计算机科学 COMPUTER SCIENCE
http://www.jsjkx.com
DOI:10.11896/jsjkx.191000156

# 一种大数据估价算法

赵会群[1]　吴凯锋[2]
1 北方工业大学信息学院　北京 100144
2 北方工业大学大规模流数据集成与分析技术北京市重点实验室　北京 100144
(zhaohq6625@sina.com)

**摘　要**　“大数据”已经成为计算机领域使用频率最高的专业词汇之一,而且已经逐渐变成了一个商品名称。无论是从学术研究角度,还是从数据交易需求角度,对大数据集的可用性进行评价都是一个新的问题。文中提出了一个大数据可用性评价模型,为学术和流通领域提供参考。结合大数据的4V(Volume,Variety,Velocity,Value)特性,分段统计样本数据的4V特性分布,从而给出基于分段分布的大数据特性概率模型,以及大数据可用性加权评价模型。文中还提出了实现大数据分块抽样的算法,以及大数据评价模型的各个特性加权系数的估计算法。结合视频大数据的可用性评价需求,展示所提模型和算法的具体应用。大数据可用性评价模型可以用于数据科学实验的数据评价,也可以用于大数据交易市场的数据集定价。给出了实际评价工作中,标准化(商品化)数据集以及确定数据评价基准等具体操作方面的解决方案。应用案例对所提模型有支持作用,进一步检验了模型的可行性。

**关键词**:大数据可用性评价;概率模型;大数据分块算法;视频大数据
**中图法分类号**　TP391

## Big Data Valuation Algorithm

ZHAO Hui-qun[1] and WU Kai-feng[2]
1 College of Computer Science and Technology,North China University of Technology,Beijing 100144,China
2 Beijing Key Laboratory of Large-scale Stream Data Integration and Analysis Technology,North China University of Technology,Beijing 100144,China

**Abstract**　With the rapid development of information technology,the generation of data has shown an exponential growth trend. Big data has become one of the most frequently used words due to the rapid emergence of big data and its great value. It is not only an academic vocabulary,but has gradually become a commodity name. Whether from academic research or data trading needs,how to evaluate the availability of big data sets is a new issue. A big data usability evaluation model is proposed to provide reference for academic and circulation fields in this paper. Combined with the 4V(Volume,Variety,Velocity,Value) characteristics of big data,the 4V characteristic distribution of the statistical data is segmented,which gives the probability model of big data based on the piecewise distribution and the availability of large data sets and weighted evaluation model. An algorithm for realizing big data block sampling and an estimation algorithm for weighting coefficients of each characteristic in the big data set evaluation model are proposed. Combined with the data availability evaluation requirements in video big data analysis,the specific applications of the proposed models and algorithms are demonstrated. The big data usability evaluation model can be used for data evaluation of data science experiments,and can also be used for data set pricing in big data transaction markets. In the actual evaluation work,how to standardize(commercialized) data sets,and how to determine the specific operational aspects of the video field evaluation benchmarks are given. The application case supports the proposed model and further tests the feasibility of the model.

**Keywords**　Big data availability evaluation,Probability model,Big data blocking algorithm,Video big data

## 1　引言

随着信息技术的快速发展,特别是“互联网”“物联网”及“云计算”等技术的突飞猛进,电子商务、生物医疗、金融投资等领域产生了海量的数据。例如,电商平台淘宝网每日成交量约为800万条;短视频应用抖音国内日活跃用户数超2.5亿人次;全球连锁沃尔玛超市每小时需要处理100余万条用户请求,同时维护着一个超过2.5PB的数据库。在高能

到稿日期:2019-10-24　返修日期:2020-02-02　本文已加入开放科学计划(OSID),请扫描上方二维码获取补充信息。
基金项目:国家自然科学基金项目(61672041)
This work was supported by the National Natural Science Foundation of China(61672041).
通信作者:吴凯锋(wkf0305@qq.com)

物理实验中，2008 年开始投入使用的大型强子对撞机每年产生超过 25PB 的数据；社交网络 Facebook 现已存储超过 500 亿张照片[1]。一方面，大数据给信息技术发展带来了新的挑战和机遇，分析和处理大数据已经成为业界的热点技术问题；另一方面，大数据也成为了一种战略资源，拥有这种资源就可以成为大数据市场的主导者。

大数据交易是大数据产业发展的重要部分，贵阳大数据交易中心的挂牌运行说明了数据是一种新的商品。随着贵阳大数据交易中心的成立，众多大数据交易平台也相继成立并投入使用，如中关村数海大数据交易平台、长江大数据交易所、东湖大数据交易平台、湖北大数据交易平台、西咸新区大数据交易所、交通大数据交易平台、华中大数据交易所、河北大数据交易中心和华东江苏大数据交易中心。随着大数据交易组织机构的迅猛增加，各大交易机构的服务体系也在不断完善。例如，被称为“中国数谷”的贵阳大数据交易中心为数据交易制订了十大标准以及规范体系，其中一项就是数据交易定价体系。交易规则中将数据定价分类为协议定价、拍卖定价和集合竞价，这 3 种定价模式都来源于传统交易。传统定价模型已经难以对大数据这种新型商品进行定价，而对大数据进行准确定价的前提是确定数据的可用性。

