基于粒度树和使用关系的大数据价值计算研究

马文胜，侯锡林，王宏波，柳森

计算机科学 COMPUTER SCIENCE
http://www.jsjkx.com
DOI:10.11896/jsjkx.230300109

基于粒度树和使用关系的大数据价值计算研究

马文胜[1]　侯锡林[2]　王宏波[2]　柳　森[2]
1 辽宁科技大学电子与信息工程学院　辽宁 鞍山 114051
2 辽宁科技大学工商管理学院　辽宁 鞍山 114051
(1391291002@qq.com)

摘　要　文中研究了大数据最基本的核心“价值数值”。首先阐述了对大数据进行粒化的粗糙集方法、基于聚类的方法、商空间法、模糊信息方法和云模型方法等,并按它们的共同特性——“划分”,对大数据进行“粒化”,按划分的粗细在大数据中建立了“粒度树”,在“粒度树”中定义了“粒空间”。然后定义了粒空间与代表项目之间的使用关系,以及不同粒空间的使用关系满足的条件。最后按照在粒空间的使用关系中每个粒及每个粒集合的使用情况,将使用情况分为3种:“正则使用”“必然使用”“相关使用”。取它们的属性及对象的平均值,并圆整到0至100,作为大数据的“正则价值”“必然价值”“相关价值”的数值。给出大数据最基本的核心“价值数值”的有效计算方法,又给出大数据最基本的核心“价值数值”计算在远程医疗、城市管理、高等院校等多个领域的应用实例。

关键词:大数据价值;价值数值;粒度树;使用关系;正则使用;必然使用;相关使用
中图法分类号　TP311

Study on Value Calculation of Big Data Based on Granular Tree and Usage Relationship

MA Wensheng[1],HOU Xilin[2],WANG Hongbo[2] and LIU Sen[2]
1 School of Electtronic and Information Engineering,Liaoning University of Science and Technology,Anshan,Liaoning 114051,China
2 School of Business Administration,Liaoning University of Science and Technology,Anshan,Liaoning 114051,China

Abstract　Study the core “data results of value” of big data. Firstly,the rough set method,cluster-based method,quotient space method,fuzzy information method and cloud model method for granulating big data are described. According to their common characteristics —“division”,the big data is “granulated”,and a “granularity tree” is established in the big data according to the size of division. “granular space” is defined in the “granular Tree”. Then it defines the usage relationship between the granular space and the representative project,and the conditions that the usage relationship of different granular spaces meets. Finally,according to the usage of each particle and each particle set in the usage relationship of the particle space,the usage is divided into three types:“regular use” “inevitable use”and “related use”. Take the average value of their attributes and objects and round them to 0～100,as the values of “data results of value”“inevitable value” and “relevant value” of big data. The effective calculation method of the core “data results of value” of big data is given,and the application examples of the core “data results of value” calculation of big data in telemedicine,urban management,universities and other fields are also given.

Keywords　Big data value,Data results of value,Granularity tree,Usage relationship,Regular use,Inevitable use,Related use

1　引言

美国思想家阿尔文·托夫勒(Alvin Toffler)于20世纪80年代,在《第三次浪潮》中首次提出了“大数据”一词[1]。从此开启了一个全新的大数据时代。大数据是改变市场、组织机构,以及政府与公民关系的方法,大数据还是人们获得新的认知、创造新的价值的源泉[2]。

随着互联网、云计算、智能终端等技术的迅猛发展,各类信息系统在各种领域的不断应用,大量的数据在开放多源的渠道中产生,逐渐汇聚成一个巨大的、精准映射并持续记录物质世界和精神世界运动状态和状态变化的数据空间[3]。在这个空间中,大数据蕴藏着巨大的科学研究价值、公共管理与服务价值、商业价值以及科学决策价值[4-5]。这些蕴藏在大数据中的各种“价值”逐渐被人们所认知,所接受,所应用。Mayer-Schönberger等[6]甚至认为,大数据价值被纳入资产负债表都将是必然的。

于是,怎么计算大数据的各种价值?大数据各种价值的数值具体是多少?采用什么方法或模型获得?上述问题就成了研究的重要课题。美国政府早在2012年就启动了“大数据研发计划”,投资2亿美元来提升从数据中获取价值的能力[7]。

目前学者已给出很多不同的研究方法和方向。其中包括:1)按照一种定价模型方法对大数据价值进行研究;2)按照一种数据资产对大数据价值进行研究;3)按一种价值评估方法对大数据价值进行研究。另外还有很多国内外的研究者为研究大数据的价值,采用了其他很多各不相同的方法[8]。

这些对大数据价值的研究,大多以资产定价、价值估值、商品交易、成本核算等方向为主,而都不是以计算大数据本身的核心“价值数值”为主,然而计算出大数据核心“价值数值”是非常有现实意义的。

1)可以作为大数据交易价格数值的参考

近年来随着大数据更广泛的应用,大数据交易也继续

通信作者:侯锡林(hou.xilin@163.com)

迅猛增加，甚至进入了井喷期。据《2019 年中国大数据产业白皮书》的统计，2015 年我国大数据相关交易的市场规模为 33.85 亿元，2016 年达到 62.12 亿元，2020 年达到 545 亿元。我国已经成立包括贵阳大数据交易所[9]、武汉东湖大数据交易中心[10]、北京国际大数据交易所[11]在内的多家大数据交易中心。同时，在国际上也出现了一大批数据交易平台，如 Dawex[12]，Xignite[13]，World Quant[14]等[15]。在如此巨大的交易中，大数据作为一种商品，应有一个合理的“价格”，然而目前交易价格仍然是按照买卖双方的交易价格意愿，按照市场价格波动随意变价，交易过程中缺乏一种对大数据定价的“核心价值”作为统一标准，难以真正准确衡量大数据应有的基准价格。根据“价格”始终是围绕着“价值”上下波动这一经济学规律，显然大数据的这个核心“价值数值”是大数据交易“价格”最好的统一的参考。

2）作为大数据失窃“索赔”的参考

由于大数据具有很高价值，因此大数据的失窃也经常会出现，就《数邦客-大数据价值构建师》[1]统计，仅 2018 年上半年较大的泄漏就有：Saks 和 Lord & Taylor 泄露数据 500 万条；PumpUp 泄露数据 600 万条；Sacramento Bee 泄露数据 1 950 万条；Ticketfly 泄露数据超过 2 700 万条；Panera 泄露数据 3 700 万条；Facebook 泄露数据至少 8 700 万条；MyHeritage 泄露数据超过 9 200 万条；Under Armour 泄露数据 5 亿条；Exactis 泄露数据超过 4 亿条；Aadhaar 泄露数据 11 亿条。这些泄露所造成的损失都需要索赔，需要计算索赔价值。例如，目前最大的一次索赔是万豪连锁酒店的喜达屋酒店的大数据泄露事件索赔，这次提起的集体行动诉讼请求赔偿 125 亿美元[2]。显然索赔的金额也需要有一个“价值数值”来作为参考。

3）作为大数据更新“删除”的参考

大数据的数据量在呈现几何级数的增加，全球产生的数据总量于 2018 年为 18 ZB，于 2019 年为 33 ZB，于 2020 年为 47 ZB，2035 年将达 2 142 ZB。时间长了，存储空间、处理时间都无法承受，需要隔一段时间把“价值”已经不大的数据删掉。但仅凭论述、描述，无法给出具体大数据的“价值”，则无法准确进行价值大小的比较，而且也不规范，不能准确地确定所需要删除的大数据。因此也极需要计算出各种大数据的“价值数值”，用价值数值的大小来决定它们的去留。

4）作为大数据行业“标准”“规范”制定的参考

“十四五”为加快标准立法建设，优化数据环境，对大数据也开展了制定“标准”、制定“规范”的工作。在这些行业标准规范的法律、法规制定过程中，大数据“价格”标准和规范的制定，显然是一个重要的环节。计算出大数据的“价值数值”，作为“价格”的参考，显然是对价格制定的规范，提供了科学、精准、统一的参考。这使制定“标准”、制定“规范”的工作更加科学、标准、全面、有效。

