC-DI(数据集成评测系统)的专用数据生成器, 已经被大数据测试基准 BigBench 广泛使用, 但是该生成器只供特定数据进行生成, 其扩展性较为一般; 工业界现有的数据生成器 Red Gate[7]、DTM[8], 可高速生成与真实业务数据相似的数据, 然而, 工业界的数据生成主要依托于相关的业务, 由于这类生成器具有通用性, 也意味着无法根据真实数据的特性随意修改生成字段间的关联.

(2) 时间字段相关性类型的研究, 需要为时间属性字段建模, 通过模拟时间相关属性特征(如网络流量自相似性、长相关性、多分形性)来生成 Web 数据. 其中包括以浙江大学尹建伟研发的 BURSE[9]为代表的工作负载数据生成器, 重点模拟数据的周期性、突发性特征来实现 Web 数据的自相似性; 法国凡尔赛大学 Laurent[10]主要通过研究天气数据的时间序列, 在不同时间尺度上依赖于多分形理论进行相应数据仿真生成; 美国新泽西理工学院 Ansari[11]研究了基于 FARIMA 的 MPEG 视频流量建模问题,采用 FARIMA 过程作为自相似流量产生器, 对 MPEG 中的 I、P 和 B 帧的自相关结构进行建模, 从而完成数据生成. 此外, 时间字段相关性质研究领域也具备较为成熟的产品用于数据生成, 加拿大西蒙菲沙大学 Michael[12]收集了蜂窝数字包数据网络 (CDPD)中的业务数据并对运用工具 OPNET 建模和仿真分析. 以上的这些时间字段相关性质研究成果, 均有强力的学科理论、技术模型支撑, 其涉及到自相似网络业务 ON/OFF 模型、时间序列分析 FARIMA 模型等, 并且也有较为成熟的产品供研究人员使用, 商用软件有 OPNET, BONeS 和 COMNET III, 科研用软件有 NS2 和 SSF NET. 而非时间字段相关性质研究的数据生成方法, 目前并没有一个通用的商用软件供研究者使用, 并且现有的数据生成器中依然局

限于简单的数据分布与粗糙的字段关联, 没有一个合适且较为完备的模型来指导非时间字段相关性质研究数据生成中字段关联的问题.

综上所述, 数据生成器的时间字段相关性质研究已趋于成熟, 而非时间字段相关性质研究中仍存在许多需要急于解决的困难问题. 本文重点对数据生成的非时间字段相关性质研究进行相关改进工作. 通过运用 SE 分布来对具有重尾现象的 Web 字段值出现次数进行刻画, 在所需关联的字段间用 MIC 系数作为关联度的描述, 建立全新的关联模型, 进而使生成的数据更具有可靠性, 从而达到逼真生成的目的.

2　理论基础

2.1 重尾数据的分布

2.1.1 Zipf-like 分布

大数据背景下, Web 日志数据中部分字段分布呈现出幂律分布的特性, 也就是人们常说的长尾现象, 本文中统一称为重尾性. Zipf-like 分布又称为类齐普夫分布, 通常用于描述具有重尾性质字段的分布, 本节图示以 Movielens-1m 数据集为例, 以排名位序值(Rank)i 作为 X 轴, 以出现次数(Times)t_i 作为 Y 轴, 如图 1 所示 userID 字段值出现次数(又称为用户活跃度)表现出重尾性, 在传统方法中通常使用 Zipf-like 分布来对其进行刻画.

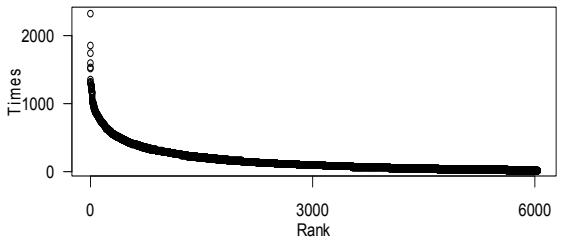

图 1　用户活跃度分布情况

假设一个数据集 D 中某字段 A 服从参数为 k 的 Zipf-like 分布, 那么对其字段值所出现的次数统计进行降序排列, 序列第 i 的字段 A_i, 其出现的次数 t_i 满足式(1):

