一种可指定分布的海量数据生成方法

李博嘉[1] 张仰森[1,2] 陈若愚[1,2]
(北京信息科技大学智能信息处理研究所 北京 100101)[1]
(网络文化与数字传播北京市重点实验室 北京 100101)[2]

摘 要 受到隐私保护等因素的影响,企业和政府数据公开缓慢;同时,由于网络带宽的限制,科研机构下载使用海量公开数据存在困难。现有的数据生成工具很少能在生成数据的分布形态、相关关系、准确性以及系统的可伸缩性等方面同时满足科研工作的要求。针对海量数据生成问题,提出了一种分布式数据生成模型,根据用户配置中指定的数据分布形态及相关关系,利用蓄水池抽样或随机采样算法对Web信息知识库进行采样、相关关系计算以及拼接等操作,生成数据属性符合用户配置的数据。通过在Apache Spark分布式计算引擎上进行数据生成实验,结果表明,生成数据符合指定的数据分布及相关关系要求,数据生成速度与数据规模、集群规模呈线性关系,从而证明该方法生成的数据具有较高的准确性和分布多样性,相应的系统具有较好的可伸缩性。

关键词 数据生成,蓄水池抽样,分布式计算,相关关系计算,数据分布检验
中图法分类号 TP391 **文献标识码** A **DOI** 10.11896/j.issn.1002-137X.2019.08.009

Method for Generating Massive Data with Assignable Distribution

LI Bo-jia[1] ZHANG Yang-sen[1,2] CHEN Ruo-yu[1,2]
(Institute of Intelligent Information Processing, Beijing Information Science and Technology University, Beijing 100101, China)[1]
(Beijing Key Laboratory of Internet Culture and Digital Dissemination Research, Beijing 100101, China)[2]

Abstract Affected by factors such as privacy protection, corporate and government data are slow to be exposed. At the same time, due to the influence of network bandwidth, it is difficult for scientific research institutions to download and use massive public data. It is rare that the existing data generation tools can concurrently meet the requirements of scientific research work in terms of the generation of data distribution pattern, correlation, accuracy and scalability of the system. Specific to the problem of mass data generation, this paper put forward a distributed data generation model. According to the data distribution pattern and correlative relation specified in the user's configuration, the reservoir sampling or random sampling algorithm is used for the sampling, calculation of relative relationship and splicing of the Web data knowledge base to generate the data of which the attribute accords with the user's configuration. Through the data generation test on the distributed computing engine Apache Spark, the generated data meets the specified data distribution and correlation requirements, and the data generation speed is linear with the data size and cluster size from the statistical point of view. It shows that the data generated by the proposed data method has high accuracy and diversity of distribution, and the proposed data generation system has good scalability.

Keywords Data generation, Reservoir sampling, Distributed computing, Correlation calculation, Data distribution test

1 引言

随着信息产业的飞速发展,数据的产生速度达到了空前的量级,根据业界预测,到2020年,全球数据总量将达到35.2ZB[1]。目前,数据大多集中在大企业或政府机构。对于Google、Amazon、百度、阿里巴巴等大型企业来说,因为自身业务量巨大,所以其掌握着大量的用户数据。而医疗、教育、民政、交通等与社会民生息息相关的领域的数据,则大多集中在政府机构的数据仓库中,中国92%以上的数据都是政府数据[1]。现在社会各界对开放数据的要求越来越迫切,然而数据公开的进展仍然十分缓慢。第一,数据是无形的财富,甚至是一些企业的核心竞争力所在,极少有企业开放核心数据。

到稿日期:2018-07-20 返修日期:2018-11-19 本文受国家自然科学基金(61772081),北京市教委科研计划项目(KM201711232014)资助。
李博嘉(1992-),男,硕士生,CCF学生会员,主要研究方向为大数据、人工智能,E-mail:1012139091@qq.com;张仰森(1962-),男,博士,教授,CCF会员,主要研究方向为自然语言处理、人工智能;陈若愚(1982-),男,博士,讲师,CCF会员,主要研究方向为自然语言处理、人工智能,E-mail:ruoyu-chen@foxmail.com(通信作者)。

第二，数据中往往包含用户隐私，公开数据之前需要做大量的隐私保护工作，阻碍了政府和企业的数据公开进程。第三，受限于网络带宽等因素，在互联网上传输大量数据困难重重，即使政府、企业开放了大量数据，对科研人员来说仍然需要花费大量时间来获取数据。因此，尽管从体量上来看，全社会的数据资源越来越丰富，然而用于科研目的的公开数据仍然十分匮乏。针对上述情况，本文提出了一种可指定分布的海量数据生成方法，基于 Apache Spark 分布式计算引擎，按照指定的数据分布、相关关系和数据规模生成数据，在一定程度上解决了上述问题。

2　相关研究工作

早期研究（2002－2010 年）主要以如何实现数据生成模块的并行化为主，生成数据的分布状态主要为奇夫分布（Zipfian distribution），模拟数据的重尾特性。2002 年，Busari 等[2]研究并提出了 proWGen 数据生成工具，通过奇夫分布模拟 Web 数据的重尾特性。Rabl 等[3]研发了 PDGF 数据生成工具，通过并行化数据生成器提高数据生成效率，被广泛应用于 ETL 测评。Rabl 等[4]对 PDGF 数据生成工具在云端的应用进行了研究。