虽然目前大数据研究已经蓬勃兴起，但是大多数研究针对大数据环境下的数据仓库架构[2]、大数据降维[3]、相关性分析[4]及海量数据应用[5]等方面，并形成了一批重要的研究成果。在技术领域和数据交易市场的研究中，学者们对如何评价大数据的价值始终没有达成共识。现阶段传统企业数据资产管理存在问题[6]，现有定价指标缺乏对大数据产品成本的考虑[7]，忽略了数据的交易维度[8]；运用竞价机制的鲁宾斯坦讨价还价模型探索交易价格制定只是概念上的提出，并没有考虑数据资产的特有属性[9]。大数据交易仍有很多不确定性，交易过程存在很多风险[10]，传统的数据确权缺乏技术可信度，很难保证个人敏感信息不被泄露[11]，且存在潜在的篡改等不可控因素[12]。大数据交易平台在数据隐私风险、交易标准化体系的构建、价格机制的探索以及专业人才队伍建设等方面有待突破[13]。从数据科学角度研究数据的价值，分析数据市场模型[14]，将定价函数的结构与函数属性关联起来，以完整地表征定价函数，进而评判数据可用性的方法，还没有经过实验论证[15]。有效地评价训练数据集的可用性，一直是一个难点问题[16]。

综上，评价大数据的价值还有很多问题需要解决。科研领域需要针对具有鲜明数据特征的数据集进行研究，技术领域需要利用典型特征大数据集来测试大数据处理系统的能力，交易市场需要根据大数据对象的可用性，确定交易的等级和价格。因此，如何评价大数据的可用性不仅是一个数据交易市场的新问题，也是科研与技术领域的新课题。

本文对大数据可用性评价模型展开讨论，从大数据 4V 特性出发，采用分段函数分别定义大数据各个特性的可用性，并采用对各个特性分别加权求和的方式定义大数据集的可用性。文中给出了大数据各个特性概率分布求解算法和大数据特性加权系数的估计算法。最后通过一个视频大数据评价案例，检验上述方法的可行性。本文第 2 节给出大数据可用性评价的相关定义；第 3 节给出实现可用性评价的算法；第 4 节进行实验研究；第 5 节分析相关工作；最后总结全文。

## 2 大数据评价的 $4V^+$ 模型

本节给出基于大数据 $4V^+$ 特性的数据评价模型。首先给出 $4V^+$ 可用性评价的定义，最后给出大数据可用性评价模型。

**定义 1** 令 $S=\langle D,V\rangle$ 为大数据体积评价系统，其中，$D$ 是大数据集合，$\forall d_k\in D$ 为随机选取的数据样本，$V_k$ 为数据样本 $d_k$ 的体积度量；$V$ 是该数据集体积度量基准集合，$\forall v_i\in V(i=1,\cdots,n)$ 是标准化的数据度量基准。令：

$$V_k=\begin{cases}v_0, & V_k\geqslant v_0\\ v_1, & v_0>V_k\geqslant v_1\\ v_2, & v_1>V_k\geqslant v_2\\ \cdots & \\ v_n, & v_{n-1}>V_k\geqslant v_n\end{cases}\tag{1}$$

则将大数据 $D$ 的数据体积 $V$ 的可用性定义为 $P_v=P\{V_k=v_k\}\times(v_k/v_0)$。其中，$P\{V_k=v_k\}$ 是随机选取的样本数据体积 $d_k$ 高于某种基准度量 $v_k$ 的概率；$(v_k/v_0)$ 是归一化的基准值比率，其值越大，$V$ 的可用性越高。

上述定义中，数据体积度量基准 $V$ 可以根据行业特点给出，也可以根据用户喜好痕迹自动排名。例如，乒乓球比赛视频数据样本的体积基准可以按照比赛各局对应的文件字节长度确定为：第 1 局——$v_0$，第 2 局——$v_1$，第 3 局——$v_2$，…，第 7 局——$v_6$。

下面给出大数据多样性可用性的定义。

**定义 2** 令 $S=\langle D,M\rangle$ 是大数据数据格式多样性评价系统，其中，$D$ 是大数据集合，$\forall d_k\in D$ 为随机选取的数据样本，$M_k$ 为数据样本 $d_k$ 的多样性度量；$M$ 是数据格式多样性度量基准集合，$\forall m_i\in M(i=1,\cdots,n)$ 是标准化的数据度量基准。令：

$$M_k=\begin{cases}m_0, & M_k\geqslant m_0\\ m_1, & m_0>W_k\geqslant m_1\\ m_2, & m_1>W_k\geqslant m_2\\ \cdots & \\ m_n, & m_{n-1}>W_k\geqslant m_n\end{cases}\tag{2}$$

则将大数据 $D$ 的数据格式多样性 $M$ 的可用性定义为 $P_m=P\{M_k=m_k\}\times(m_k/m_0)$。其中，$P\{M_k=m_k\}$ 是随机选取的样本数据的多样性度量 $d_k$ 高于某种基准度量 $m_k$ 的概率；$(m_k/m_0)$ 是归一化的比率，其值越大，$M$ 的可用性越高。

例如，视频大数据的样本格式多样性基准被确定为媒体文件格式种类，包括 Mp4，Mkv，Avi 等。体育总局科研队分析体育比赛视频时，如果视频格式稀有，后期转换格式以及播放需要耗费大量的时间。