还有很多领域也需要大数据的“价值数值”，因此计算出核心价值的“价值数值”是非常必要的。然而，现今 Fama[16]，Hansen[17]，Shiller 等[18]的各种资本资产定价模型，David[19]，Chiu[20]，Chen[21]，Lin[22]，Jorge[23]，Niyato[24]等的价值评估方法，都未能给出大数据的核心“价值数值”的计算方法，因此研究大数据核心“价值数值”的计算方法是非常迫切的，也是非常有意义的。侯锡林教授认为，对大数据的价值进行科学的评估和计算，创建大数据的价值模型，给出大数据的“价值数值”，无论在理论上还是在实践中，都是亟待解决的最重要的问题[25]。

如何计算出大数据的核心“价值数值”呢？在大数据的应用过程中，人们逐渐发现，不论在哪个领域，其实只有使用大数据，大数据才能体现出价值，大数据使用得越多，其体现出来的价值就越多。反之，如果没有使用，不论是纸质的，还是电子的，都没有任何价值，不用大数据就体现不出价值，其就都只是一堆文字和符号[8]。因此，人们逐渐深刻地认识到，大数据核心价值的多少应体现在大数据使用程度上。

人们还逐渐发现，虽然在众多领域中，大数据在各领域有各领域的特殊价值，例如科学决策价值、商业价值、经济价值、医疗价值、教育人文价值、公共管理服务价值、科研价值等，但体现大数据“使用”情况的价值，才是最核心最基本的价值。其他各个领域的各种价值都是这个核心基本价值的外在表现。例如科学决策价值：挖掘了关联规则的价值、挖掘了分类规则的价值、挖掘了聚类规则的价值等，无非都是“使用”了大数据后，才有这些挖掘，它们都是“使用”的外在表现。商业价值、经济价值、医疗价值等，也都是各自领域的外在表现价值。

本文将探讨基于“使用”的大数据最核心最基本的“价值数值”计算。首先将大数据“粒化”，在大数据中建立“粒度树”。然后考虑每个粒及每个粒集合的使用情况，并将使用情况分为 3 种，即“正则使用”“必然使用”“相关使用”，取它们的属性及对象的平均值作为大数据的“正则价值”“必然价值”“相关价值”的数值，并给出多个应用实例。

2 粒计算与粒度树

大数据是巨量数据、海量数据，是无法在一定时间范围内通过人工或计算机进行捕捉、管理和处理的数据集合[26]。因此大数据的“价值”计算也必须使用非传统的方法。在非传统的方法中，Chen 等将“粒计算”列为驾驭大数据的第一方法[27]。

粒计算的基本思想是把初始形式的数据分为不同的粒度进行处理。用粒度合适的“粒”作为处理对象，从而在保证求得满意解的前提下，提高解决问题的效率[28]。

目前对大数据进行粒化的方法，主要有粗糙集方法[29]、基于聚类的粒化方法[30]、商空间法[31]、模糊信息粒化方法[32]和云模型法[33-36]等。

1）粗糙集的粒化方法[29]

粗糙集是 1982 年波兰科学院院士帕拉克（Pawlack）教授提出的。该方法是将所有大数据称为论域 U，在其中建立一个等价关系 E，由这个等价关系形成等价类，每个等价类就是一个粒。$u\in U$，包含 u 的等价类，用 $[u]_E$ 表示，即 $[u]_E=\{v\in U|(u,v)\in E\}$。这些粒 $[u]_E$ 形成了 U 的一个“划分”（“划分”是 U 的一些非空子集 $X_1,\cdots,X_n\subseteq U$，这些子集满足 $\bigcup_{i=1}^{n}X_i=U$，及 $X_i\bigcap X_j=\varnothing, i\neq j, 1\leqslant i,j\leqslant n$）。可证明“等价关系”与“划分”是等效的。因此也可以说粗糙集方法是通过对论域 U 进行“划分”来形成粒的。同一个粒中的元素在这个等价关系

[1] https://www.databanker.cn

[2] https://www.prnewswire.com/news-releases/class-action-lawsuit-filed-on-behalf-of-plaintiffs-whose-sensitive-personal-information-was-stolen-in-breach-of-marriott-servers-300758440.html

中不可区分。对于论域 U 的任意子集，可能无法用一些粒来表示，即不恰好为某些粒的并，这时用上、下近似来表示。设 X 为 U 的任意子集，X 的下近似 $\underline{apr}(X)$ 和上近似 $\overline{apr}(X)$ 分别定义为：

$$\underline{apr}(X)=\bigcup\{[u]_E\mid[u]_E\subseteq X,u\in U\}$$

$$\overline{apr}(X)=\bigcup\{[u]_E\mid[u]_E\cap X\neq\varnothing,u\in U\}$$

用不同的等价关系可以产生不同的粒化层次。如果 E_1，E_2 是两个等价关系，而且对所有的 $u\in U$ 都有 $[u]_{E_1}\subseteq[u]_{E_2}$，这时必有 $E_1\subseteq E_2\Rightarrow\underline{apr}_{E_2}(X)\subseteq\underline{apr}_{E_1}(X)\subseteq X\subseteq\overline{apr}_{E_1}(X)\subseteq\overline{apr}_{E_2}(X)$，这时称等价关系 E_1 比 E_2 更细，形成了两个粒化层次，粒度越细，准确度越高（但反之不一定成立）。更一般地，可以考虑 n 个嵌套的等价关系序列 $E_1\subseteq E_2\subseteq\cdots\subseteq E_n$，可以形成 n 个粒化层次，满足各种准确度的需求。

2）基于聚类的粒化方法[30]

聚类的粒化也是基于“划分”，但需要有先验知识，即必须事先知道对象间的相似程度。根据这种相似程度寻求一种方法，对论域进行“划分”，使得同一类中的对象之间相似程度尽量大，不同类中的对象之间相似程度尽量小。根据聚类结果的结构，可以将聚类分为“划分聚类”和“层次聚类”。划分聚类得到论域上的一个“划分”；层次聚类是得到 n 个层次，每个层次都是论域上的一个“划分”，而下一层是把上一层中最接近的两个类合并成一类，从而总的类数少了一个而形成的，设形成的 n 个层次为 $H_1,H_2,\cdots,H_n$，如果 $1\leqslant i<j\leqslant n$，$\omega_p\in H_i$，$\omega_q\in H_j$，则 $\omega_q\subseteq\omega_p$ 或 $\omega_q\cap\omega_p=\varnothing$。两个类合并的方法有：single-linkage 方法，complete-linkage 方法以及 average-linkage 方法等。SL 方法，即 single-linkage 方法（也称 connectedness 或 minimum 方法）是类间距离等于两类对象之间的最小距离，若用相似度衡量，则是类间距离等于各类中的任一对象与另一类中任一对象的最大相似度。CL 方法，即 complete-linkage 方法（也称 diameter 或 maximum 方法）是类间距离等于两类对象之间的最大距离；AL 方法，即 average-linkage 方法是类间距离等于两类对象之间的平均距离。这种层次聚类称为“凝聚”法，它是逐步合并一些类，还有一种“分离”层次聚类法，与“凝聚”相反，它先将所有对象放在同一类中，然后逐步分离成一些更小的类，分离法一般很少使用。还有很多种基于知识的聚类方法，包括模糊聚类、半监督聚类、协同聚类、方向聚类等。Lin[37] 于 2010 年提出源自于谱聚类的幂迭代聚类（PowerIteration C lustering，PIC）方法。2013 年 Yan 等[38] 提出了一种大数据上的并行幂迭代聚类方法 p_PIC。

3）商空间法[31]

商空间也是对论域进行“划分”来形成粒。商空间法认为，一个问题是一个三元组 (X,f,T)。X 是问题的论域，函数 $f:X\to Y$ 表示论域的属性情况，这里 Y 为属性的取值。T 为论域的结构（例如 X 的子集形成的拓扑），是论域 X 中各元素之间的关系。问题的不同粒度就表示为不同的关系 R，也就是说，不同的粒度就是对论域的不同“划分”。因此“划分”也是商空间法中构成不同粒度世界的基本方法，可以根据 $f:X\to Y$ 的结果 Y 对 $X=f^{-1}(Y)$ 进行“划分”，也可按其他方法直接对 X 进行“划分”。例如可以有下述“划分”法。（1）属性“划分”法：即将属性相同或相似的元素归为一个类。投影“划分”法：若元素 x 的属性函数是多维的，例如有 n 个属性函数分量 $f_1,\cdots,f_n$，若暂不考虑其中的 k 个属性 $f_1,\cdots,f_k$，用将 $f_{k+1},\cdots,f_n$ 几个属性相同的元素归为一类的方法进行分类。结构“划分”法：把结构上或功能上关系密切的元素分为一类。约束“划分”法：设有 n 个约束条件 $C_1,\cdots,C_n$，那么可按 C_k 进行“划分”。当利用分类技术在粗粒度世界讨论问题时，若问题无解，那么在细粒度的原问题上也无解（保假原理），一个命题在两个较粗粒度的商空间中是真，则（在一定条件下）在其合成的商空间中对应的问题也是真的（保真原理）。这样就可缩小求解范围，加快求解进度，因为粗粒度世界通常比原世界简单。