Hadoop 等分布式计算引擎的兴起（2010－2016 年左右）降低了分布式计算算法的开发难度，研究重点转移到设计算法以使其生成的数据符合多种分布和相关关系。Ghazal 等[5]基于 PDGF 特别针对 Web 数据生成提出了 BigBench 数据生成模型。Huang 等[6]提出了 HiBench 工具，利用 Hadoop MapReduce 对生成数据时的系统负载进行了研究。Ming 等[7]提出了用于大数据 Benchmark 的数据生成工具 BDGS，基于大数据的“4V 原则”生成可定义结构的数据。Yin 等[8]研发了 BURSE 数据生成器，以模拟数据突发性、自相似性等特点分布式地生成 Web 数据，且能动态适应属性特征的改变。丘志鹏等[9]通过现有的研究工作提出了一种基于 MIC 的字段优先关联的 Web 数据逼真生成算法。

近两年的研究（2016－2018 年）主要是应用神经网络模拟真实数据，并扩充数据集。随着人工智能的普遍应用，出现了一些基于神经网络生成数据的新方法。赵会群等[10]提出了基于贝叶斯网络的复杂事件大数据处理系统测试数据生成方法，并将其特别地应用于在事件流中模拟感兴趣的事件模式。徐鹏等[11]提出了基于循环神经网络的模糊测试用例生成技术，应用长短记忆神经网络（LSTM）和 GRU 算法模型生成 PDF 文件输入型测试用例。王坤峰等[12]和 Goodfellow 等[13]提出了生成式对抗网络（Generative Adversarial Networks，GAN），基于大量的无标记数据无监督地学习生成器 G，其具备生成各种形态（图像、语音、语言等）数据的能力。近期，Chen 等[14]提出了通过最大化观测数据和隐变量之间的互信息来提高模型可解释性的 Info GAN 模型。

数据生成技术随着数据的新特点、数据科学研究的新需求不断发展。本文结合现有研究的优点，提出了数据生成模型，实现了分布式数据生成系统。利用蓄水池抽样算法能够满足流式数据的知识库数据抽取，并通过可配置的生成数据模块，生成指定分布形态、指定数据属性相关关系的海量数据。

3　数据生成模型

3.1　数据生成模型概述

一般地，任意的结构化数据都可以用二维数组进行逻辑表达，对于其中的任意一条数据来说，可以表示为一维数组。本节给出以下定义。

定义 1（数据属性段）

$$sec=\langle att_1,att_2,\cdots,att_k\rangle$$

其中，sec 是由多个相关数据属性组成的数组，是反映数据分布形态的最小单位。

定义 2（知识库）

$$kb_i=\langle sec_1^i,sec_2^i,\cdots,sec_n^i\rangle$$

其中，sec_j^i 为第 i 个知识库的第 j 个数据属性段，即知识库是同一类型数据属性段的集合。

定义 3（采样）　本文将采样定义为一个函数：

$$samp(kb,alg_s)\rightarrow sec'$$

其中，输入参数 alg_s 为采样算法，kb 为采样的知识库，输出结果 $sec\in kb$。即采样是从知识库中按照一定的采样算法抽取数据属性段的过程。

定义 4（相关关系计算）　定义相关关系计算为一个函数：

$$corr(sec,kb,alg_r)\rightarrow sec'$$

其中，sec 为原始数据属性段；kb 为相关的知识库；alg_r 为相关关系计算算法；$sec'=\langle sec,att_1,att_2,\cdots,att_n\rangle$ 为相关关系计算的结果，即 sec' 由原始数据属性段拼接相关关系计算算法生成的数据属性得到。

定义 5（单条生成数据）

$$Rec=\langle sec^1,sec^2,\cdots,sec^n\rangle$$

即生成数据是由 n 个数据属性段线性拼接而成。而其中：

$$sec^k=\begin{cases}samp(kb_k,alg_k)\\corr(samp(kb_k,alg_k),kb_j,alg_r)\end{cases}$$

即数据属性段，既可以通过对知识库 kb_k 应用采样算法 alg_k 得到，又可以在采样的基础上进一步应用相关关系算法 alg_r 并结合知识库 kb_j 进行相关关系计算得到。

定义 6（生成数据集）

$$DB=\langle Rec_1,Rec_2,\cdots,Rec_n\rangle$$

其中，Rec_i 为第 i 条生成数据，生成数据集是由 n 条生成数据组成的序列。

定义 7（数据生成）　数据生成是从已有或动态生成的单个或多个知识库 $kb_1,kb_2,\cdots,kb_m$ 中采样数据属性段 $sec_1,sec_2,\cdots,sec_m$，依据相关关系对部分数据属性段进行补充，并将这些数据属性段依次拼接成单条生成数据 Rec_i，再将海量单条生成数据序列化为生成数据集，最后将其交付的过程。

本文研究的数据生成模型主要分为两个部分，分别是采样和相关关系计算，具体过程如图 1 所示。数据知识库作为

生成数据的基础数据，首先根据用户配置的分布形态进行采样；再根据相关关系计算扩展数据属性段；然后依据数据序列格式将多个属性的字段进行拼接得到目标格式的数据；最后通过数据存储/分发接口模块对数据进行交付，交付到本地（数据库等）、存储系统（HDFS、HIVE 等）、消息队列（Kafka 等），同时也支持用户自行开发数据交付接口。