**定义 3** 令 $S=\langle D,T\rangle$ 是大数据数据处理紧迫性评价系统，其中，$D$ 是大数据集合，$\forall d_k\in D$ 为随机选取的数据样本，$T_k$ 为数据样本 $d_k$ 的紧迫性度量；$T$ 是该数据集体积度量基准集合，$\forall t_i\in T(i=1,\cdots,n)$ 是标准化的数据度量基准。令：

$$T_k=\begin{cases}t_0, & T_k\leqslant t_0\\ t_1, & t_0<T_k\leqslant t_1\\ t_2, & t_1<T_k\leqslant t_2\\ \cdots & \\ t_n, & t_{n-1}<T_k\leqslant t_n\end{cases}\tag{3}$$

则将大数据 $D$ 的数据处理紧迫性 $T$ 的可用性定义为 $P_t=P\{T_k=t_k\}\times(t_k/t_n)$。其中，$P\{T_k=t_k\}$是随机选取样本数据 $d_k$ 的紧迫性度量低于某种基准度量 $t_k$ 的概率；$(t_k/t_n)$是归一化的比率，其值越小，$T$ 的可用性越高。

例如，视频数据处理紧迫性基准可依靠帧率来定，因为每秒传输帧数(Frames Per Second，FPS)越大，视频就会越流畅。分别按照各种视频的帧率，如 10 fps，25 fps，1 000 fps，设定紧迫性基准度量。

大数据的第 4 个特性是数据有效价值比较低，这一特性表现在数据的内容新鲜程度不高，如体育比赛中的很多技术动作变化不大，体现在视频内容上数据的差异性不大。为此，本文用数据内容的差异性刻画和度量这一特性的可用性。

**定义 4** 令 $S=\langle D,F\rangle$是大数据数据差异性评价系统，其中，$D$ 是大数据集合，$\forall d_k\in D$ 为随机选取的数据样本，$F_k$ 为数据样本 $d_k$ 的差异性度量；$F$ 是数据集差异性度量基准集合，$\forall f_i\in F(i=1,\cdots,n)$是标准化的数据差异性度量基准。令：

$$F_k=\begin{cases}f_0, & F_k\leqslant f_0\\ f_1, & f_0<F_k\leqslant f_1\\ f_2, & f_1<F_k\leqslant f_2\\ \cdots & \\ f_n, & f_{n-1}<F_k\leqslant f_n\end{cases}\tag{4}$$

则将大数据 $S=\langle D,F\rangle$的数据差异性 $F$ 的可用性定义为 $P_v=P\{F_k=f_k\}\times(f_k/f_n)$。其中，$P\{F_k=f_k\}$是随机选取样本数据 $d_k$ 的差异性评价高于某种基准度量 $f_k$ 的概率；$(f_k/f_n)$是归一化的比率，其值越小，$F$ 的可用性越高。

除了上述根据大数据 4V 特性定义的可用性之外，还可以从领域应用需求凝练出用于评价数据可用性的其他因素，比如数据的真实性和数据的清洁性等。把综合多种领域因素的大数据可用性评价模型称为 $4V^+$ 模型。

**定义 5** 令 $S=\langle D,4V^+\rangle$是大数据可用性评价系统，大数据可用性评价模型为：

$$V_D=\alpha_1\times P_v+\alpha_2\times P_m+\alpha_3\times P_t+\alpha_4\times P_f\tag{5}$$

其中，$\Sigma\alpha_i=1(i=1,\cdots,4)$，$\alpha_i$ 称为价值加权系数。

在上述定义中，可以根据各个大数据特征的关注度调整相应的 $\alpha_i$ 值。虽然 $V_D$ 由代数表达式求得，但其仍然具有概率属性。

## 3 大数据可用性评价的实现算法

### 3.1 大数据 4V 特性的获取

依照大数据可用性评价的 $4V^+$ 模型，获取数据集的 4V 特性至关重要。大数据 4V 特性的获取可分解为以下 3 个步骤：

1)为数据集中的每条数据建立索引；

2)基于文件索引依次获取每条数据的 4V 特性；

3)将获取结果按顺序存储，形成与数据集对应的 4V 特性文件。

上述步骤对应的算法如算法 1 所示。

**算法 1** 大数据 4V 特性获取算法

输入：数据文件地址

输出：数据集 4V 属性文件

```
Begin
1. create index on bigdata/*为数据集每条数据建立索引，共有 N 条数据 */;
2. Volume[]=""/*用来存储数据的 Volume 属性*/
3. Velocity[]=""
4. Variety[]=""
5. Value[]=""
6. for(i=0;i<N;i++)
7. {
8.     Volume[i]=GetVolume(i);/*获取文件 Volume 属性，一般取文件大小*/
9.     Velocity[i]=GetVelocity(i);
10.    Variety[i]=GetVariety(i);/*获取文件 Variety 属性，一般取文件格式*/
11.    Valve[i]=GetValue( i )
12. }
End
```

大数据 4V 特性获取算法的性能主要取决于遍历时 $N$ 的大小，而 $N$ 反映了数据集中数据的个数，与单条数据大小无关，因此，该算法的时间复杂度为 $O(N)$。

### 3.2 大数据样本的块划分

依照大数据可用性评价的 $4V^+$ 模型，各项特性指标的评价都需要通过大数据样本进行。因此，如何划分大数据样本块是一个需要明确的问题。大数据样本可以有多种分块方法：按照传统的文件体积为计量单位进行分块；按照多媒体数据播放的时间度量进行分块；按照网络数据流量进行分块；按照数据内容完整性度量进行分块；等。其中，前 3 种分块方法延用了操作系统中文件组织方式，而第 4 种方法是一种新的大数据分块方法。上述方法也可以综合使用，从而形成多维度度量数据块。本节重点讨论基于内容的分块方法和技术。