4）模糊信息粒化[32]

模糊信息粒化理论（Theory of Fuzzy Information Granulation，TFIG）的出发点是广义约束，粒的主要类型有 3 种：可能性的，真实性的，概率性的。用广义约束来刻画粒的特征。模糊信息粒化理论已有一些推广模式：模糊化，粒化，模糊粒化。另外模糊信息粒化理论与 FIG 相关的推广模式还有：模糊化（Fuzzification，f-generalization），把一个清晰集用模糊集代替；粒化（Granulation，g-generalization），一个集合被划分成粒；随机化（Randomization，r-generalization），变量被随机变量代替；通常化（Usualization，u-generalization），命题 X is A 被 $Usually(X\text{ is }A)$ 代替。

5）基于云模型的粒化[31-35]

云模型是由我国李德毅教授创建的不确定性知识表示和推理模型，它可以揭示概念的随机性、模糊性以及随机性和模糊性之间的关联性，用期望熵和超熵作为数字特征表示定性概念，并通过云变换实现定性概念（概念内涵）和定量数据（概念外延）之间的相互转换。云模型的定义如下：设 U 是一个用精确数值表示的定量论域，C 是 U 上的定性概念，若定量值 $x\in U$，且 x 是定性概念 C 的一次随机实现，x 对 C 的确定度 $\mu(x)\in[0,1]$ 是有稳定倾向的随机数，即 $\mu:U\to[0,1]$，$\forall x\in U,x\to\mu(x)$，则 x 在论域 U 上的分布称为云模型，简称为云，每一个 x 称为一个云滴[32]，云模型用期望 Ex、熵 En 和超熵 He 这 3 个数字特征来表征一个概念，它们反映了定性概念 C 整体上的定量特征。使用云模型进行粒化的思路是使用逆向云发生器，从数据集中学习得到反映定性概念的数字特征集 $\{(Ex_i,En_i,He_i)\}$，$i=1,\cdots,n$[35]，然后将数据划分成 n 个粒。对各种类型数据集中对象的处理，经过如下的关键步骤就可以实现粒化：(1)云变换；(2)泛概念树的自动生成和爬升；(3)使用极大判定法[34]。

由这些方法来看，很多都是采用“划分”来形成粒。例如粗糙集方法、基于聚类的方法、商空间法等。本文将借鉴这些方法，利用基于“划分”的粒度树来对大数据进行粒化，计算大数据核心基本价值的价值量。

定义 1 设 G 是一个集合，$g_1,g_2,\cdots,g_n$ 是 G 的非空子集，若 $g_1\cup g_2\cup\cdots\cup g_n=G$，且 $g_i\cap g_j=\varnothing(1\leqslant i,j\leqslant n,i\neq j)$，则称 $\Pi=\{g_1,g_2,\cdots,g_n\}$ 是 G 的一个“划分”，这时每个 $g_i(1\leqslant i\leqslant n)$ 都称为是 G 的一个“粒”，称 Π 是“粒空间”。记 G 的所有“划分”为 $\Pi(G)$。如果 $\Pi_1,\Pi_2\in\Pi(G)$，且对 Π_1 的每一个粒 g'，都有 Π_2 的某一个粒 g，使得 $g'\subseteq g$，则称“划分”Π_1 比另一“划分”Π_2 更细，或者“划分”Π_2 比另一“划分”Π_1 更粗，记作 $\Pi_1\leqslant\Pi_2$。

例 1 $G=\{1,2,3,4,5,6,7,8\}$，$\Pi_0=\{\{1,2,3,4,5,6,7,8\}\}$，$\Pi_1=\{\{1,2,3,4\},\{5,6\},\{7,8\}\}$，$\Pi_2=\{\{1,2,3,4\},\{5,6\},\{7\},\{8\}\}$，$\Pi_3=\{\{1\},\{2,3,4\},\{5,6\},\{7,8\}\}$。它们都是

G 的“划分”，而且这里 $\Pi_2 \leqslant \Pi_1 \leqslant \Pi_0$，$\Pi_3 \leqslant \Pi_1 \leqslant \Pi_0$. 但 $\Pi_2 \leqslant \Pi_3$ 且 $\Pi_3 \leqslant \Pi_2$。

定理 1 设 G 是集合，$\Pi_1, \Pi_2 \in \Pi(G)$，$\Pi_1 \leqslant \Pi_2$，$g_0 \in \Pi_2$，则 $\exists g_1, g_2, \cdots, g_k \in \Pi_1$，使得 $\{g_1, g_2, \cdots, g_k\}$ 是 g_0 的“划分”。

证明：由定义 1 知对 Π_1 的每一个粒 g'，都有 Π_2 的某一个粒 g，使得 $g' \subseteq g$。若 Π_1 中所有是 g_0 子集的粒是 $g_1, g_2, \cdots, g_k$。由于 $g_i \subseteq g_0$，$1 \leqslant i \leqslant k$，因此 $\bigcup_{1 \leqslant i \leqslant k} g_i \subseteq g_0$，即或者 $\bigcup_{1 \leqslant i \leqslant k} g_i \subset g_0$ 或者 $\bigcup_{1 \leqslant i \leqslant k} g_i = g_0$。如果 $\bigcup_{1 \leqslant i \leqslant k} g_i \subset g_0$，则存在 $d \in g_0$ 但 $d \notin \bigcup_{1 \leqslant i \leqslant k} g_i$。若 Π_1 中含 d 的粒是 g_{k+1}，由于 $d \in g_0$，而 Π_1 的每一个粒 g' 都有 Π_2 的某一个粒 g，使得 $g' \subseteq g$，因此 $g_{k+1} \subseteq g_0$，这与 Π_1 中所有是 g_0 子集的粒是 $g_1, g_2, \cdots, g_k$ 矛盾，故 $\bigcup_{1 \leqslant i \leqslant k} g_i \not\subset g_0$，从而 $\bigcup_{1 \leqslant i \leqslant k} g_i = g_0$。由于 Π_1 中所有粒是 Π_1 的划分，因此 $g_i \cap g_j = \varnothing (1 \leqslant i, j \leqslant k, i \neq j)$，故 $\{g_1, g_2, \cdots, g_k\}$ 是 g_0 的划分。

定义 2 设大数据 $\mathbb{D} = \{d_1, d_2, \cdots, d_N\}$，其中每个 $d_i (1 \leqslant i \leqslant N)$ 都是原始文件（例如：一个 txt 文件，一个 doc 文件，一个 bmp 文件，一个 gif 文件，一个 png 文件，一个 mp3 文件，一个 mp4 文件，等等）。它们是对 $\mathbb{D}$ 进行粒划分时的最小单位。一个以 $\mathbb{D}$ 的子集为节点的，满足以下条件的树 T_D 称为 $\mathbb{D}$ 的“粒度树”。

1) T_D 的根节点是 $\mathbb{D}$ 本身。

2) 每一个节点都有一个“名称”，根节点的“名称”是 *BigData*。

3) T_D 中若 $z_1, \cdots, z_n$ 是 z 的所有子节点，则 $\{z_1, \cdots, z_n\}$ 是 z 的一个划分。

粒度树 T_D 中有一个节点集 C，若满足 T_D 的每一个叶子节点 z 到根节点的路上，都存在且只存在唯一的节点 v 属于 C，则称 C 是 T_D 的一个“视角”。