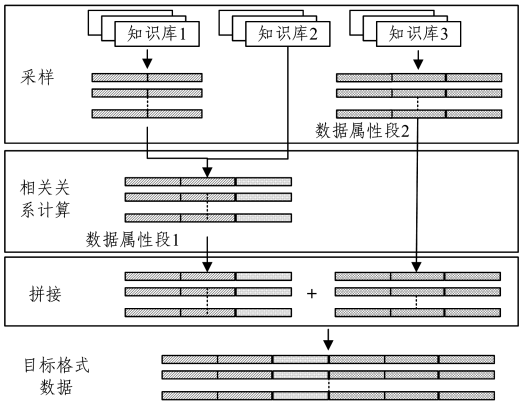

图 1 数据生成模型

Fig. 1 Data generation model

3.2 采样算法

在从数据知识库抽取生成数据时，对于规模固定的知识库进行采样一般是根据参数为 p_a 的(0-1)分布从知识库中抽取相应数据，若采样的数据量为 n_{samp}、原始知识库数据量为 n_{kb}，则参数 $p_a=n_{samp}/n_{kb}$。但如果知识库数据量 n_{kb} 无法事先确定，例如知识库来自于流式数据，则上述方法就不再适用。

此外，在生成数据规模较大时，上述方法的数据分布存在误差。设判断每次采样数据是否被抽取的随机变量为 X_i，自然地，随机变量 $X_1,X_2,\cdots,X_n$，相互独立，服从参数为 p_a 的(0-1)分布，且数学期望为 $E(X_k)=p_a$，方差为 $D(X_k)=p_a\cdot(1-p_a)$。则根据独立同分布的中心极限定理，随机变量之和 $\sum_{k=1}^{n}X_k$ 的标准化变量为：

$$Y_n=\frac{\sum_{k=1}^{n}X_k-E(\sum_{k=1}^{n}X_k)}{\sqrt{D(\sum_{k=1}^{n}X_k)}}=\frac{\sum_{k=1}^{n}X_k-n\cdot p_a}{\sqrt{n\cdot p_a\cdot(1-p_a)}}$$

$$=\frac{\frac{1}{n}\sum_{k=1}^{n}X_k-p_a}{\frac{1}{\sqrt{n}}\cdot\sqrt{p_a\cdot(1-p_a)}}=\frac{\overline{X}-p_a}{\frac{1}{\sqrt{n}}\sqrt{p_a\cdot(1-p_a)}}$$

当 n 充分大时，$\overline{X}$近似地满足 $N(p_a,\frac{1}{n}\cdot p_a\cdot(1-p_a))$，但并不满足方差 $D(X_k)=p_a\cdot(1-p_a)$。由此可见，采样数据数量是存在偏差的，这种偏差在生成少量数据时并不明显，但在生成大量数据时，将直接导致生成的数据分布出现抖动。对于这种数据知识库样本总数不断增长的情况，本文采用蓄水池抽样[15]算法，具体如算法 1 所示。

算法 1 蓄水池抽样算法

输入：$\langle n_{kb},j\rangle$，其中 $n_{kb}=[n_1,n_2,\cdots,n_i]$，为生成知识库 j 为采样数据数量，其中 $j\leqslant i$

输出：采样数据集合 RDD_{sec}

Reservoir-Sampling

$RDD_{sec}\leftarrow[n_1,n_2,\cdots,n_j]$ //从原始数据集中选取前 j 个数据属性段，保存在 RDD_{sec} 中；

While $k<i$

if rand(1)>j/k //从第 k 个数据属性段开始，以 p=j/k 的概率判断是否留下该数据属性段

$l\leftarrow$rand(j) //等概率地替换掉 RDD_{sec} 中的 j 个元素之一

$n_l\leftarrow n_k$

Return RDD_{sec}

3.3 数据分布和相关关系

数据生成模块主要是从知识库中抽取属性数据段，按特定的数据分布进行采样，根据数据相关关系计算相关数据，并最终拼接生成目标数据。数值型知识库主要应用于生成数值数据，例如时间戳、统计量等。非数值型知识库根据知识库的不同含义生成对应的非数值型数据，例如 URL、地理位置信息等。

在数据生成的过程中，根据用户指定的分布生成对应属性数据段是生成数据中的重要部分。根据不同业务的不同特点，属性数据段的分布往往表现出不同的分布形态。较为常见的分布形态包括均匀、单峰（如午间高峰）、多峰（如早晚高峰）。

在生成数据时，同一数据属性段的各个数据属性之间有些可以认为是相互独立的随机变量，有些则满足相关关系，例如有些用户对主题是娱乐的 URL 的访问量较大，有些用户对主题是新闻的 URL 的访问量较大。为了刻画数据之间的相关关系，设其中一个随机变量为 X，另一个随机变量为 Y，两者之间满足相关关系函数 $Y=f(X)$，这一相关函数可以是线性的，也可以是非线性的。

为了模拟不同的分布形态和相关关系，本文给出多分布数据生成模型。给定生成数据量，知识库 $KB=\langle kb_1,kb_2,\cdots,kb_i\rangle$，该序列的分段概率密度函数为：

$$F=\begin{cases}f_1, & n_1\\ f_2, & n_2\\ \quad\vdots\\ f_i, & n_i\end{cases}$$

根据用户需求对数据量进行分配使得每个区间上的数据量为 n_j，其中 $n_j=p_j*n$，$\sum_{1}^{i}p_k=1$。在某一区间上的生成数据，不仅来源于对应的知识库 kb_j，而且分布形态满足概率密度函数 f_j，且 $j\leqslant i$，概率密度函数 f_j 可以根据需求配置正态分布（Normal Distribution）、指数分布（Exponential Distribution）、对数正态分布（logarithmic normal Distribution）、伽马分布（Gamma Distribution）等。多分布数据生成算法如算法 2 所示。