基于内容的分块过程可分解为以下 3 个步骤：

1)建立数据样本内容描述；

2)基于内容表述对大数据进行过滤和汇聚；

3)存储汇聚的数据，形成数据样本块。

上述步骤对应的算法如算法 2 所示。

**算法 2** 基于内容的大数据分块算法

输入：数据内容描述和大数据集

输出：数据块和分块数

```
Begin
1. DataKey[m][]; /*二维数组用于存放内容描述关键词，构成 m 个内容描述关键词向量*/
2. DataBlockAdd[]; /*存放数据块地址*/
3. File[]; /*存放数据文件名字*/
4. Thread theFilters[];/*定义一组线程*/
5.    for( i=1;i<N;i++) /*N 为抽样次数*/
6.       File[i]=RandomAccessFile(Dataset,"r");
7.    for(i=1;i<N;i++)
8.       for( j=1;j<M;j++)
9.          read key[i][j];/*输入一组内容描述*/
10.   While key[i][j]<>"#" {
11.   DataBlockAdd [k]=Filter(key[i][j],File[]);/*对所有的数据源进行过滤*/
12.   IF DataBlockAdd [k]<>""
13.      Collector(Key[i][j],DataBlockAdd[k])/*保存数据块样本*/
14.      k++;
```

```
15.    }
16.    Filter(String key[][],File[])
17.    { for(int i=1;i<N;i++)
18.    theFilters[i](Key[][],File[i])/*采用线程的方式分步过滤满足内容描述的数据样本*/
19.    return DataBolckAdd;/*数据块地址*/
20.}
End
```

上述算法设计了 Filter( )和 Collector( )两个函数，分别用于实现基于内容的数据过滤和收集，其中 Filter( )函数可以通过线程同步实现，这与 Hadoop 中的 MapReduce 策略相似。上述算法的设计也借鉴了基于内容的推荐算法，因此如何确定基于内容的数据描述是分块的关键，一般通过内容标签对数据进行划分。

分块算法的时空性能依赖于内容表述的粒度和大数据体积，但由于采用了分布式计算技术，算法 2 的时空性能仅依赖于内容表述和体积最大的本地文件。因此，算法 2 的时间复杂度为 $O(n\times m)+O(L)$，其中 $n$ 和 $m$ 是内容表述向量的维度，$L$ 是最大局部文件的长度；类似地，其空间复杂度为 $O(n\times m)+O(L)$。

### 3.3 大数据可用性评价算法

根据提出的 $4V^+$ 模型，计算大数据可用性的步骤如下：

1）随机选取大数据样本；

2）根据 $4V^+$ 模型计算各特性的可用性；

3）计算各个特征值的权重系数；

4）计算大数据集的可用性。

上述步骤中，第 1）步的实现可以参考算法 2 进行，这里不再累述；第 2）步可以根据数据样本的外部特征和基准度量来计算，即根据数据样本的体积 $P_v$、处理时间紧迫性 $P_t$、结构化多样性 $P_m$、处理多样性 $P_r$ 和基准指标来计算该数据样本特征值；第 3）步和第 4）步是需要进一步明确的算法，如算法 3 所示。

**算法 3** 大数据可用性评价算法

```
输入：选择条件 Sele[]，数据样本的特征值(v,t,m,f)
输出：基于特征的可用性度量 V_D
Begin
1. Banch[][]; /*用于存放抽样的分布，即每行依次为(V,T,M,F)编号，每列为落入基准区间的次数*/
2. for(i=1;i<N;i++)/*N 为数据特征个数*/
3.    for( j=1;j<M;j++)/*M 为基准区间个数*/
4.       for( k=1;k<O;k++)/*O 为抽样次数*/
5.          { Banch[i][j]=Statistic(i,BlockDataAdd[k]);
6.          P_i=P{θ_i=b_k}×(b_k/b_n);/*P_i∈(P_v,P_t,P_m,P_f),θ_i∈(V_k,T_k,M_k,F_k),b_k 和 b_n 是基准值*/
7.       F_ij=Num(θ_i)/Max(Banch[i][j]);/*θ_i∈(v,t,m,r)*/
8.       IDF_i=Log(O/Num(θ_i));
9.       α_i=F_ij(Log(O/ Num(θ_i))
10.}
11.  V_D=α_1P_v+α_2P_t+α_3P_m+α_4P_r;
End
```

算法 3 中的 $F_{ij}$ 是某个特性在 $N$ 次抽样后出现的频率，$\theta_i$ 表示本次抽样的某个特征值，$Num(\theta_i)$ 为选定分块的该特征值在基准区间出现的次数，$Max(\theta_j)$ 记录每个特征区间的数值，则 $Max(O)$ 是最大值，代表样本出现最多基准区间的次数；$IDF_i=\mathrm{Log}(O/Num(\theta_i))$ 称为逆向频率，$O$ 是累计抽样次数；$\alpha_i=F_{ij}\times\mathrm{Log}(O/Num(\theta_i))$ 是数据样本可用性公式中各个特征值的权重系数。

上述算法的主要计算任务是特征值权重系数的计算，这里参考了著名的 TF-IDF(特征频率/逆向频率(Item Frequency/Inverse Frequency，TF-IDF))公式，这是一种用于信息检索与数据挖掘的常用加权技术。算法 3 的计算代价完全依赖于抽样的次数，因此其执行时间为 $O(N)$。