$$t_i = \frac{l}{i^k z} \tag{1}$$

其中 l 为数据集的总记录数, 参数 z 的表达如式(2)所示:

$$z = \sum_{i=1}^{N} \frac{1}{i^k} \tag{2}$$

若数据集 D 中某字段的所有值出现次数服从 Zipf-like 分布, 那么根据对象出现次数降序排列, 在坐标系中, 以排名位序值(Rank)i 作为 X 轴, 以出现次数(Times)t_i 作为 Y 轴, 分别对 X 轴、Y 轴上的对应所有数据进行取自然对数处理, 那么应当呈现出一条直线. 如图 2 可发现, 用户活跃度在双对数坐标系下并非呈现一条直线, 说明用户活跃度并不服从 Zipf-like 分布.

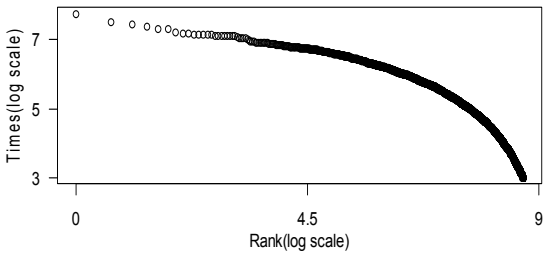

图 2　双对数坐标系下用户活跃度分布情况

2.2.2 SE 分布

SE 分布(Stretched Exponential Distribution), 中文全称为广延指数分布, 最早由 Kohlrausch 于 1847 年研究发现, 适用于描述不同复杂系统的动态衰减现象, 其中包括自然、经济、互联网等领域. 美国俄亥俄州立大学张晓东[3]对不同 Web 系统的用户行为日志数据进行分析, 发现 Zipf-like 分布不适合描述 Web 日志行为数据的重尾性, 而 SE 分布能对其进行很好的刻画. 说明该分布适用于描述幂律模型无法准确刻画的情况.

式(3)表示 SE 分布的概率密度函数:

$$p(x) = c\frac{x^{c-1}}{x_0^c} e^{-(\frac{x}{x_0})^c} \tag{3}$$

累计分布函数如式(4)所示:

$$p(x) = e^{-(\frac{x}{x_0})^c} \tag{4}$$

其中 c 为广延参数, 其参数范围在(0, 1), x_0 为尺度参数.

为了方便描述, 我们约定将 X 轴上的对应所有数据进行取自然对数处理, Y 轴上的对应所有数据进行取原值的 c 次幂处理, 这样得到的坐标系称为 SE 坐标系. 若数据集 D 中某字段的所有值出现次数服从 SE 分布, 那么根据对象出现次数降序排列, 在坐标系中, 以位序值 i 作为 X 轴, 以出现次数 t_i 作为 Y 轴, 再将 X、

Y 的值转化置 SE 坐标系中, 那么应当呈现出一条直线. 如图 3, 可以清楚的看出用户活跃度在 SE 坐标系下呈现一条近似直线, 说明用户活跃度服从 SE 分布.

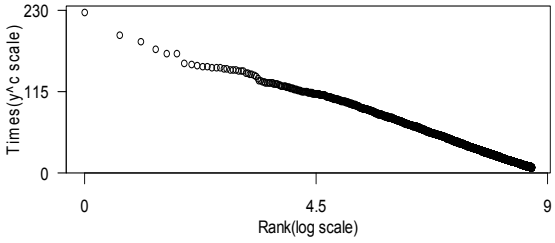

图 3 SE 坐标系下用户活跃度分布情况

采用公式(5)对该直线进行描述:

$$y_i^c = -a\log i + b(1 \le i \le N) \tag{5}$$

其中 a= x_0^c , b = y_1^c , 又因为 c 为经验常量, 因此通过最小二乘法可以拟合出 a,b 的值, 从而求得 x_0, 代入式(3)中, 计算得到完整的累计概率分布函数, 完成 SE 分布的建模.