算法 2 多分布数据生成算法

输入：$\langle n,f_{all},F,KB,Alg_s,Alg_r\rangle$，其中 $n\in\mathbb{N}$ 为生成数据量，$f_{all}=\langle p_1,p_2,\cdots,p_i\rangle$ 为数据在第 i 个区间上的概率集合，$F=\langle f_1,f_2,\cdots,f_i\rangle$ 为分段概率密度函数集合，$KB=\langle kb_1,kb_2,\cdots,kb_i\rangle$ 为知识库集

合，$Alg_s=\langle alg_s^1, alg_s^2, \cdots, alg_s^i\rangle$ 为采样算法集合，$Alg_r=\langle alg_r^1, alg_r^2, \cdots, alg_r^i\rangle$ 为相关关系计算算法集合

输出：RDD_{DB}

for(j←1;j≤i;j++)

　for(n_j←n * p_j;n_j>0;n_j−−)

　　sec←samp(kb_j,alg_s^j) //对第 j 个知识库应用采样算法 alg_s^j 得到数据属性段 sec

　　if correlation exist //如果数据属性段 sec 存在相关关系属性

　　sec′←corr(sec,kb_k,alg_r^j)

　　sec←sec′　　//基于数据属性段 sec 对第 k 个知识库应用相关关系计算算法 alg_r^j 得到数据属性段 sec′并存入 sec

　　RDD_j←RDD_j+f_j(sec) //将数据属性段 sec 以概率密度函数 f_j 存入 RDD_j 中

　end for

end for

RDD_{DB}←RDD_1+RDD_2+…+RDD_i//将生成的多个 RDD 合并

4　实验结果与分析

本节以网络访问日志数据的生成为例介绍本文提出的数据生成方法的实际应用，并对所提出的数据生成工具[1)]进行以下几个方面的实验验证：生成数据分布的多样性、生成数据的分布检验、生成数据的相关关系检验、数据生成速度、集群规模对数据生成时间的影响、数据属性段个数对数据生成时间的影响。

4.1　网络访问日志数据结构分析

在互联网上，由于用户的网络访问行为产生了体量庞大的网络访问日志，其中蕴藏着丰富的社会和商业价值，也包含着大量的用户隐私信息。在科研工作中，因为网络带宽的限制，下载这些网络日志需要耗费大量时间。常见的网络访问日志数据包含时间戳、地理位置信息、URL 信息等数值统计信息[16]，本节将以生成特定网络访问日志为例，介绍生成的目标数据属性，如表 1 所列。数据生成的过程如表 2 所列。

表 1　目标数据属性

Table 1　Target data properties

属性名	数据属性段	来源	分布形态或相关关系
访问时间	属性段 1	依据用户给出的时间戳范围，例如：2018 年 7 月 12 日这一天的时间戳范围为[1531324800, 1531411200]，来确定时间戳知识库 kb_{time}	访问时间在午间到达访问量高峰，早晚访问量递减，即这一属性在一天内呈现正态分布
离开时间		由访问时间进行相关关系计算得到	离开时间 T_{leave} 后访问时间 T_{access} 约一小时，将约一小时表示为正态分布 N(1,1)，即两者存在相关关系：$T_{leave}=T_{access}+\frac{1}{\sqrt{2\pi}}\exp(-\frac{(x-1)}{2})$
源 IP	属性段 2	提供源/目标 IP、国家、省份、城市属性的 IP 地址知识库以纯真 IP 数据库[17]为原始数据集整理得到，知识库示例如表 3 所列。本例中根据访问数据量不同将源 IP 的来源分为两个 IP 地址知识库：$kb_{IP\text{-}HF}$ 和 $kb_{IP\text{-}LF}$	由于网络访问的目标有集中性的特点，总访问量 n 中，95% 的访问量目标 IP 存储在 IP 地址知识库 $kb_{IP\text{-}HF}$ 中，另外 5% 的访问量目标 IP 来源于 $kb_{IP\text{-}LF}$，即目标 IP 属性段满足概率密度函数为：$F_{IP}=\begin{cases}\frac{1}{0.95\cdot n}, & 0.95\cdot n\\ \frac{1}{0.05\cdot n}, & 0.05\cdot n\end{cases}$
目标 IP			
端口			
国家			
省份			
城市			
URL	属性段 3	URL、DOMAIN 知识库 kb_{URL} 基于 Alexa 站点统计访问量前 2000 的网址[18]整理标注得到，知识库示例如表 4 所列	访问量 n 中 80% 的访问量访问搜索引擎类的 URL，20% 的访问量访问其他类型的 URL，即 URL 属性段满足概率密度函数为：$F_{URL}=\begin{cases}\frac{1}{0.8\cdot n}, 0.8\cdot n\\ \frac{1}{0.2\cdot n}, 0.2\cdot n\end{cases}$
DOMAIN			