例 2 设 $\mathbb{D} = \{\text{File}_1, \text{File}_2, \text{File}_3, \text{File}_4, \text{File}_5, \text{File}_6, \text{File}_7, \text{File}_8\}$，为了书写方便，以下只写下标：$\mathbb{D} = \{1,2,3,4,5,6,7,8\}$。图 1 给出了以 $\mathbb{D}$ 的子集为节点的树，满足以下条件：1) 根节点是 $\mathbb{D}$ 本身；2) 每个节点都有“名称”，根节点的“名称”是 *BigData*，其他节点名称是 $A, B, C, D, E, F, G, H, I$；3) 满足节点 z 的所有子节点是 z 的一个划分。例如：*BigDate* 有两个子节点 A, B，于是 $\{A, B\} = \{\{1,2,3,4\}, \{5,6,7,8\}\}$ 是 $|BigData| = \{1,2,3,4,5,6,7,8\}$ 的划分，A 有 3 个子节点 C, D, E，于是 $\{C, D, E\} = \{\{1\}, \{2,3\}, \{4\}\}$ 是 $|A| = \{1,2,3,4\}$ 的划分。

D 有两个子节点 H, I，于是 $\{H, I\} = \{\{2\}, \{3\}\}$ 是 $|D| = \{2,3\}$ 的划分。

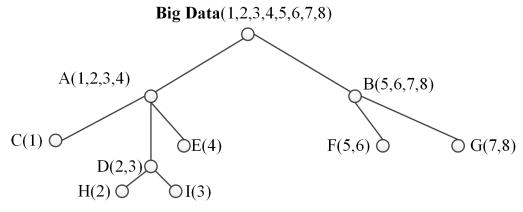

图 1 大数据的粒度树

Fig. 1 Granularity tree of big data

图 1 中，$\{C, H, I, E, F, G\}$ 是视角，$\{C, H, I, E, B\}$ 是视角，$\{C, D, E, B\}$ 是视角，$\{C, D, E, F, G\}$ 是视角，$\{C, A, F, G\}$ 不是视角，因为在从叶节点 C 到根节点 *BigDate* 的路上，存在两个节点 A 和 C。同理 $\{C, D, I, E, B\}$ 也不是视角，因为在从叶节点 I 到根节点 *BigDate* 的路上，存在两个节点 I 和 D。显然这个粒度树共有 7 个视角，$C_0 = \{BigData\}$，$C_1 = \{A, B\}$，$C_2 = \{A, F, G\}$，$C_3 = \{C, D, E, B\}$，$C_4 = \{C, D, E, F, G\}$，$C_5 = \{C, H, I, E, B\}$，$C_6 = \{C, H, I. E, F, G\}$。它们决定的粒空间分别是：$\Pi_0 = \{BigData\} = \{\{1,2,3,4,5,6,7,8\}\}$，$\Pi_1 = \{A, B\} = \{\{1,2,3,4\}, \{5,6,7,8\}\}$，$\Pi_2 = \{A, F, G\} = \{\{1,2,3,4\}, \{5,6\}, \{7,8\}\}$，$\Pi_3 = \{C, D, E, B\} = \{\{1\}, \{2,3\}, \{4\}, \{5,6,7,8\}\}$，$\Pi_4 = \{C, D, E, F, G\} = \{\{1\}, \{2,3\}, \{4\}, \{5,6\}\{7,8\}\}$，$\Pi_5 = \{C, H, I, E, B\} = \{\{1\}, \{2\}, \{3\}, \{4\}, \{5,6,7,8\}\}$，$\Pi_6 = \{C, H, I, E, F, G\} = \{\{1\}, \{2\}, \{3\}, \{4\}, \{5,6\}\{7,8\}\}$。

定理 2 设 T_D 是大数据 $\mathbb{D}$ 的“粒度树”，C 是 T_D 的一个视角，则 C 是 $\mathbb{D}$ 的一个“划分”。并称其为 C 决定的“粒空间”，或 C 决定的“粒层次”。

证明：显然 T_D 可以这样来形成：令 $T_D^{(0)} = \mathbb{D}$，而对 $i = 0, 1, 2, \cdots$，把 $T_D^{(i)}$ 的某一个叶子节点细化，得到 $T_D^{(i+1)}$，最后直到某次 $T_D^{(m)} = T_D$ 为止。首先，对 T_D 的形成过程 $T_D^{(0)}, \cdots, T_D^{(i)}, T_D^{(i+1)}, \cdots, T_D^{(m)}$ 进行归纳，证明：若 $z_1, \cdots, z_p$ 是 T_D 的所有叶子节点，则 $z_1 \cup \cdots \cup z_p = \mathbb{D}$。（基础）对于 $T_D^{(0)}$，因 $T_D^{(0)} = \mathbb{D}$，所以结论正确。（归纳）若 $T_D^{(i)}$ 的叶子节点为 $w_1, w_2, \cdots, w_k$，且 $w_1 \cup \cdots \cup w_k = \mathbb{D}$，那么 w_1 细化为 $w_{1_1}, \cdots, w_{1_t}$ 后，因 $\{w_{1_1}, \cdots, w_{1_t}\}$ 是 w_1 的划分，所以 $w_{1_1} \cup \cdots \cup w_{1_t} = w_1$，故 $T_D^{(i+1)}$ 的叶子节点满足 $w_{1_1} \cup \cdots \cup w_{1_t} \cup w_2 \cup \cdots \cup w_k = w_1 \cup w_2 \cup \cdots \cup w_k = \mathbb{D}$。其次若“视角”$C$ 是 $\{v_1, v_2, \cdots, v_k\}$，而 T_D 的所有叶子节点 $z_1, \cdots, z_p$ 中是 v_i 的子节点的是 $z_{i_1}, \cdots, z_{i_q}$，则 $v_i = z_{i_1} \cup \cdots \cup z_{i_q}$。由于从每一个叶节点 z 到根节点 BigData 的路上，都存在唯一的节点 v 属于 C，因此 $i \neq j$ 时 $v_i \cap v_j = \varnothing$，另外因为只存在唯一的节点 v 属于 C，故 $v_1 \cup \cdots \cup v_k = z_1 \cup \cdots \cup z_p$，而前证 $z_1 \cup \cdots \cup z_p = \mathbb{D}$，因此 $v_1 \cup \cdots \cup v_k = \mathbb{D}$，故 C 是 $\mathbb{D}$ 的一个划分。

如何建立大数据的粒度树呢？关键是其第（3）步：“若 $z_1, z_2, \cdots, z_n$ 是节点 z 的所有子节点，则 $\{z_1, z_2, \cdots, z_n\}$ 是 z 的一个划分”。怎样对 z 进行划分呢？由计算公式来看，把 z 划分为 $\{z_1, z_2, \cdots, z_n\}$ 没有特殊要求，只要是“划分”就可以。但在实际工作中人们总是把属性相近的、建立时间相近的或同一课题的等，放在同一粒中。就如同放在同一个文件夹中的文件没有限制，但人们总是把属性相近的、建立时间相近的、同一课题的文件放在同一文件夹中一样。节点 z 的不同“划分”下，计算出的“价值数值”将会略有不同。这是正常的，因不同的“划分”方法会产生不同的“视角”。

3 使用关系与核心基本价值

如前文所述，大数据 $\mathbb{D}$ 如果没有被使用，则没有任何价值。大数据被使用得越多，则其价值越大。在大数据 $\mathbb{D}$ 中建立粒度树 T_D，对 T_D 中的每一个视角 C 决定的粒空间 $\Pi = \{g_1, g_2, \cdots, g_n\}$ 都可给出一个使用关系 I_Π. 我们将利用这些对应不同粒度 Π 的使用关系 I_Π 来描述大数据的使用情况。

定义 3 设 $\mathbb{D}$ 是大数据，$\Pi = \{g_1, g_2, \cdots, g_n\}$ 是 $\mathbb{D}$ 的一个粒空间。$\mathbb{U} = \{u_1, u_2, \cdots, u_m\}$ 是选取的使用大数据的一些代表项目（简称“项目”）的集合。$\mathbb{I}_\Pi \subseteq \mathbb{U} \times \Pi$（这里 $\times$ 是笛卡尔积）是 $\mathbb{U}$ 与 Π 间的关系，当且仅当项目 u_i 使用了粒 g_j 中的内容时 $(u_i, g_j) \in \mathbb{I}_\Pi$。称 $\mathbb{I}_\Pi$ 为大数据 $\mathbb{D}$ 对应 Π 的“使用关系”，简称“关系”。若 $\Pi_1 \leqslant \Pi_2$，则对于 $g_0 \in \Pi_2$，都存在 $g_1, g_2, \cdots, g_k \in \Pi_1$，使得 $\{g_1, g_2, \cdots, g_k\}$ 是 g_0 的划分（见定理 1）。这时 $\mathbb{I}_{\Pi_1}$，$\mathbb{I}_{\Pi_2}$ 应满足：

$$(u,g_0)\in\mathbb{I}_{\Pi_2}\Leftrightarrow(u,g_1)\in\mathbb{I}_{\Pi_1}\vee(u,g_2)\in\mathbb{I}_{\Pi_1}\vee\cdots\vee(u,g_n)\in\mathbb{I}_{\Pi_1}$$