2.2 字段关联性度量

记录是由若干个字段组合而成, 而字段间必然存在着某种关联. 为了能准确量化描述两个字段间的关联性, 研究者们提出了 pearson 系数、spearman 系数、核密度估计(KDE)、互信息等度量标准. 这些度量方法复杂、不适用非线性数据, 缺乏普适性、健壮性低等问题, 难以适用于数据生成算法中. 为此本文采用 MIC(The Maximal Information Coefficient)系数作为字段关联性度量.

2011 年, Reshef[13]在 Science 首次提出 MIC 系数, 中文又称为最大信息系数. 该系数是在互信息的基础上衍化而来, 能对不同类型的关联关系进行评估, 其范围为[0,1], 且具有对称性、良好的普适性和公平性. 如果变量 X 与 Y 独立, 则 MIC(X,Y)=0; 如果 X 与 Y 之间具有确定的关系, 则 MIC(X,Y)=1, 此时不存在任何噪声影响.

计算方法主要是通过对变量对(X,Y)中所有样本点的构成的散点图进行划分, 利用动态规划的方式计算并搜索不同划分方式下所能达到的最大互信息值. 最后, 对最大互信息值进行标准化处理, 所得结果即为 MIC, 记作 e_{MIC}. 记 D'为给定数据集, x 和 y 分别表示在 X 和 Y 变量轴上的划分份数, n 为变量对(X,Y)的样本容量, G 表示某种划分. 因此在划分 G 下等($x \times y$)轴划分的最大互信息为式(6):

$$I^*(D',x,y) = \max I(D'|_G) \tag{6}$$

标准化处理得到的特征矩阵如式(7)所示:

$$M(D')_{x,y} = \frac{I^*(D',x,y)}{\log\min\{x,y\}} \tag{7}$$

最终得到的 MIC 值如式(8)所示:

$$e_{MIC} = \max_{xy<B(n)}\{M(D')_{x,y}\} \tag{8}$$

其中 $B(n)$为网格划分细度, 通常取值为 $n^{0.6}$, 以上方法步骤简称 MINE 方法.

由式(8)可以发现, MIC 随着网格划分细度的变化而变化, 当样本容量越大的时候估计值也越准确, 这适用于当前大数据的时代背景. 表 1 列出四种相关系数的应用对比, 由表 1 可知 MIC 系数具有适用范围广、计算复杂度低, 鲁棒性高, 标准化结构特性. 因此, 本文算法采用 MIC 作为字段关联度参考.

表 1 四种相关系数优劣对比

相关系数	应用情况		鲁棒性	计算复杂度	可标准化
	线性	非线性			
Pearson	√	×	L	L	Y
Spearman	√	×	M	L	Y
KDE	√	√	H	H	N
MIC	√	√	H	L	Y

注: L—低; M—中; H—高; Y—是; N—否.

3 基于MIC的字段优先关联模型

假设要生成由两列字段组成, 共计 l 条记录的日志数据集, 其中字段名分别用 A、B 表示. 令字母 S 表示为集合, 那么字段 A 对应的值所在集合 S_A={A_1,A_2,A_3…A_m}, 共有 m 种取值; 字段 B 对应的值所在集合 S_B={B_1,B_2,B_3…B_n}, 共有 n 种取值. 每条记录的形式为{A_x,B_y}(1≤x≤ m,1≤ y≤ n). 令字母 t 代表次数, 则字段 A 值 A_m 出现的次数为 t_{A_m}次,字段 A 中所有值分别出现次数构成集合 S_{t_A}, 字段 B 中值 B_n 出现的次数为 t_{B_n}次, 字段 B 中所有值分别出现次数构成集合 S_{t_B}, 且满足式(9)表示字段 A 所有值出现次数累加和等于字段 B 所有值出现次数累加和等于日志数据集总记录数 l.

$$\sum_{i=1}^{m} t_{A_i} = \sum_{i=1}^{n} t_{B_i} = l \tag{9}$$

对于数据生成而言, 首先分别对字段的所有值出现次数的集合进行建模, 根据章节 2.1 的方法, 得到出现次数降序排列的集合 S_{t_A} 与 S_{t_B}. 然后累积分布函数