表 2　数据生成过程

Tabel 2　Data generation process

属性名	采样算法	生成属性段	生成过程	存储集合
访问时间	根据正态分布的分布形态确定采样算法 alg_s^{access}	sec_{access}	$samp(kb_{time}, alg_s^{access})\rightarrow sec_{access}$	RDD_{access}
离开时间	根据访问时间与离开时间的线性相关关系确定相关关系计算算法 alg_r^{leave}	sec'_{leave}	$corr(sec_{access}, kb_{time}, alg_r^{leave})\rightarrow sec'_{leave}$	RDD_{leave}
源 IP	依据配置的分段概率密度函数确定采样算法 $alg_s^{IP\text{-}HF}$ 和 $alg_s^{IP\text{-}LF}$	$sec_{IP\text{-}HF}$, $sec_{IP\text{-}LF}$	$samp(kb_{IP\text{-}HF}, alg_s^{IP\text{-}HF})\rightarrow sec_{IP\text{-}HF}$ $samp(kb_{IP\text{-}LF}, alg_s^{IP\text{-}LF})\rightarrow sec_{IP\text{-}LF}$	RDD_{IP}
目标 IP				
端口				
国家				
省份				
城市				
URL	依据配置的分段概率密度函数确定采样算法 alg_s^{URL}	sec_{URL}	$samp(kb, alg_s^{URL})\rightarrow sec_{URL}$	RDD_{URL}
DOMAIN				

1) https://github.com/libojia-aug/GenerateDataset

表 3 IP 地址知识库示例

Table 3 Example of IP address knowledge base

IP	端口	国家	省份	城市	经度坐标	纬度坐标
1.12.119.41	80	中国	北京	北京	116.40514055052417	39.723946957916
1.13.111.104	80	中国	天津	天津	117.2120884567028	38.9539903115013
⋮						

表 4 URL、DOMAIN 知识库示例

Table 4 Example of URL、DOMAIN knowledge base

URL	DOMAIN	类型	网站名	主要地区
google.com	google.com	搜索引擎	Google	美国
youtube.com	youtube.com	视频分享	YouTube	美国
facebook.com	facebook.com	社交网络	Facebook	美国
baidu.com	baidu.com	搜索引擎	百度	中国
⋮				

将这些数据属性段集合 RDD_{access}，RDD_{leave}，RDD_{IP}，RDD_{URL} 按照目标数据格式拼接得到目标数据集合 RDD_{DB}，最后数据存储模块将其交付到数据仓库。

4.2 生成数据统计学检验

以生成数据时间戳为例，刻画数据记录在时间域上的分布形态，模拟均匀、单峰、多峰等数据分布特点。

图 2、图 3 展示了生成数据满足指数分布、正态分布的分布形态。图 4 展示了在连续时间窗口上两次正态分布的分布形态。

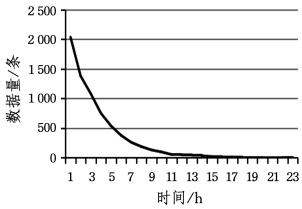

图 2 指数分布展示

Fig. 2 Exponent distribution example

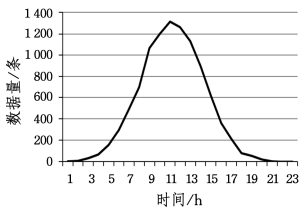

图 3 正态分布展示

Fig. 3 Normal distribution example

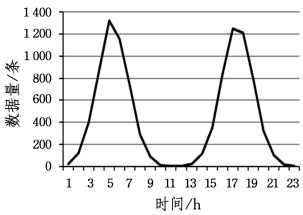

图 4 分段正态分布(双峰)展示

Fig. 4 Piecewise normal distribution example

图 5 展示了生成数据满足均匀分布的分布形态。图 6 模拟了夜晚高峰的数据分布形态，从中可以看到 0 点到 7 点、22 点到 23 点的数据量较高，其他时间的数据量处于较低水平。通过已有生成算法和用户自定义的生成算法，可以支持多样的数据分布形态。

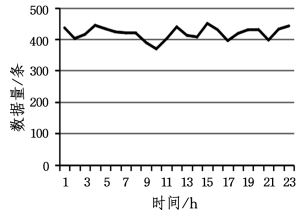

图 5 均匀分布展示

Fig. 5 Uniform distribution example

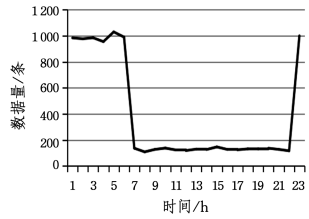

图 6 分段均匀分布展示

Fig. 6 Piecewise uniform distribution example

生成数据的分布检验以生成数值型属性为例，配置生成的数值属性满足正态分布。生成数值属性后，从中抽取 30 条数据进行 T 检验，并计算每组数据的 p 值，以验证数据是否满足均值为 0 的正态分布。

T 检验[19]主要应用于在样本总体的标准差未知的情况下，验证随机变量的数学期望是否等于某一已知值。p 值[20](Probability Value)是由检验统计量的样本观察值计算得出的原假设可被拒接的最小显著水平 α。对于任意的显著性水平，若 $p \leqslant \alpha$，则在显著性水平 α 下拒绝原假设，反之，接受原假设。