例3(继续例2) 设项目为$\mathbb{U}=\{u_1,u_2,u_3,u_4,u_5,u_6,u_7\}$。$\mathbb{U}$与$\Pi_1,\Pi_2,\Pi_3,\Pi_4,\Pi_5,\Pi_6$间的使用关系$\mathbb{I}_{\Pi_1},\mathbb{I}_{\Pi_2},\mathbb{I}_{\Pi_3},\mathbb{I}_{\Pi_4},\mathbb{I}_{\Pi_5},\mathbb{I}_{\Pi_6}$如表1所列。

表1 大数据对应粒度$\Pi_1,\Pi_2,\Pi_3,\Pi_4,\Pi_5,\Pi_6$的使用关系$\mathbb{I}_{\Pi_1},\mathbb{I}_{\Pi_2},\mathbb{I}_{\Pi_3},\mathbb{I}_{\Pi_4},\mathbb{I}_{\Pi_5},\mathbb{I}_{\Pi_6}$

Table 1 Usage relationship $\Pi_1,\Pi_2,\Pi_3,\Pi_4,\Pi_5,\Pi_6$ of big data corresponding to granularity $\mathbb{I}_{\Pi_1},\mathbb{I}_{\Pi_2},\mathbb{I}_{\Pi_3},\mathbb{I}_{\Pi_4},\mathbb{I}_{\Pi_5},\mathbb{I}_{\Pi_6}$

$\mathbb{I}_{\Pi_1}$	A	B
u_1	×	×
u_2	×	
u_3		×
u_4	×	×
u_5	×	×
u_6	×	
u_7		×

$\mathbb{I}_{\Pi_2}$	A	F	G
u_1	×	×	×
u_2	×		
u_3			×
u_4	×	×	×
u_5	×	×	
u_6	×		
u_7		×	×

$\mathbb{I}_{\Pi_3}$	C	D	E	B
u_1	×	×		×
u_2		×	×	
u_3				×
u_4	×	×		×
u_5	×	×	×	×
u_6	×		×	
u_7				×

$\mathbb{I}_{\Pi_4}$	C	D	E	F	G
u_1	×	×		×	×
u_2		×	×		
u_3					×
u_4	×	×		×	×
u_5	×	×	×	×	
u_6	×		×		
u_7				×	×

$\mathbb{I}_{\Pi_5}$	C	H	I	E	B
u_1	×	×	×		×
u_2		×		×	
u_3					×
u_4	×		×		×
u_5	×	×	×	×	×
u_6	×			×	
u_7					×

$\mathbb{I}_{\Pi_6}$	C	H	I	E	F	G
u_1	×	×	×		×	×
u_2		×		×		
u_3						×
u_4	×		×		×	×
u_5	×	×	×	×	×	
u_6	×			×		
u_7					×	×

定义4 设大数据$\mathbb{D}$对应粒度Π及使用项目$\mathbb{U}$的使用关系为$\mathbb{I}$,若$u\in\mathbb{U},g\in\Pi$,则规定函数$\varphi(u)$为:$\varphi(u)=\{g\in\Pi|(u,g)\in\mathbb{I}\}$,并规定$\varphi(\varnothing)=\Pi$。

我们考察粒空间Π的任一个子集$G\subseteq\Pi$的使用情况。设$u\in\mathbb{U}$是一个使用者,则可能有3种情况:

1)u使用了G中的每一个粒,即$\varphi(u)\supseteq G$,这时称u为G的正则使用者;

2)u使用了G中的一部分粒,而且没有使用G以外的粒,即$\varphi(u)\subseteq G$,则称u为G的必然使用者;

3)u使用了G中的一部分粒,而且还使用了G以外的粒,即$\varphi(u)\cap G\neq\varnothing$,则称$u$为$G$的相关使用者。

注1:3种使用者的集合不是互相排斥的,例如若$\varphi(u)=G$,则u既是G的正则使用者,又是G的必然使用者,又是G的相关使用者。

定义5 对于各种使用者的集合:记G的所有正则使用者的集合为$N(G)$,即:

$$N(G)=\{u\in\mathbb{U}|\varphi(u)\supseteq G\}$$

记G的所有必然使用者的集合为$C(G)$,即:

$$C(G)=\{u\in\mathbb{U}|\varphi(u)\subseteq G\}$$

记G的所有相关使用者的集合为$R(G)$,即:

$$R(G)=\{u\in\mathbb{U}|\varphi(u)\cap G\neq\varnothing\}$$

于是根据定义5,各种使用者的总和分别为(这里$|S|$表示集合S的元素个数):

正则使用者的总和为$\sum\limits_{\substack{G\subseteq\Pi\\G\neq\varnothing}}|N(G)|$;

必然使用者的总和为$\sum\limits_{\substack{G\subseteq\Pi\\G\neq\varnothing}}|C(G)|$;

相关使用者的总和为$\sum\limits_{\substack{G\subseteq\Pi\\G\neq\varnothing}}|R(G)|$。

例4(继续例3) 我们取粒空间为Π_3,(以下将$\{C\}$简写为C,将$\{C,D\}$简写为CD等等),则$N(C)=\{u_1,u_4,u_5,u_6\}$,$N(D)=\{u_1,u_2,u_4,u_5\}$,$N(CD)=\{u_1,u_4,u_5\}$,$N(E)=\{u_2,u_5,u_6\}$,$N(CE)=\{u_5,u_6\}$,$N(DE)=\{u_2,u_5\}$,$N(CDE)=\{u_5\}$,$N(B)=\{u_1,u_3,u_4,u_5,u_7\}$,$N(CB)=\{u_1,u_4,u_5\}$,$N(DB)=\{u_1,u_4,u_5\}$,$N(CDB)=\{u_1,u_4,u_5\}$,$N(EB)=\{u_5\}$,$N(CEB)=\{u_5\}$,$N(DEB)=\{u_5\}$,$N(CDEB)=\{u_5\}$,$\sum\limits_{G\subseteq\Pi_3,G\neq\varnothing}|N(G)|=37$,$C(C)=\varnothing$,$C(D)=\varnothing$,$C(CD)=\varnothing$,$C(E)=\varnothing$,$C(CE)=\{u_6\}$,$C(DE)=\{u_2\}$,$C(CDE)=\{u_2,u_6\}$,$C(B)=\{u_3,u_7\}$,$C(CB)=\{u_3,u_7\}$,$C(DB)=\{u_3,u_7\}$,$C(CDB)=\{u_1,u_3,u_4,u_7\}$,$C(EB)=\{u_3,u_7\}$,$C(CEB)=\{u_3,u_6,u_7\}$,$C(DEB)=\{u_2,u_3,u_7\}$,$C(CDEB)=\{u_1,u_2,u_3,u_4,u_5,u_6,u_7\}$,$\sum\limits_{G\subseteq\Pi_3,G\neq\varnothing}|C(G)|=29$,$R(C)=\{u_1,u_4,u_5,u_6\}$,$R(D)=\{u_1,u_2,u_4,u_5\}$,$R(CD)=\{u_1,u_2,u_4,u_5,u_6\}$,$R(E)=\{u_2,u_5,u_6\}$,$R(CE)=\{u_1,u_2,u_4,u_5,u_6\}$,$R(DE)=\{u_1,u_2,u_4,u_5,u_6\}$,$R(CDE)=\{u_1,u_2,u_4,u_5,u_6\}$,$R(B)=\{u_1,u_3,u_4,u_5,u_7\}$,$R(CB)=\{u_1,u_3,u_4,u_5,u_6,u_7\}$,$R(DB)=\{u_1,u_2,u_3,u_4,u_5,u_7\}$,$R(CDB)=\{u_1,u_2,u_3,u_4,u_5,u_6,u_7\}$,$R(EB)=\{u_1,u_2,u_3,u_4,u_5,u_6,u_7\}$,$R(CEB)=\{u_1,u_2,u_3,u_4,u_5,u_6,u_7\}$,$R(DEB)=\{u_1,u_2,u_3,u_4,u_5,u_6,u_7\}$,$R(CDEB)=\{u_1,u_2,u_3,u_4,u_5,u_6,u_7\}$,$\sum\limits_{G\subseteq\Pi_3,G\neq\varnothing}|R(G)|=83$。