## 4 实验分析

本节以视频大数据为例，具体介绍以 $4V^+$ 模型评价视频大数据可用性的过程。实验选取的样本数据是互联网上的常见视频，包含体育竞技视频、短视频、电影集，涵盖了多种格式，来源众多。表 1 列出了数据源的相关信息。

表 1 数据源相关信息

Table 1 Data source related information

| VideoClass | File Size /GB | File Format | Store |
|---|---|---|---|
| Short Video | 13.4 | Mp4 | /shore |
| Movie Set | 23.6 | Rmvb,Mkv | /movies |
| Sports Competition | 83.9 | Flv,Avi,Rmvb,Mp4 | /sports |
| Others | 5.0 | Mp4,Avi | /others |

视频数据集的总文件大小约 125 GB，其中短视频来源于抖音短视频、火山小视频、全民小视频和快手等各大短视频网站；体育竞技视频主要包括国乒队历年赛事和其他运动竞技视频，主要下载源为“中国体育”；电影集样本来自视频网站优酷和爱奇艺等。

### 4.1 视频大数据 4V 特性的获取

结合视频数据集，本次实验选取文件大小衡量 Volume 属性，选取视频格式衡量 Variety 属性，选取视频帧率衡量 Volocity 属性。视频大数据的价值属性体现在视频的传播量上。在实验中，我们通过爬取对应视频的社交属性信息，发现视频的社交属性信息包括：评论数、点赞数、分享数、下载数和转发数。在此次研究中，我们定义影响因子和传播因子共同作用视频的价值属性。其中，影响因子包括点赞和评论，传播因子包括分享和转发。这样，可以对数据进行降维，迅速聚焦研究重点。

依据算法 1 对视频数据集的 4V 进行提取，部分结果如表 2 所列。

表 2 视频大数据 4V 特性

Table 2 Video big data 4V features

| $i$ | Volume | Variety | Velocity | $F_{k1}$ | $F_{k2}$ |
|---|---|---|---|---|---|
| 0 | 7668643 | Mp4 | 20 | 1897 | 225 |
| 1 | 45852 | Avi | 30 | 264398 | 27575 |
| 2 | 7455921 | Flv | 25 | 56 | 9 |
| ⋮ | ⋮ | ⋮ | ⋮ | ⋮ | ⋮ |
| $n-1$ | 731200 | Rmvb | 60 | 1432 | 265 |
| $n$ | 44228 | Mov | 10 | 189513 | 1603 |

表 2 中，Volume 的单位是字节，Velocity 的单位是帧/秒。每一行对应单个视频数据所包含的 4V 属性；最后一列价值属性分为 $F_{k1}$ 和 $F_{k2}$，分别表示此样本的影响因子和传播因子。

根据图 1 到图 4 的统计图表，对每个“V”进行分析，如 K-means 聚类分析、AP(Affinity Propagation)聚类分析、基于统

计学的抽样调查和基于 Apriori 算法的技术关联分析等。实验得出的一组大数据的特征评价基准如表 3 所列。


图 1 数据体积统计图

Fig. 1 Data volume diagram

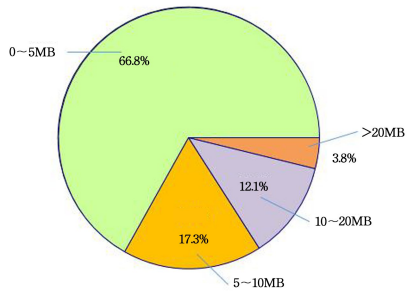


图 2 体积区间统计图

Fig. 2 Volume partition map

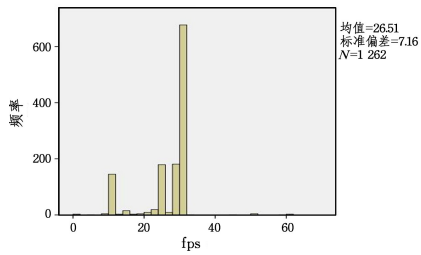


图 3 数据帧率统计图

Fig. 3 Data frame rate diagram

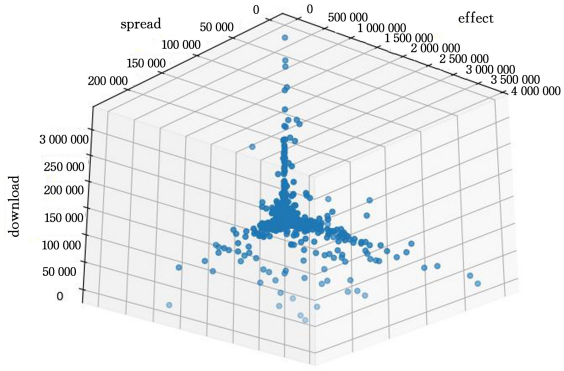


图 4 价值属性三维散点图

Fig. 4 Value attribute three-dimensional scatter plot

表 3 视频数据评价基准

Table 3 Video data evaluation criteria

| 特征 | 区间 1 | 区间 2 | 区间 3 | 区间 4 | 区间 5 | 区间 6 | 区间 7 |
|---|---|---|---|---|---|---|---|
| $V_k$ | <5 M | 5~10M | 10~20M | >20M | null | null | null |
| $M_k$ | Mp4 | Avi | Mov | Wmv | Rmvb | Mkv | Flv |
| $T_k$ | <10 fps | 10~25fps | 25~30fps | >30fps | null | null | null |
| $F_k$ | a | b | c | d | e | f | null |