表 5 生成数据的 p 值

Table 5 Probability value of Generated data

	第 1 组	第 2 组	第 3 组	第 4 组	第 5 组
p	0.6750	0.5963	0.9499	0.2545	0.3724

从表 5 可知，当显著性水平 $\alpha=0.05$ 时，生成的数据都是满足原假设的，即分布均值为 0，这说明生成数据的分布形态是符合配置的。

在生成数据的相关关系检验中，为了更直观地表现数据之间的相关关系，本文选用两组相互独立的正态分布数据生成二维坐标，结果如图 7－图 10 所示。对二维坐标的每一维数据进行线性或非线性变化，得到新的坐标，再计算其线性相关系数，并使用线性回归计算线性变换方程，以此证明线性相关关系刻画的准确性。图 7－图 10 中，三角型坐标点是变换之前的数据，圆型坐标点是变换之后的数据，四组实验的二维坐标都选用相互独立的正态分布。4 组实验中，前 3 组在生成数据时配置了线性变换方程，如图 7－图 9 所示；第四组配

置了非线性变换方程，如图 10 所示。

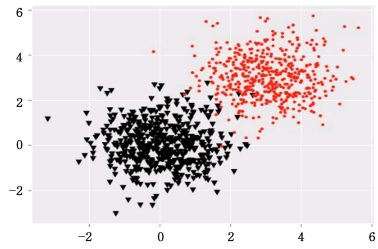

图 7　变换方程：$Y=X+3$

Fig. 7　Transformation function: $Y=X+3$

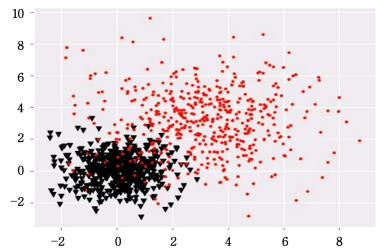

图 8　变换方程：$Y=2X+3$

Fig. 8　Transformation function: $Y=2X+3$

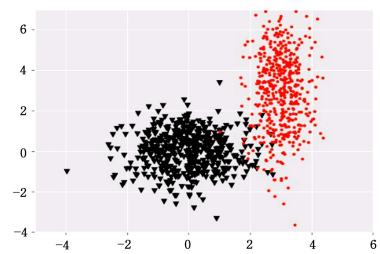

图 9　变换方程：横坐标 $Y=0.5X+3$，纵坐标 $Y=2X+3$

Fig. 9　Transformation function: abscissa: $Y=0.5X+3$, ordinates: $Y=2X+3$

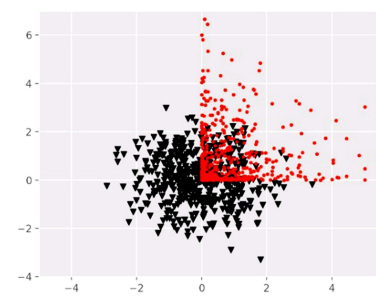

图 10　变换方程：$Y=X^2$

Fig. 10　Transformation function: $Y=X^2$

从表 6 中不难看出，当给出两组数据间的相关关系函数后，本系统可以充分表现多样的相关关系。特别地，因为随机变量 Y 是由随机变量 X 按线性关系函数变换出的，所以线性相关系数表现出了极强的线性相关性，而第四组实验并不是线性变换的，表现出了较弱的线性相关性。同时，应用线性回归方法，对线性相关数据都能求出与配置值相同的线性参数。这说明这种使数据间具有相关性的数据生成方法是有效的。

表 6　相关关系分析表

Table 6　Correlation analysis table

组号	维度	线性相关系数	配置斜率	线性回归斜率	配置截距	线性回归截距
第一组	x 轴	1	1	1.00	3	3.00
	y 轴	1	1	1.00	3	3.00
第二组	x 轴	1	2	2.00	3	3.00
	y 轴	1	2	2.00	3	3.00
第三组	x 轴	1	0.5	0.50	3	3.00
	y 轴	1	2	2.00	3	3.00
第四组	x 轴	0.0599				
	y 轴	0.0132				

4.3　数据生成模型的可伸缩性

数据的生成速度是考量数据生成工具的重要指标。本文实验选用 4 台物理机作为运算节点，每台机器硬件的配置如下：CPU 使用 Intel(R) Xeon(R) CPU E5-2640 v2 @ 2.00 GHz，内存为 8 GB。生成单列均匀分布数据，生成速度如图 11 所示。

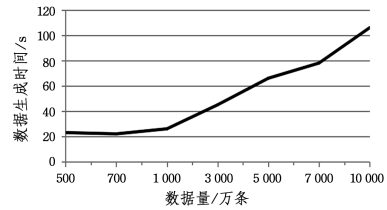

图 11　数据生成速度

Fig. 11　Data generation speed

当生成数据量在 1000 万条以下时，生成数据量与生成时间的关系并不明显，生成时间约 22 s 左右，这主要是因为生成的数据量较少，程序执行时间大多用于集群调度，与生成数据量的关联较少。当生成数据量在 1000 万条到 1 亿条时，生成数据量和生成时间基本呈线性关系，这表明在本数据量范围内，数据生成时间与生成数据量呈线性关系。当生成 1 亿行数据时，所用时间为 106 s，生成速度约每秒 94.34 万条。

针对不同类型的数据，本文选取了 4 个开源数据集作为数据生成的原始数据[1)]。数据结构如表 7 所列。

表 7　数据结构

Table 7　Summary of data sets

数据集	数据类型	数据源	数据大小
Wikipedia 词条	非结构化	文本数据	4 300 000 条英文词条
Amazon 电影评论	半结构化	文本数据	7 911 684 条评论
Facebook 社交网络	非结构化	图表数据	4 039 个点、88 234 条边
Google 网络数据	非结构化	图表数据	875 713 个点、5 105 039 条边