$p(x)$, 其中 x 表示字段值出现次数的排名位序. 以 A 字段为例, 累积分布函数具体如式(10), 到该步便完成字段建模的步骤.

$$p^A(x)=\sum_{i=1}^{x}t_{A_i}\Big/\sum_{i=1}^{m}t_{A_i}\ (1\le x\le m) \tag{10}$$

记录是由字段组合而成, 在完成字段建模之后, 需要将两个字段进行关联操作, 进而形成一条完整的记录. 关联操作即为取集合 S_A 与 S_B 笛卡尔积的一个元素的过程. 假定符号 ξ 表示(0,1)上均匀分布的随机数, 字母 r 表示关联取值数, 则在生成一条记录时, 首先生成随机数 ξ_A, 令 $\xi_A=p^A(x)$, 通过式(10)的逆函数解析式 $p_{-1}^{A}(x)$,计算可得唯一的实数位序 x, 根据位序与字段值映射关系, 求得字段值 A_x. 然后, 根据 AB 字段间的相关性, 通过关联模型计算得到 r_{AB}, 令 $r_{AB}=p^B(y)$, 同理可得字段值 B_y, 即得到记录$\{A_x,B_y\}$.

关联过程存在三种情况, 分别为正相关、负相关与零相关, 其中正相关表示自变量增长, 因变量也跟着增长; 负相关表示自变量增长, 因变量反而减少; 因变量的增减与自变量的增减无关, 相互独立. 现阶段数据生成算法中主要使用关联模型分为正相关模型与负相关模型, 其中正相关模型为 $r_{AB}=\xi_A$, 负相关模型为 $r_{AB}=1-\xi_A$, 该模型的不足之处在于关联度量简单, 不具备的物理意义, 且未考虑字段间零相关情况. 因此, 本文提出一种基于 MIC 的字段优先模型 PRF(the Priority Relevance of Field based on maximal information coefficient, PRF). 令 r'_{AB} 表示经过 PRF 模型得到的关联取值数, 且 r'_{AB} 由优先关联部分与独立部分组合而成. 正相关 PRF 模型如式(11)所示:

$$r'_{AB}=e_{MIC}\sum_{i=1}^{g}t_{A_i}\Big/\sum_{i=1}^{m}t_{A_i}+(1-e_{MIC})\,h/n \tag{11}$$

负相关 PRF 模型如式(12)所示:

$$r'_{AB}=1-[e_{MIC}\sum_{i=1}^{g}t_{A_i}\Big/\sum_{i=1}^{m}t_{A_i}+(1-e_{MIC})\,h/n] \tag{12}$$

其中 $g\in[1,m]$, $h\in[1,n]$, 参数 $e_{MIC}\in[0,1]$为字段 A 与字段 B 之间的 MIC 系数, 用于衡量字段间相关程度, 在模型中的物理意义表示优先关联部分所占比例. $\sum_{i=1}^{g}t_{A_i}\Big/\sum_{i=1}^{m}t_{A_i}$ 表示随机字段值 A_a 出现次数的累积分布概率 $p(x)$. h/n 表示在 B 字段中 n 个取值内, 随机选取第 h 个值作为字段值的概率. 令 $\xi_A=\sum_{i=1}^{g}t_{A_i}\Big/\sum_{i=1}^{m}t_{A_i}$, $\xi_B=h/n$ 分别带入式(11)、式(12), 化简得到式(13)、式(14).