表 3 中，$V_k$ 的最小数据体积基准值是 5 M，以此顺延有 5~10 M，10~20 M 等区间；$M_k$ 的最小数据格式多样性基准值是视频格式种类，且其占比分别为 76.2%，9.9%，0.2%，2.1%，1.3%，3.4%，6.9%；$T_k$ 的最小数据处理时间紧迫性基准值为 10 帧/秒，以此递增，最大值上限为 30fps；$F_k$ 共有 6 个区间，对应于表 4 的第一列，可根据影响因子和价值因子进行区间的确定。其中，影响因子 $F_{k1}$ 和传播因子 $F_{k2}$ 具有相关性，即 $F_{k1}$ 的值变大时 $F_{k2}$ 也会变大；同时，实验数据表明，$F_{k1}$ 和 $F_{k2}$ 会出现在对等区间中，如表 4 所列。

表 4 价值属性的基准定义

Table 4 Value attribute benchmark definition

| 区间 | $F_{k1}$ | $F_{k2}$ |
|---|---|---|
| a | <307 274 | <6 563 |
| b | [307 274, ] | [6 563, 15 762] |
| c | [957 211, 1 694 309] | [15 762, 31 879] |
| d | [1 694 309, 2 431 097] | [31 879, 50 489] |
| e | [2 431 079, 3 247 111] | [50 489, 86 438] |
| f | >3 247 111 | >86 438 |

### 4.2 大数据样本的块划分

通过简单随机抽样从数据源中抽取数据，依据算法 2 记录符合过滤条件的数据并分块存储，依据算法 1 对分块存储的数据获取 4V 特性。简单随机抽样方法适用于样本容量较大而各个体之间差异较小的情况，符合实验条件。

表 5 列出了一组视频大数据分块的实际数据，此时对每个分块过滤条件抽样 100 次。其中，第一列依据“PGC+UGC”内容创作模式进行分块，“PGC+UGC”是以内容为核心的新媒体传播平台内容创作模式（PGC（Professionally-Generated Content）指专业生产内容，UGC（User-Generated Content）指用户生产内容），是互联网内容的主要来源，此处仅列出了 5 种常用的情形；第二列是在进行数据抽样时，符合第一列的分块过滤条件的次数，其中“日常美食生活”在抽样中出现次数最多，为 66 次；第三列是各特征值分别在分块抽样中出现最多的区间次数，比如（$22/I_1$，$40/I_1$，$36/I_3$，$40/I_1$）分别表示体积特性 $V_k$、数据格式多样性 $M_k$、处理时间紧迫性 $T_k$ 和数据内容差异性 $F_k$ 在分块抽样中在某区间出现次数的最大值。

表 5 大数据可用性评价的抽样数据

Table 5 Sample data for evaluating big data usability

| 分块过滤条件 | 次数 | 最大值（$V_k$，$M_k$，$T_k$，$F_k$） |
|---|---|---|
| 体育竞技比赛 | 31 | （$23/I_4$，$23/I_1$，$18/I_2$，$30/I_3$） |
| 日常美食生活 | 66 | （$22/I_1$，$40/I_1$，$36/I_3$，$40/I_1$） |
| 数码科技电脑 | 33 | （$11/I_1$，$16/I_1$，$16/I_2$，$16/I_2$） |
| 舞蹈动漫动画 | 41 | （$10/I_1$，$17/I_1$，$12/I_3$，$15/I_5$） |
| 时尚美妆健身 | 53 | （$22/I_1$，$30/I_1$，$30/I_3$，$28/I_1$） |

### 4.3 视频可用性评价

依据算法 3 对表 5 所列抽样数据进行可用性评价，结果如表 6 所列。

表 6 评价可用性的抽样数据

Table 6 Sample results of usability evaluation

| $(V,T,M,F)$ | $(P_v,P_t,P_m,P_f)$ | $(\alpha_1,\alpha_2,\alpha_3,\alpha_4)$ | $V_d$ |
|---|---|---|---|
| $(I_4,I_1,I_2,I_3)$ | (0.18,0.74,0.23,0.17) | (0.73,0.73,0.64,0.84) | 0.96 |
| $(I_1,I_1,I_3,I_1)$ | (0.33,0.60,0.16,0.60) | (0.34,0.45,0.43,0.45) | 0.72 |
| $(I_1,I_1,I_2,I_2)$ | (0.33,0.48,0.19,0.15) | (0.44,0.56,0.56,0.56) | 0.60 |
| $(I_1,I_1,I_3,I_5)$ | (0.24,0.41,0.08,0.03) | (0.33,0.46,0.37,0.43) | 0.31 |
| $(I_1,I_1,I_3,I_1)$ | (0.41,0.57,0.17,0.52) | (0.42,0.49,0.49,0.48) | 0.78 |

表 6 中的第二列数据依据算法 3 中的 $P_i=P\{\theta_i=b_k\}\times(b_k/b_n)$得出，第三列数据依据算法 3 中的 $\alpha_i=F_{ij}\times\alpha_i=F_{ij}\times IDF_i$ 得出。数据可用性 $V_d$ 是不同分块抽样下的大数据集可用性度量值，不同的 $V_d$ 值从不同的应用角度给出了客观的度量，符合实际情况。