本系统与 BDGS[7] 数据生成工具的对比如图 12、图 13 所示，BDGS 实验环境选用 1 台搭载 2 个 Xeon E5645 处理器、32 GB 内存的主机。对于文本数据，选用每秒生成文件的大小为评价标准；对于图表数据，选用每秒生成边的数量为评价标准。根据实验数据可以看出，本系统在生成数据效率上有显著提升，提升约为 25%。这主要是因为本系统在生成数据前，对数据分布形态和相关关系进行了分析，由程序配置文件进行设置，不需要在生成数据时再对数据集进行分析。

1) http://prof.ict.ac.cn/BigDataBench/#downloads

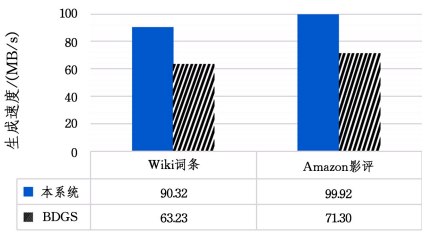

图 12 文本数据生成速度对比

Fig. 12 Comparison of text data generation speed

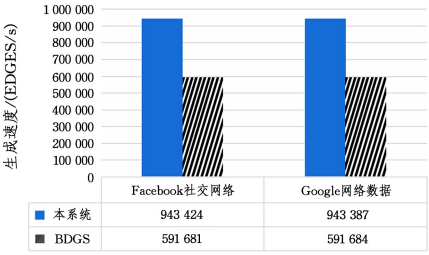

图 13 图表数据生成速度对比

Fig. 13 Comparison of graph data generation speed

通过分布式系统生成数据时，集群规模直接决定了计算资源的量，对数据的生成时间有直接的影响。为适应不同的数据量生成需求，动态调整系统资源尤为重要。图 14 展示了生成多字段数据时，机器节点数对生成时间的影响。分别生成 1000 万行、700 万行、500 万行、300 万行、100 万行数据时，集群规模的增大使得生成时间非线性减少。整体来看，增加节点对生成速度有积极影响。但如图 14 所示，生成 100 万行数据时，增加节点对数据生成时间的优化程度不大；与计算节点由 3 个增加到 4 个相比，计算节点由 1 个增加到 2 个时生成数据的时间并没有呈现同样的线性减小关系。这说明对于不同的生成数据需求，集群节点数的选取尽管整体表现为节点数越多生成数据的速度就越快，但综合考虑占用资源、时间优化等各方面的系统性定量标准还有待探究。

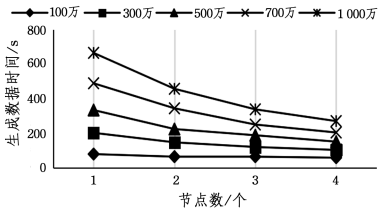

图 14 生成数据时间与节点的关系

Fig. 14 Relationship of generating data time and node

下面探究数据属性个数对数据生成时间的影响。生成多列数据时，数据属性之间往往存在着相关关系（正相关、负相关、零相关），为了刻画这类相关关系，在生成多列存在相关关系的数据属性时，要以一定的规则进行数据属性拼接。这种拼接的实现需要存在相关关系的两列数据进行笛卡尔积运算。这在大量数据上进行时会产生较大的系统开销，从而影响数据的生成时间。图 15 对生成数据属性数量与生成时间的关系进行了探究。生成数据中相关属性的数量与生成时间呈现线性增长关系，且数据量越大产生的影响越明显。这主要是因为当生成数据量较少时，系统负载较小，集群计算资源没有完全使用。当数据量较大时，集群计算资源紧缺，新增属性拼接所带来的额外计算量需要在计算队列中等待，从而延长了数据的生成时间。虽然数据属性的拼接会对数据生成时间产生负面影响，但这种影响只有当数据属性间存在相关关系时才会出现，完全无关的属性拼接不会进行笛卡尔积的匹配运算，自然不会明显延长运算时间。

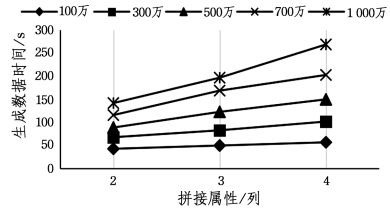

图 15 生成数据时间与拼接属性列数的关系

Fig. 15 Relationship of generating data time and attribute columns

结束语 本文提出并实现了一种可指定数据分布、可指定数据属性间相关关系的海量数据生成模型，介绍了利用蓄水池抽样算法从 Web 信息数据知识库中抽取数据属性段，实现了可配置分布形态和相关关系的数据生成模型。通过大量实验，展示了生成数据分布形态的多样性、生成数据在统计学上的可靠性、生成数据属性段之间线性和非线性相关关系的准确性。在系统性能上，针对生成数据的极限速度、集群规模、生成数据相关关系属性个数对生成时间的影响进行了分析，并阐述了相应影响的原因。

在今后的研究中，将着重关注以下问题：1）如何刻画更加逼真的数据分布形态；2）如何更加优化地使用系统资源，减少系统的资源消耗。

参 考 文 献

[1] PAN W. The current situation and trend of big data development in China[J]. The Science of Leadership Forum, 2017(4): 28-44. (in Chinese)
潘文. 我国大数据发展现状与趋势[J]. 领导科学论坛, 2017(4): 28-44.