正则使用者的总数为$\sum\limits_{\substack{G\subseteq\Pi\\G\neq\varnothing}}|N(G)|$。$\Pi$的非空子集共有$2^{|\Pi|}-1$个,因此正则使用者的总数对属性的平均值为$\sum\limits_{\substack{G\subseteq\Pi\\G\neq\varnothing}}|N(G)|\times(2^{|\Pi|}-1)^{-1}$,此值再对对象个数$|\mathbb{U}|$平均,$\frac{1}{|\mathbb{U}|}\sum\limits_{\substack{G\subseteq\Pi\\G\neq\varnothing}}|N(G)|\times(2^{|\Pi|}-1)^{-1}$,将所得结果圆整到0～100之间:$\frac{1}{|\mathbb{U}|}\sum\limits_{\substack{G\subseteq\Pi\\G\neq\varnothing}}|N(G)|\times(2^{|\Pi|}-1)^{-1}\times100$就称为是大数据的"正则价值数值"。

同理,定义"必然价值数值"为$\frac{1}{|\mathbb{U}|}\sum\limits_{\substack{G\subseteq\Pi\\G\neq\varnothing}}|C(G)|\times(2^{|\Pi|}-1)^{-1}\times100$

"相关价值数值"为$\frac{1}{|\mathbb{U}|}\sum\limits_{\substack{G\subseteq\Pi\\G\neq\varnothing}}|R(G)|\times(2^{|\Pi|}-1)^{-1}\times100$

由此我们有以下定义。

定义6 设$\mathbb{D}$是大数据,Π是$\mathbb{D}$的一个粒空间。$\mathbb{U}$是选取的使用大数据的一些代表项目的集合。$I_\Pi\subseteq\mathbb{U}\times\Pi$为大数据$\mathbb{D}$对应$\Pi$的"使用关系",$\mathbb{K}=(\mathbb{U},\Pi,I_\Pi)$是一个形式背景,则大数据的正则价值数值为:

$$value_N(\mathbb{D},\Pi)=\frac{1}{|\mathbb{U}|}\sum_{\substack{G\subseteq\Pi\\G\neq\varnothing}}|N(G)|\times(2^{|\Pi|}-1)^{-1}\times100$$

必然价值数值为:

$$value_C(\mathbb{D},\Pi)=\frac{1}{|\mathbb{U}|}\sum_{\substack{G\subseteq\Pi\\G\neq\varnothing}}|C(G)|\times(2^{|\Pi|}-1)^{-1}\times100$$

相关价值数值为:

$$value_R(\mathbb{D},\Pi)=\frac{1}{|\mathbb{U}|}\sum_{\substack{G\subseteq\Pi\\G\neq\varnothing}}|R(G)|\times(2^{|\Pi|}-1)^{-1}\times100$$

4 实际应用

例5 一个远程医疗的大数据$\mathbb{D}$的粒度树(只给出节点名称)如图2所示。

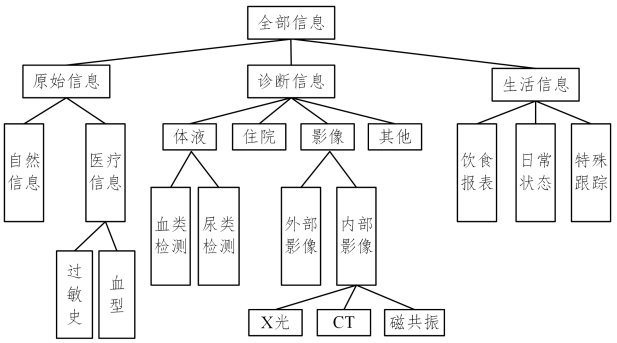

图 2 远程医疗的大数据$\mathbb{D}$的粒度树

Fig. 2 Granularity tree of big data $\mathbb{D}$ in telemedicine

令 $\Pi_0=\{$全部信息$\}$，$\Pi_1=\{$原始信息，诊断信息，生活信息$\}$，$\Pi_2=\{$原始信息，体液，住院，影像，其他，生活信息$\}$，$\Pi_3=\{$原始信息，体液，住院，外部影像，内部影像，其他，生活信息$\}$，$\Pi_4=\{$原始信息，体液，住院，外部影像，X 光，CT，磁共振，其他，生活信息$\}$，$\Pi_5=\{$原始信息，血类检测，尿类检测，住院，外部影像，X 光，CT，磁共振，其他，生活信息$\}$，$\Pi_6=\{$原始信息，诊断信息，饮食报表，日常状态，特殊跟踪$\}$，则它们都是粒度树的“视角”。我们分别考虑 Π_3，Π_5（着重诊断的一些视角）和 Π_6（着重生活的一些视角）来计算“价值数值”。

1）Π_3：使用关系$\mathbb{I}_{\Pi_3}$如表 2 所列，其中 a 是原始信息，b 是体液，c 是住院，d 是外部影像，e 是内部影像，f 是其他，s 是生活信息。选择 9 个使用项目：u_1 是医疗单位，u_2 是医疗设备生产企业，u_3 是养老院，u_4 是药品研究院，u_5 是药品销售单位，u_6 是食品生产部门，u_7 是服装生产部门，u_8 是保险公司，u_9 是陪护、家政服务公司。计算得出：$value_N(\mathbb{D},\Pi_3)=28.5880$，$value_C(\mathbb{D},\Pi_3)=30.6513$，$value_R(\mathbb{D},\Pi_3)=93.7883$。

表 2 粒空间 Π_3 的使用关系

Table 2 Usage relationship of grain space Π_3

$\mathbb{I}_{\Pi_3}$	a	b	c	d	e	f	s
u_1	×		×	×	×	×	×
u_2	×	×		×	×		×
u_3	×	×		×	×	×	×
u_4	×		×				×
u_5	×		×				×
u_6	×		×				×
u_7	×	×		×			
u_8	×		×		×	×	×
u_9	×			×	×	×	×

2）Π_5：使用关系$\mathbb{I}_{\Pi_5}$如表 3 所列，其中 a 是原始信息，b_1 是血类检测，b_2 是尿类检测，c 是住院，d 是外部影像，e_1 是 X 光，e_2 是 CT，e_3 是磁共振，f 是其他，s 是生活信息。使用项目 $u_1,\cdots,u_9$ 同 1）。计算得出：$value_N(\mathbb{D},\Pi_5)=20.5949$，$value_C(\mathbb{D},\Pi_5)=26.9157$，$value_R(\mathbb{D},\Pi_5)=93.9937$。

表 3 粒空间 Π_5 的使用关系

Table 3 Usage relationship of grain space Π_5

$\mathbb{I}_{\Pi_5}$	a	b_1	b_2	c	d	e_1	e_2	e_3	f	s
u_1	×			×	×	×	×	×	×	×
u_2	×	×	×		×	×	×	×		×
u_3	×	×	×		×	×	×	×	×	×
u_4	×			×						×
u_5	×			×						×
u_6	×			×						×
u_7	×		×		×					
u_8	×			×		×	×	×	×	×
u_9	×				×	×			×	×

3）Π_6：使用关系$\mathbb{I}_{\Pi_6}$如表 4 所列，其中 a 是原始信息，z 是诊断信息，g 是饮食报表，h 是日常状态，i 是特殊跟踪。使用项目 $u_1,\cdots,u_9$ 同 1）。计算得出：$value_N(\mathbb{D},\Pi_6)=48.8095$，$value_C(\mathbb{D},\Pi_6)=37.5000$，$value_R(\mathbb{D},\Pi_6)=93.5484$。

表 4 粒空间 Π_6 的使用关系

Table 4 Usage relationship of grain space Π_6

$\mathbb{I}_{\Pi_6}$	a	z	g	h	i
u_1	×	×	×		×
u_2	×	×		×	
u_3	×	×	×	×	×
u_4	×	×		×	×
u_5	×	×	×	×	
u_6	×	×	×	×	
u_7	×	×			
u_8	×	×			×
u_9	×	×	×		×