## 5 相关工作

本节从数据资产管理、服务评价模型和服务推荐系统 3 个方面概述与本文相关的研究工作，并分析本文工作与相关工作的联系和区别。

### 5.1 数据资产管理

大数据研究正蓬勃兴起，近年来对大数据资产管理方面的研究逐渐增多，而数据的可用性评价是关乎数据资产管理和服务的核心要素。

Cheng[6]认为现阶段传统企业数据资产管理的六大问题是：1)数据黑盒化；2)数据多头管理；3)数据标准不统一；4)数据缺乏有效治理；5)数据生命周期运营不完整；6)数据流通渠道不通畅。他们结合相关数据管理经验，阐述了数据资产管理五星模型，包括数据架构、数据治理、数据运营、数据共享和数据变现，为众多传统企业数据资产管理的相关实践提供了参考。该文虽然从多方面进行了分析，但局限于概念层次，并没有深入到算法层面。Wu 等[17]提出大数据“HAO 治理”模型，该模型以支持人类智能 HI、人工智能 AI 和组织智能 OI 的三者协同为目标，并以公安的数据治理为例来介绍 HAO 治理的应用。该模型对结构化数据的处理效果较好。

邹照菊提出了大数据资产计价应当采用以产出价值为基础的计量模式，采用现金流量现值作为计量属性，货币计量手段与非货币计量手段相结合的原则。此研究将传统资产与大数据资产进行了对比，从计价方面对大数据资产管理进行了分析。Guo 等[18]提出了一种个人大数据资产管理与增值服务系统，包括数据确权、汇聚、管理、交易和增值服务等功能，以个人为主体对象组织数据，有效连接人与数据；同时提出了个人数据银行的定义，探讨了个人数据银行平台的组成架构和关键技术，从理论和实践两方面分析了个人数据银行建设的可行性，为个人数据资产化管理奠定了基础。

### 5.2 服务评价模型

大数据作为一种资源，可以用于数据服务。数据可用性评价，是评价服务质量的关键。目前，与服务评价相关的研究工作并不太多。

易安信[19]是一家美国信息存储资讯科技公司，其白皮书中预测，到 2021 年将有 15%的互联网投资人把业务转移到数据服务中来，大数据服务的市场份额将增加到每年 300 亿美元。该报道虽然没有给出具体数据服务的评价模型，但提示我们大数据市场对数据服务的需求将迅速增加。Liu 等[20]给出了一个新的相似度度量模型 NHSM，该模型从 3 个方面评价服务数据推荐的相似度，包括接近程度、重要性和奇异性，而不是用传统的皮尔逊系数。该评价模型对新用户的冷点数据服务推荐准确度较好。Zhao 等[21]针对网络软件可信性评价问题，提出基于分段分布统计的概率模型，分别用区间函数定义网络软件的可靠性、可维护性和安全性等特性，并给出了上述特性加权和的可信性计算公式，但没有考虑相关系数的求解问题。贵阳大数据交易所给出了一种 6 维大数据定价方法，该方法将数据划分成 6 个维度，包括数据品种、时间跨度、数据深度、数据完整性、数据样本覆盖和数据实时性，不同品种的数据实行不同的价格机制。实时价格主要取决于数据的样本量和单一样本的数据指标项价值，再通过交易系统自动定价，价格实时浮动。该评价方法与本文提出的评价方法都采用抽样数据推断大数据集的可用性，但该方法没有考虑大数据的 4V 特性。

### 5.3 服务推荐系统

本文提出的大数据可用性评价模型中，首先需要统计数据抽样在大数据特性评价基准中的概率分布，其中数据抽样算法与服务推荐系统(Recommender System，RS)有相似之处。下面仅概述与本课题相关的研究成果。

Le[22]认为新用户冷起点问题是推荐系统中的研究前沿，并综述了该方向的最新研究成果；通过对几种典型推荐算法的比较和分析，给出了对这些推荐算法的实际评测数据，认为 NHSM 算法在冷起点推荐中最为准确。Rahul 等[23]综述了情感推荐系统(Affective Recommender System)最近 15 年来的研究成果，从人的行为、大脑思维、感觉、情绪、面部表情、肢体语言和人机交互的心理特征等方面对上述成果进行分类和归纳，给出了目前研究的最新状态，以及今后可能的研究方向，还给出了情感推荐的算法、数据集、研究平台和应用领域。Tommaso 等[24]针对推荐系统对内容差异性的要求，提出了一种用户偏好差异性模型，通过基于内容的属性特性计算内容的差异性；给出了根据差异性计算推荐列表的算法，并通过 MovieLens 数据集测试了所提算法的有效性。María 等[25]介绍了一款上下文知晓推荐系统测试集的生成工具 DataGenCARS，针对目前推荐系统算法中对象排名信息缺乏上下文信息而导致推荐不准确等问题，提出上下文信息自动生成算法和排名与上下文信息混合推荐的算法，为评价此类推荐系统的功能提供了高可用性的数据集。