[2] BUSARI M, WILLIAMSON C. PRoWGen: A synthetic workload generation tool for simulation evaluation of web proxy caches[J]. Computer Networks, 2002, 38(6): 779-794.

[3] RABL T, POESS M, DANISCH M, et al. Rapid development of data generators using meta generators in PDGF[C]// International Workshop on Testing Database Systems. ACM, 2013: 1-6.

[4] RABL T, FRANK M, SERGIEH H M, et al. A data generator for cloud-scale benchmarking[C]// Performance Evaluation, Measurement and Characterization of Complex Systems. Springer, 2011: 41-56.

[5] GHAZAL A, RABL T, HU M Q, Raab F, Meikel Poess, Alain Crolotte, and Hans-Arno Jacobsen. Bigbench: Towards an industry standard benchmark for big data analytics[C]// Proceesings of the 2013 ACM SIGMOD International Conference on Management of Data. ACM, 2013.

[6] HUANG S S, HUANG J, DAI J Q, et al. The hibench bench-

mark suite：Characterization of the mapreduce-based data analysis[C]// 2010 IEEE 26th International Conference on Data Engineering Workshops (ICDEW). IEEE，2010：41-51.

[7] MING Z，LUO C，GAO W，et al. BDGS：A scalable big data generator suite in big data benchmarking[C]// Advancing Big Data Benchmarks. Springer International Publishing，2014：138-154.

[8] YIN J，LU X，ZHAO X，et al. BURSE：A bursty and self-similar workload generator for cloud computing[J]. IEEE Trans. on Parallel & Distributed Systems，2015，26(3)：668-680.

[9] QIU Z P，XIAO R P，ZHANG R. Simulate generating web log algorithm using fields' priority relevance[J]. Computer Systems & Applications，2017，26(3)：126-133. (in Chinese)
丘志鹏，肖如良，张 锐. 优先关联的 Web 日志数据逼真生成算法[J]. 计算机系统应用，2017，26(3)：126-133.

[10] ZHAO H Q，LIU J L. Research on complex event big data processing system test data generation method based on Bayesian network[J/OL]. Application Research of Computers，2018(8)：1-2. [2018-06-26]. http://kns. cnki. net/kcms/detail/51. 1196. TP. 20180507. 1706. 040. html. (in Chinese)
赵会群，刘金銮. 基于贝叶斯网络的复杂事件大数据处理系统测试数据生成方法研究[J/OL]. 计算机应用研究，2018(8)：1-2. [2018-06-26]. http://kns. cnki. net/kcms/detail/51. 1196. TP. 20180507. 1706. 040. html.

[11] XU P，LIU J Y，LIN B，et al. Generation of fuzzing test case based on recurrent neural networks[J/OL]. Application Research of Computers，2019(10)：1-3. [2018-06-26]. http://kns. cnki. net/kcms/detail/51. 1196. TP. 20180619. 1517. 062. html. (in Chinese)
徐鹏，刘嘉勇，林波，等. 基于循环神经网络的模糊测试用例生成[J/OL]. 计算机应用研究，2019(10)：1-3. [2018-06-26]. http://kns. cnki. net/kcms/detail/51. 1196. TP. 20180619. 1517. 062. html.

[12] WANG K F，ZUO W M，TAN Y，et al. Generative confrontation network：from generating data to creating intelligence[J]. ACTA Automatic Sinica，2018，44(5)：769-774. (in Chinese)
王坤峰，左旺孟，谭营，等. 生成式对抗网络：从生成数据到创造智能[J]. 自动化学报，2018，44(5)：769-774.

[13] GOODFELLOW I J，POUGET-ABADIE J，MIRZA M，et al. Generative adversarial nets[C]// Proceedings of the 27th International Conference on Neural Information Processing Systems. Montreal，Canada：Curran Associates，Inc.，2014：2672-2680

[14] CHEN X，DUAN Y，HOUTHOOFT R，et al. Info GAN：interpretable representation learning by information maximizing generative adversarial nets[C]// Proceedings of the 30th Conference on Neural Information Processing Systems. Barcelona，Spain：Curran Associates，Inc.，2016.

[15] LI F J，YANG Z Q. An integrated sampling method over imbalanced network flows[J]. Fire Contrd & Command Contrd.，2015，40(20)：74-79. (in Chinese)
李富景，杨志强. 一种面向不均衡网络流的综合抽样方法[J]. 火力与指挥控制，2015，40(20)：74-79.

[16] GUAN L，HU G J，WANG Z. Research on network security situational awareness technology base on big data[J]. Netinfo Security，2016(9)：45-50. (in Chinese)
管磊，胡光俊，王专. 基于大数据的网络安全态势感知技术研究[J]. 信息网络安全，2016(9)：45-50.

[17] 纯真 IP 数据库[DB/OL]. http://www. onlinedown. net/soft/19051. html.

[18] Alexa 站点流量统计[DB/OL]. http://www. alexa. cn/.

[19] THOMAS W. MacFarland. Student's t-Test for Independent Samples[M]. Springer International Publishing：2014-06-15.

[20] FAN D M. The P value in hypothesis testing[J]. Journal of Zheng Zhou Economic Management Institute，2002(4)：70-71. (in Chinese)
樊冬梅. 假设检验中的 P 值[J]. 郑州经济管理干部学院学报，2002(4)：70-71.