$$r'_{AB}=e_{MIC}\xi_A+(1-e_{MIC})\xi_B \tag{13}$$

$$r'_{AB}=1-[e_{MIC}\xi_A+(1-e_{MIC})\xi_B] \tag{14}$$

若字段间存在关联, 模型优先采用 ξ_A 对字段 B 进行关联取值, 若字段间相互独立, 则重新生成随机数 ξ_B, 进行关联取值. 当 $e_{MIC}\to 1$时, 说明字段 A 与字段 B 存在线性相关关系, 表示每个字段 A 的值都关联着在各自累积分布函数 $p(x)$下相同累积概率的字段 B 的值, 以正相关模型为例, PRF 模型转化为 $r'_{AB}=\xi_A$. 当 $e_{MIC}\to 0$ 时, 说明字段 A 与字段 B 相互独立, 表示每个字段 A 的值都与字段 B 的值不存在关联, 呈现随机关系. 以正相关模型为例, PRF 模型转化为 $r'_{AB}=\xi_B$. 当 $e_{MIC}\in(0,1)$ 时, 优先关联部分所占比例为 e_{MIC}, 独立部分所占比例为(1-e_{MIC}), 通过两部分的和, 根据式(11)计算得出 r'_{AB}, 以 r'_{AB} 作为字段 B 中某值的累积概率, 从而可以求出字段 B 的值, 最终完成一次字段 A 与字段 B 的关联.

PRF 模型具有一般性与明确的物理意义, 以 MIC 系数作为主要参考, 能合理的描述数据间的关联情况, 适用于大部分数据生成算法中的字段关联步骤.

4　基于PRF的Web日志数据生成算法SGWL

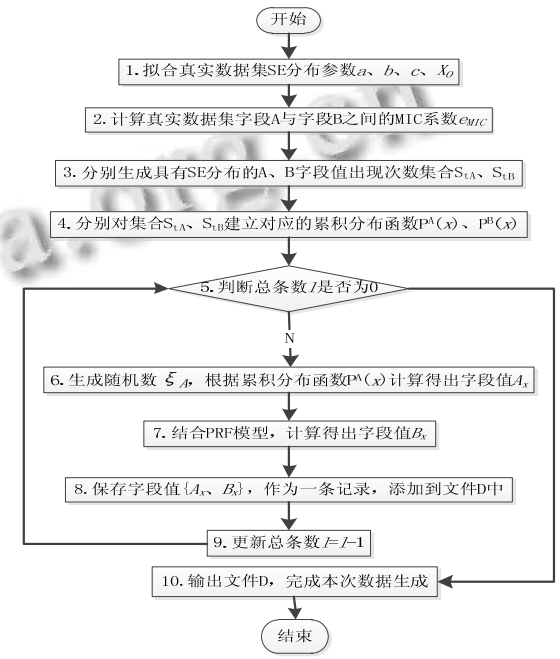

图 4　基于 PRF 模型的 Web 日志数据生成算法 SGWL

本文提出一种基于 PRF 的 Web 日志数据逼真生成算法 SGWL. 该算法通过提取真实数据集的相关参数,

利用 SE 分布模拟具有重尾性质的字段值出现次数集合, 在数据生成过程中根据 PRF 模型完成字段关联, 每次生成完一条记录之后对总条数进行更新, 从而达到控制生成记录总量的目的. 算法描述如图 4 所示.

在图 4 SGWL 算法流程中, 步骤 1 至步骤 2 为 Web 日志数据字段特征提取过程, 步骤 3 至步骤 4 表示对字段进行建模, 步骤 6 至步骤 8 为生成一条完整记录的过程, 其中步骤 7 表示字段关联.

5 实验结果与分析

5.1 实验数据集介绍

在生成 Web 日志数据结束之后需要测评仿真数据集的可靠度, 采用真实数据集作为参照比对. 实验采用四个不同领域具有代表性的数据集进行实验分析, 旨在验证 SGWL 算法的一般性, 其分别是 Movielens-1M 电影评分数据集、NASA 网络请求数据集、Epinions 社会网络数据集和 Xiami 音乐用户行为数据集. 其中 MovieLens 1M 为 6040 个用户对 3952 个电影产生的 1000209 条评分记录; NASA 为 54770 个请求节点对 8937 个路径产生的 1048576 服务日志数据记录; Epinions 为 40163 个用户对 139738 个物品产生的 664823 条评分记录; Xiami 为 162273 个用户对 8377 首歌曲产生的 11098957 条行为记录, 其统计结果如表 2 所示.