例 6 苏州市所辖 6 区 4 市：姑苏区、吴中区、吴江区、高新区、工业园区、相城区、昆山市、太仓市、常熟市、张家港市的市政大数据$\mathbb{D}$的粒度树（只给出节点名称）如图 3 所示。

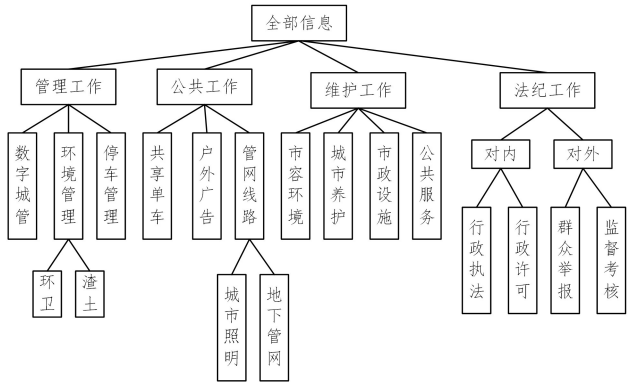

图 3 苏州城管的大数据$\mathbb{D}$的粒度树

Fig. 3 Granularity tree of big data $\mathbb{D}$ of Suzhou Urban Management

令 $\Pi_0=\{$全部信息$\}$，$\Pi_1=\{$管理工作，公共工作，维护工作，法纪工作$\}$，$\Pi_2=\{$数字城管，环境管理，停车管理，公共工作，维护工作，法纪工作$\}$，$\Pi_3=\{$数字城管，环卫，渣土，停车管理，公共工作，维护工作，法纪工作$\}$，$\Pi_4=\{$管理工作，公共工作，维护工作，对内，对外$\}$，$\Pi_5=\{$管理工作，公共工作，维护工作，行政执法，行政许可，群众举报，监督考核$\}$，则 $\Pi_3\leqslant\Pi_2\leqslant\Pi_1\leqslant\Pi_0$ 是着重管理工作的一些视角，$\Pi_5\leqslant\Pi_4\leqslant\Pi_1\leqslant\Pi_0$ 是着重执法工作的一些视角。我们考虑 Π_3，Π_5 来计算价值数值。

1）Π_3：使用关系$\mathbb{I}_{\Pi_3}$如表 5 所列。

表 5 粒空间 Π_3 的使用关系

Table 5 Usage relationship of grain space Π_3

$\mathbb{I}_{\Pi_3}$	a	b	c	d	e	f	g
u_1	×						×
u_2				×	×	×	×
u_3	×					×	×
u_4					×		
u_5		×	×	×	×	×	
u_6				×	×		×
u_7	×	×		×	×	×	×
u_8				×			×

表 5 中，a 是数字城管，b 是环卫，c 是渣土，d 是停车管理，e 是公共工作，f 是维护工作，g 是法纪工作。选择 8 个使用项目：u_1 数据预警、u_2 数据监管、u_3 数据决策、u_4 数据

服务、u_5 环境问题、u_6 执法问题、u_7 社区问题、u_8 治安问题。计算得出：$value_N(\mathbb{D},\Pi_3)=23.6607$，$value_C(\mathbb{D},\Pi_3)=21.7105$，$value_R(\mathbb{D},\Pi_3)=87.7953$。

2）Π_5：使用关系$\mathbb{I}_{\Pi_5}$如表 6 所列，其中 h 是管理工作，e 是公共工作，f 是维护工作，i 是行政执法，j 是行政许可，k 是群众举报，l 是监督考核。选择 8 个使用项目：$u_1,\cdots,u_8$ 同 1）。计算得出：$value_N(\mathbb{D},\Pi_5)=20.3125$，$value_C(\mathbb{D},\Pi_5)=24.7154$，$value_R(\mathbb{D},\Pi_5)=83.6614$。

表 6　粒空间 Π_5 的使用关系

Table 6　Usage relationship of grain space Π_5

$\mathbb{I}_{\Pi_5}$	h	e	f	i	j	k	l
u_1	×			×		×	×
u_2	×	×	×			×	×
u_3	×		×		×		
u_4		×					
u_5	×	×	×		×		
u_6	×	×	×	×			
u_7	×	×				×	
u_8	×			×		×	

例 7　辽宁省鞍山市各高等院校的大数据$\mathbb{D}$的粒度树（只给出节点名称）如图 4 所示。

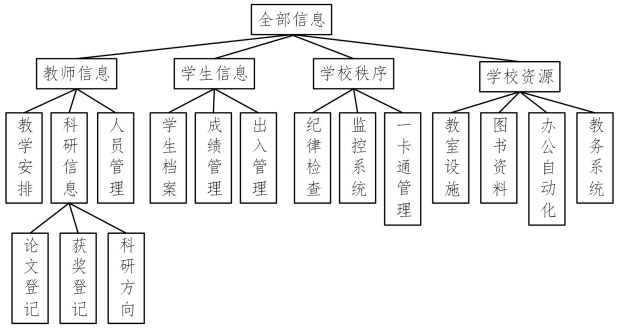

图 4　辽宁省鞍山市各高等院校的大数据$\mathbb{D}$的粒度树

Fig. 4　Granularity tree of big data $\mathbb{D}$ of colleges and universities in Anshan, Liaoning Province

令 $\Pi_0=\{$全部信息$\}$，$\Pi_1=\{$教师信息，学生信息，学校秩序，学校资源$\}$，$\Pi_2=\{$教学安排，科研信息，人事管理，学生信息，学校秩序，学校资源$\}$，$\Pi_3=\{$教学安排，论文登记，获奖登记，科研方向，人事管理，学生信息，学校秩序，学校资源$\}$，$\Pi_4=\{$教学安排，论文登记，获奖登记，科研方向，人事管理，学生档案，成绩管理，出入管理，学校秩序，学校资源$\}$，$\Pi_5=\{$教学安排，论文登记，获奖登记，科研方向，人事管理，学生档案，成绩管理，出入管理，纪律检查，监控系统，一卡通管理，教室设施，图书资料，办公自动化，教务系统$\}$。$\Pi_5\leqslant\Pi_4\leqslant\Pi_3\leqslant\Pi_2\leqslant\Pi_1\leqslant\Pi_0$，是一些视角。取两个视角 Π_2（着重教师）及 Π_5（考虑教师与学生）来计算价值数值。

1）Π_2：使用关系$\mathbb{I}_{\Pi_2}$如表 7 所列。

表 7 中，a 是教学安排，b 是科研信息，c 是人事管理，d 是学生信息，e 是学校秩序，f 是学校资源。选择 10 个使用项目：u_1 是心理孤独学生分析、u_2 是贫困学生分析、u_3 是学生爱好分析、u_4 是学生行为分析、u_5 是个性化学习分析、u_6 是消费行为分析、u_7 是教师结构分析、u_8 是教学质量分析、u_9 是科研方向分析、u_{10}是图书推荐分析。计算得出：$value_N(\mathbb{D},\Pi_2)=24.1667$，$value_C(\mathbb{D},\Pi_2)=34.1176$，$value_R(\mathbb{D},\Pi_2)=83.1746$。

2）Π_5：使用关系$\mathbb{I}_{\Pi_5}$如表 8 所列。

表 7　粒空间 Π_2 的使用关系

Table 7　Usage relationship of grain space Π_2

$\mathbb{I}_{\Pi_2}$	a	b	c	d	e	f
u_1				×	×	
u_2				×	×	
u_3		×		×		×
u_4		×		×	×	
u_5		×		×		×
u_6				×	×	
u_7	×	×	×			×
u_8	×					×
u_9		×	×			×
u_{10}	×	×	×			×

表 8　粒空间 Π_5 的使用关系

Table 8　Usage relationship of grain space Π_5

$\mathbb{I}_{\Pi_5}$	a	g	h	i	c	j	k	l	m	n	o	p	q	r
u_1						×	×	×	×	×	×			
u_2						×		×		×	×			
u_3		×	×			×	×						×	
u_4		×	×			×	×	×	×		×			
u_5		×	×			×	×						×	
u_6						×		×		×	×			
u_7	×	×	×	×	×								×	
u_8	×										×	×	×	×
u_9		×	×	×		×						×		×
u_{10}	×	×	×	×		×						×	×	×