**结束语** 本文提出了一个大数据可用性评价模型，给出了大数据可用性求解的算法，并结合视频大数据可用性评价展示了该模型的实际应用效果。大数据可用性评价模型可以用于数据科学实验的数据评价，也可以用于大数据交易市场的数据集定价。

从学术角度研究大数据可用性评价的方法和技术还不多见，存在很多技术难题。比如，大数据的标准化需要根据不同领域进行，没有统一的标准，这也是今后我们继续研究的问题。

## 参 考 文 献

[1] LI J Z，LIU X M. An important aspect of big data：data availabi-

lity [J]. Computer Research and Development, 2013, 50(6): 1147-1162.
[2] WANG S, WANG H J, XI X P, et al. Architectural Big Data: Challenges, Status Quo and Prospects[J]. Chinese Journal of Computers, 2011, 34(10): 1741-1752.
[3] LIANG J Y, WANG F, DANG C Y, et al. An efficient rough feature selection algorithm with a multigranulation view[J]. Interational Journal of Approximate Reasoning, 2012, 53: 912-926.
[4] ZHOU H X, CHEN S C. A Canonical Correlation Analysis of Ordered Discrimination[J]. Journal of Software, 2014, 25(9): 2018-2025.
[5] HUO W, MENG X F. Research on Trajectory Privacy Protection Technology [J]. Chinese Journal of Computers, 2011, 34(10): 1820-1830.
[6] CHENG Y X. Methodology and Practice of Data Asset Management in the Age of Big Data[J]. Computer Applications and Software, 2018, 35(11): 326-329.
[7] ZHAO Z R. Analysis of Domestic Big Data Transaction Pricing [J]. Information Security & Communication Secrecy, 2017(5): 61-67.
[8] CHEN Y, ZHOU J E, DU J Q. A Credit Evaluation Method Based on Transaction Data[J]. Computer Applications and Software, 2018, 35(5): 168-171.
[9] VINAYAK R, BORKAR, MICHAEL J. Big Data Platforms: What's The Next? [J]. XRDS • FALL, 2012(1): 44-49.
[10] WANG W, ZHANG M J, WANG J. Research on Risk Factor Identification in Big Data Transaction Business Process [J/OL]. [2019-07-08]. http://kns.cnki.net/kcms/detail/11.1762.G3.20190603.0844.004.html.
[11] YE Q Q, MENG X F, ZHU M J, et al. A Review of Localized Differential Privacy Research[J]. Journal of Software, 2018, 29(7): 1981-2005.
[12] WANG H L, TIAN Y L, YIN X. Big Data Confirmation Scheme Based on Blockchain[J]. Computer Science, 2018, 45(2): 15-19, 24.
[13] HE C, WANG Y R. Research on the Difficulties and Countermeasures of Big Data Trading Platform in China[J]. Modern Love Newspaper, 2017, 37(8): 98-105, 153.
[14] NIYATOD, ABUALSHEIKHM, PING WING, et al. Market model and optimal pricing scheme of big data and internet of things(IOT)[J/OL]. Arxiv, 2016: 1-6. https://xueshu.baidu.com/usercenter/paper/show?paperid=8038a12a20a285199b002c907070d4f9&site=xueshu_se.
[15] DEEP S, KOUTRIS P. The design of arbitrage-free data pricing schemes[J]. Schloss Dagstuhl-Leibniz-Zentrum für Informatik, 2017(12): 1-18.
[16] TAN X T, GU Y Y, RUAN T, et al. Confidence Interval Method for Data Set Classification Availability Evaluation[J]. Computer Science, 2019, 46(1): 78-85.
[17] WU X D, DONG B B, CAO X Z, et al. Data Governance Technology [J/OL]. [2019-07-02]. https://doi.org/10.13328/j.cnki.jos.005854.
[18] GUO B, LI Q, DUAN X L, et al. Personal Data Banking—A New Model of Personal Big Data Asset Management and Value-added Services Based on Bank Architecture[J]. Computer Journal, 2017, 40(1): 126-143.
[19] EMC Solution Group. Big data-as-a-service: A market and technology perspective[R]. 2012.
[20] LIU H F, ZHENG H, AHMAD M, et al. A new user similarity model to improve the accuracy of collaborative filtering[J]. Knowledge-Based Systems, 2014(56): 156-166.
[21] ZHAO H Q, SUN J, ZHAO R X. A Model for Assessing the Dependability of Internetware Software Systems[C]// IEEE 39th Annual International Computers, Software & Applications Conference. 2015: 578-581.
[22] LE H S. Dealing with the new user cold-start problem in recommender systems: A comparative review[J]. Information Systems, 2016, 58: 87-104.
[23] KATARYA R, VERMA O P. Recent developments in affective recommender systems[J/OL]. Physica A Statal Mechanics & Its Applications, 2016: 182-190. https://xueshu.baidu.com/usercenter/paper/show?paperid=8038a12a20a285199b002c907070d4f9&site=xueshu_se.
[24] TOMMASO D N, JESSICA R, PAOLO T, et al. Adaptive multi-attribute diversity for recommender systems[J]. Information Sciences, 2017, 3: 234-253.
[25] MARÍA D C R H, SERGIO I, RAMÓN H R T L. DataGenCARS: A generator of synthetic data for the evaluation of context-aware recommendation systems[J]. Pervasive and Mobile Computing, 2017, 7: 516-541.
[26] LI J Z, WANG H Z, GAO H. Research Progress in Big Data Usability[J]. Journal of Software, 2016, 27(7): 1605-1625.
[27] Guiyang Big Data Trading Center. 2016 China Big Data Transaction White Paper [OL]. http://www.gbdex.com/website/view/bigData.jsp.

**ZHAO Hui-qun**, born in 1960, Ph. D, professor. His main research interests include software architecture, big data generation, internet of things, cloud computing, and sports computing.

**WU Kai-feng**, born in 1994, master. His main research interests include big data pricing, big data asset management, and big data services.