表 2 四个数据集基本统计结果

数据集	类别	用户数量	对应类别数量	请求数
Movielens-1m	Movie	6040	3952	1000209
NASA	Traces	54770	8937	1048576
Epinions	Reviews	40163	139738	664823
Xiami	Music	162273	8377	11098957

5.2 评估指标

5.2.1 字段均衡性指标: 基尼系数

基尼系数(Gini Coefficient)[13]是意大利经济学家基尼于 1992 年提出, 定量测定收入分配差异程度. 基尼系数是比例数值, 在 0 和 1 之间, 是国际上用来综合考察居民内部收入分配差异状况的一个重要分析指标. 假定一定数量的人口按收入由低到高排序, 分为人数相等的 m 组, 从第 1 组到第 i 组人口累计收入占全部人口总收入的比重为 Wi 其计算方法如式(15)所示. 按照联合国有关组织规定: 0.2 表示绝对平均, 0.3-0.4 表示相对合理, 0.5 以上表示严重不均衡. 而如今, 基尼系数也可以用来测度各种意义下的资源分配均衡度. 正因为数据生成的时候需要对字段值出现次数进行建模, 同理基尼系数也适用于评估字段值出现次数的均衡性, 可以通过式(15)计算 Gini 系数.

$$Gini = 1 - \frac{1}{m}\left(2\sum_{i=1}^{m-1} W_i + 1\right) \tag{15}$$

5.2.1 节点相似性指标: PA 指数、AA 指数

若用二分网络结构来描述数据集, 那么字段上的值即对应为网络中的节点这一概念. 节点相似性指标[14]在链路预测、节点聚类、个性化推荐方面应用都很广泛. 电子科技大学周涛[14]罗列了十五种相似性指标, 本文采用其中两种稳定性较好的指标作为实验评判标准, 分别是 PA 指数与 AA 指数. 令 $S_{\alpha\beta}$ 表示节点相似性度量, α 与 β 分别表示字段值 B_α 与 B_β, u 表示字段值 A_u, $U_{\alpha\beta}$ 表示字段 A 中既关联了 α 又关联了 β 的字段值的集合, t_α 表示 α 出现的次数, t_β 表示 β 出现的次数, t_u 表示 u 出现的次数.

PA(Preferential Attachment)指数计算如式(16):

$$S_{\alpha\beta} = t_\alpha t_\beta \tag{16}$$

AA(Adamic-Adar)指数计算方法如式(17)所示:

$$S_{\alpha\beta} = \sum_{u \in U_{\alpha\beta}} \frac{1}{\log t_u} \tag{17}$$

5.3 实验结果及分析

本节主要生成从生成数据的整体分布拟合刻画、局部上字段均衡性、个体中节点相似性三个层面与真实数据进行对比评估, 在整体和局部层面与传统方法进行纵向对比, 在个体评估层面上对 SGWL 算法采用两种指数实验评估横向对比.

首先, 就整体层面而言, 本节选用 Movielens-1m 数据集与 Epinions 数据集作为真实数据集参考, 对具有重尾性质的字段 User 值出现次数分别进行 Zipf-like 分布刻画与 SE 分布刻画, 然后选取合适分布对字段进行拟合, 并计算出拟合函数与真实数据集的拟合优度 R^2 评估拟合效果, 其实验结果如图 5、图 6 所示, 其中点线为双对数坐标系下真实数据分布刻画, 虚线为 SE 坐标系下真实数据分布刻画, 实线为拟合直线.

由图 5 与图 6, 可以看出两个数据集字段 User 值出现次数分布在双对数坐标系下均呈现出“胖头瘦尾“的曲线形状, 而在 SE 坐标系下均呈现出一条近似直线的情况, 因此根据章节 2.1 所述, 验证了 SE 分布在描述重尾特征的数据字段上优于传统的 Zipf-like 分布.

然后对虚线进行拟合, 通过R语言中的nls方法计算得到 a、b 值, 然后在图上绘制对应直线, 计算出虚线与实线之间的拟合优度 R^2. 图 5 中, R^2=0.9748, 图 6 中, R^2=0.9719, 均接近于 1, 说明回归直线对真实数据的拟合程度很高. 总体而言, 采用 SE 分布的 SGWL 在重尾性刻画上描述要优于 proWgen, 生成的数据与真实数据集更为接近, 能更准确的描述真实数据集的重尾性, 从整体上把握数据的逼真生成.