表 8 中，a 是教学安排，g 是论文登记，h 是获奖登记，i 是科研方向，c 是人事管理，j 是学生档案，k 是成绩管理，l 是出入管理，m 是纪律检查，n 是监控系统，o 是一卡通管理，p 是教室设施，q 是图书资料，r 是办公自动化，s 是教务系统。选择 10 个使用项目：$u_1\cdots u_{10}$同 1）。计算得出：$value_N(\mathbb{D},\Pi_5)=15.5233$，$value_C(\mathbb{D},\Pi_5)=20.5232$，$value_R(\mathbb{D},\Pi_5)=97.0324$。

结束语　大数据蕴藏的巨大的科学研究价值、公共管理与服务价值、商业价值以及支持科学决策的价值等逐渐被人们所认知。而且为满足大数据交易“定价”的需求、失窃“索赔”的需求、删除“取舍”的需求、制订“规范”的需求等，人们不仅需要对价值进行论述、引用、挖掘、拆分、合并、分析等，而且需要用一个基准的“价值数值”来衡量它核心价值的大小。其他的外在价值都在这个核心价值的基础上附加了相关领域的系数或相关领域的附加值而产生的不同领域价值。人们逐渐认识到大数据的“价值数值”并不是“信息量”，它与概率无关，与使用有关。人们还认识到，各种各样的各领域价值都是由使用决定的“价值”的外在表现。使用决定的“价值”是大数据最基本最核心的价值，而且实践表明使用可以有“正则使用”“必然使用”“相关使用”，对应了“正则价值”“必然价值”“相关价值”这 3 种价值。其中：“正则价值”是经常应用的价值，“必然价值”是强调各粒的集合独立使用时的价值，“相关价值”是强调各粒的集合联合使用时的价值，这 3 种价值相辅相成，配合使用，形成了较全面的大数据核心基本“价值数值”谱系。

本文探讨了基于“使用”的大数据价值计算以及多个实际应用。作为大数据价值计算领域的新的方向探索，引玉之砖，供大家参考。

参 考 文 献

[1]　TOFFLER A. The third wave[M]. New York: Bantam Books, 1981: 167-168.

[2]　MAYER-SCHÖNBERGER V, CUKIER K. Big data: a revolu-

tion that will transform how we live, work, and think [M]. New York: Houghton Mifflin Harcourt, 2013: 7, 47.
[3] BORGATTI S P, MEHRA A, BRASS D J, et al. Net-work analysis in the social sciences[J]. Science, 2009, 323(5916): 892-895.
[4] PORTER M E. Competitive advantage: creating and sustaining superior performanc[M]. New York: Free Press, 1985: 2-4.
[5] XU Z B, FENG Z Y, GUO X H, et al. Frontier issues of management and decision making driven by Big data[J]. Management World, 2014(11): 158-163.
[6] MAYER-SCHÖNBERGER V, CUKIER K. Bigdata: a revolution that will transform how we live, work, and think[M]. Hought on Mifflin Harcourt, 2013.
[7] WANG N, LI T Z, CAO S Y, et al. Research on the formation path of big data value: a biological analogy [J]. China Science and Technology Forum, 2020, 294(10): 142-149.
[8] MA W S, HOU X L, WANG H B, et al. Big data value calculation method based on granular computing and usage times [J]. Journal of Liaoning University of Science and Technology, 2021, 44(3): 196-207.
[9] Global big data exchange[OL]. https://www.gzdex.com.cn/.
[10] Donghu Big Data[OL]. http://www.chinadatatrading.com/.
[11] Beijing International Data Exchange [OL]. https://www.bjidex.com/.
[12] Dawex: Sell, buy and share data [OL]. https://www.dawex.com/en/.
[13] Xignite[OL]. https://www.xignite.com/.
[14] World Quant [OL]. https://www.worldquant.com/data-exchange/.
[15] JIANG D, YUAN Y, ZHANG X W, et al. Summary of data pricing and trading research [J/OL]. Journal of Software: 1-29. [2013-03-04]. http://221.203.21.203: 8001/rwt/CNKI/https/MSYXTLUQPJUB/10.13328/j.cnki.jos.006751.
[16] FAMA E F, FRENCH K R. The value premium and the CAPM [J]. Journal of Finance, 2006, 61(5): 2163-2185.
[17] HANSEN L P, SARGENT T J. Formulating and estimating dynamic linear rational expectations models[J]. Australian Journal of Otolaryngology, 1980, 2(1): 7-46.
[18] SHILLER R J. Theuse of volatility measures in assessing market efficiency[J]. Journal of Finance, 1981, 36(2): 291-304.
[19] DAVID T. Valuing intellectual property assets[J]. Computer and Internet Lawyer, 2002, 19(2): 1-8.
[20] CHIU Y J, CHEN Y W. Using AHP in patent valuation[J]. Mathematical and Computer Modelling, 2007, 46(7/8): 1054-1062.
[21] CHEN Z Z, WANG H Z, XIONG F, et al. Pricing strategy and method of big data auction [J]. Journal of China University of Science and Technology, 2018, 48(6): 486-494.
[22] LIN G T R, TANG J Y H. Appraising intangible assets from the viewpoint of value drivers[J]. Journal of Business Ethics, 2009, 88(4): 679-689.
[23] JORGE M, ISMAEL C, BIBIANO R, et al. A data quality in use model for big data[J]. Future Generation Computer Systems, 2016, 63: 123-130.
[24] NIYATO D, ALSHEIKH M A, WANG P, et al. Market model and optimal pricing scheme of big data and internet of things (IoT)[C]// IEEE International Conference on Communications (ICC). IEEE, Kuala Lumpur, 2016.
[25] HOU X, SHEN J. Construction and analysis of big data value model in relay innovation[J]. Journal of University of Science and Technology Liaoning, 2019, 42(2): 149-153, 160.
[26] VANCE A. Start-up goes after big data with Hadoophelper [OL]. http://bits.blogs.nytimes.com/2010/04/22/start-up-goes-after-big-data-with-hadoophelper/?dbk.
[27] CHEN C L P, ZHANG C Y. Data-intensive applications, challenges, techniques and technologies: Asurvey on bigdata[J]. Information Sciences, 2014, 275: 314-347.
[28] XU J, WANG G Y, YU H. Review of Big Data Processing Based on Granular Computing [J]. Chinese Journal of Computers, 2015, 38(8): 1497-1517.
[29] YAO Y Y. Information granulation and rough set approximation [J]. International Journal of Intelligent Systems, 2001, 16(1): 87-104.
[30] HUA Q Y. Granulation Mechanism and Data Modeling for Complex Data[D]. Taiyuan: School of Computer and Information Technology, Shanxi University, 2011.
[31] ZHANG Y P, ZHANG L, WU T. The Representation of Different Granular Worlds: A Quotient Space[J]. Chinese Journal of Computers, 2004, 27(3): 328-333.
[32] ZADEHL A. Toward a theory of fuzzy information granulation and its centrality in human reasoning and fuzzy logic[J]. Fuzzy Sets and Systems, 1997, 90: 111-127.
[33] MA H Y, WANG G Y, ZHANG Q H, et al. Multi-granularity color image segmentation based on cloud model[J]. Computer Engineering, 2012, 38(20): 184-187.
[34] QIN K, LI D Y, XU K. Imagesegmentationbasedon cloudmodel [J]. Journal of Geomatics, 2006, 31(5): 3-5.
[35] LIU C Y, FENG M, DAI X J, LI D Y. A New Algorithm of Backward Cloud [J]. Journal of System Simulation, 2004, 16(11): 2417-2420.
[36] ZHANG L, ZHANG B. Theory of fuzzy quotient space(methods of fuzzy granular computing) [J]. Journal of Software, 2003, 14(4): 770-776.
[37] LIN F, COHEN W W. Power iteration clustering [C]// Proceedings of 2010 International Conference on Machine Learning (ICML). Haifa, ISRAEL, 2010: 655-662.
[38] YAN W, BRAHMAKSHATRIYA U, XUE Y, et al. p-PIC: Parallel power iteration clustering for big data[J]. Journal of Parallel and Distributed Computing, 2013, 73(3): 352-359.
[39] WANG G Y, LI D Y, YAO Y Y, et al. Cloud Model and Granular Computing[M]. Beijing: Science Press, 2012.

MA Wensheng, born in 1971, Ph.D candidate. His main research interest is big data application.

HOU Xilin, born in 1960, Ph.D, professor. His main research interest include big data application, enterprise innovation system.