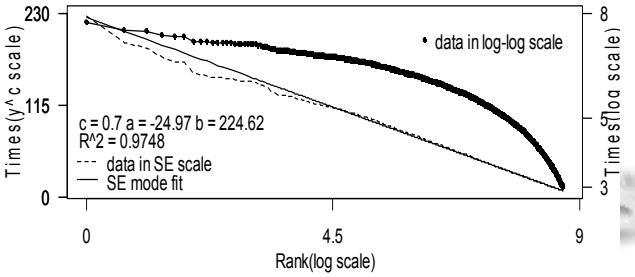

图 5　Movielens-1m 字段 User 整体分布拟合刻画图

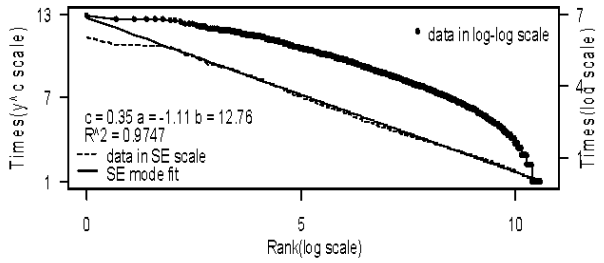

图 6　Epinions 字段 User 整体分布拟合刻画图

在局部层面上, 基尼系数是研究字段均衡性的一个重要特征, 选用四个不同领域的真实数据集的某一字段分别采用SGWL算法与proWgen算法进行数据仿真, 最终与真实数据集通过计算其基尼系数进行字段均衡性对比分析. 实验结果如表 3 所示.

表 3　真实数据集与生成数据集字段基尼系数(Gini)对比

数据集	字段	Real Gini	Gini by SGWL	Gini by proWgen
Movielens-1m	t_{UserID}	0.53	0.52	0.43
NASA	t_{host}	0.95	0.93	0.79
Epinions	t_{UserID}	0.68	0.67	0.58
Xiami	t_{SongID}	0.91	0.89	0.77

根据表中数据可以直观的看到列“Gini by SGWL”的每一个数值都明显逼近于列“Real Gini”的值, 进一步通过数据计算可以得到SGWL生成数据的基尼系数与真实数据集的平均误差为 1.5%, 而 proWgen 生成数据的基尼系数与真实数据集的平均误差却达到了 11%, 由此说明, SGWL 算法生成的数据在字段均衡性上要优于 proWgen, 且适用于不同领域背景下的数据生成, 具有一般性.

最后在个体评估层面上对节点间相似性进行实验分析. 以 Movielens-1m 数据集作为真实数据集参考, 根据 5.2.2 介绍的方法, 令字段”UserID”代表字段 A, 字段”MovieID”代表字段 B, 在字段 B 中随机选取 10000 对节点$\{\alpha, \beta\}$, 依次分别在真实数据集与 SGWL 算法生成数据集中计算对应的相似性度量 $S_{\alpha\beta}$, 令真实数据集中所有 $S_{\alpha\beta}$ 组成的序列为 $S^1_{\alpha\beta}$, SGWL 算法生成数据集中所有 $S_{\alpha\beta}$ 组成的序列为 $S^2_{\alpha\beta}$. 实验需在坐标轴上绘制 10000 个散点, 其中以 $S^1_{\alpha\beta}$ 上所有的 10000 个值归一化后依次作为散点的 X 坐标, 以 $S^2_{\alpha\beta}$ 上所有的 10000 个值归一化后依次作为散点的 Y 坐标. 若两个数据集具有相同的节点相似性, 那么散点将全部散落在倾斜度为 45 度的实线 y=x 上, 偏离斜线越远则代表两个数据集的节点间相似性差异越大, 从而说明算法生成的数据越不可靠. 实验结果如图 7、图 8 所示.

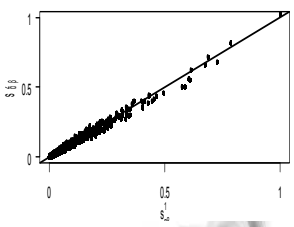

图 7 PA 下节点相似性对比

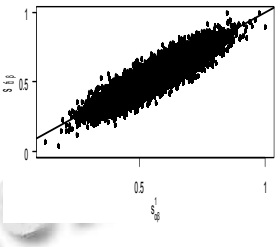

图 8 AA 下节点相似性对比

如图所示, 图 7、图 8 中大多数的散点位置均落在 y=x 这条斜线的附近, 部分点甚至于斜线重合. 图 7 中真实数据 PA 指数与 SGWL 算法生成数据 PA 指数误差为 0.18%, 图 8 中真实数据 AA 指数与 SGWL 算法生成数据 AA 指数误差为 6%, 因此图 7 散点分布较图 8 更为稠密. 从而说明生成的数据能较好的保持真实数据集中节点间的相似性, 表明 SGWL 算法生成的数据具有一定的可靠性. 在图中 7 中 99.8%节点对 PA 指数集中于(0,0.4)这个区间内, 这种情况的产生源于数据集中 User 字段的重尾性, 由式(16)可以看到, PA 指数依赖于节点对值出现次数的乘积, 因此重尾性导致该乘积值普遍较小, 从而使得散点集中落在 X 坐标上(0,0.4)这个区间内. 这也进一步说明了 SGWL 算法能逼真刻画字段的重尾性.

6 结论

合理的字段关联是 Web 日志数据生成算法中的关键. 本文提出了基于 MIC 系数的字段优先关联的 Web 日志数据逼真生成算法 SGWL, 该方法以 SE 分布代替 Zipf-like 分布来模拟 Web 数据的重尾性, 并提出一个全新且物理意义明确的字段关联模型 PRF, 指导字段关联. SGWL 算法可保证生成的数据集具有同真实数据集一致的字段间关联和字段值的分布, 为 Web 数据驱动的软件研发, 提供了可靠的逼真数据生成.

参考文献

1 Busari M, Williamson C. ProWGen: A synthetic workload generation tool for simulation evaluation of web proxy caches. Computer Networks, 2002, 38(6): 779–794.

2 Sarla P, Doodipala MR, Dingari M. Self similarity analysis of web users arrival pattern at selected web centers. American Journal of Computational Mathematics, 2016, 6(1): 17–22.

3 Guo L, Tan E, Chen S, et al. The stretched exponential distribution of internet media access patterns. Twenty-Seventh ACM Symposium on Principles of Distributed Computing (PODC 2008). Toronto, Canada. August, 2008. 283–294.

4 Ming Z, Luo C, Gao W, et al. BDGS: A scalable big data generator suite in big data benchmarking. Advancing Big Data Benchmarks. Springer International Publishing, 2014: 138–154.

5 Tay YC, Dai BT, Wang DT, et al. UpSizeR: Synthetically scaling an empirical relational database. Information Systems, 2013, 38(8): 1168–1183.

6 Rabl T, Poess M, Danisch M, et al. Rapid development of data generators using meta generators in PDGF. International Workshop on Testing Database Systems. 2013. 1–6.

7 Campbell MK. SQL data generator. Sql Server Magazine, 2009.

8 Lear D, Hebbes S. Database Tools, EP1606735. 2005.

9 Yin J, Lu X, Zhao X, et al. BURSE: A bursty and self-similar workload generator for cloud computing. IEEE Trans. on Parallel & Distributed Systems, 2015, 26(3): 668–680.

10 Akrour N, Mallet C, Barthes L, et al. A rainfall simulator based on multifractal generator. EGU General Assembly Conference. EGU General Assembly Conference Abstracts. 2015.

11 Ansari N, Liu H, Shi Y Q, et al. On modeling MPEG video traffics. IEEE Trans. on Broadcasting, 2002, 48(4): 337–347.

12 Jiang M, Nikolic M, Hardy S, et al. Impact of self-similarity on wireless data network performance. IEEE ICC. IEEE. 2001. 477–481.

13 Przanowski K, Mamczarz J. Consumer finance data generator-a new approach to credit scoring technique comparison. General Information, 2012. arXiv: 1210.0057.

14 Liu JG, Lei H, Xue P, et al. Stability of similarity measurements for bipartite networks. Science Reports, 2016.