“大数据—小数据”问题：以小见大的洞察*

□陈国青　张　瑾　王　聪　卫　强　郭迅华

摘要：在数字经济时代，基于数据的管理决策变成了新的决策范式，并为管理实践赋予了创新源动力。把握决策范式转变机遇的一个重点是如何利用大数据这种新兴的生产要素，通过大数据赋能以提升管理决策的水平。然而，当大数据可能为决策提供全局视图的同时，在很多应用场景中，出于数据的可获性以及成本、时间的限制，乃至人们的认知能力、阅读心理等相关因素的影响，人们所面对和能够直接处理的数据往往是有限的、部分的(即小数据)。针对这种决策信息的不对称性，本文基于一系列的研究，围绕如何通过小数据反映大数据语义内容这一核心，提出了“大数据—小数据”问题。进而，从语义反映的“代表性”、“一致性”、“多样性”的视角出发，系统性地梳理和阐释了这一问题的科学内涵、求解路径、实践意义和管理启示。通过“大数据—小数据”问题提炼而成的以小见大的洞察，可以为数据驱动的决策和创新性价值创造开拓广阔的空间。

关键词：“大数据—小数据”问题　语义反映　管理决策
DOI:10.19744/j.cnki.11-1235/f.2021.0028

一、引言

随着大数据、人工智能、物联网、5G、云计算、区块链等新兴科技与社会经济、产业生态、企业管理、用户生活的深入融合，数字经济正逐渐成为一种重要的经济形态。尤其是在新冠肺炎疫情给经济发展带来的巨大冲击和不确定性背景下，数字经济呈现了巨大的发展韧性，从一个特定角度诠释了“百年未有之大变局”的独特内涵。数字经济发展的核心动力为科技创新。在数字经济时代，科技的快速迭代使经济、社会、生活等各个领域发生了日新月异的变革，经济社会的形态、企业管理的场景、人们生活的方式都正在或即将在数字空间进行着重构。数字空间中的数字场景产生了大量的数据，例如，虚拟化生产中的数字组装日志、智能交通中的参与者时空轨迹、用户直连制造(C2M)中的需求订单，信息服务平台上的企业组织流程、电子商务活动中的消费者评论、社交媒体上的多媒体内容动态等。据IDC研究显示，到2025年全球数据圈将扩展至175ZB(1ZB等于1万亿GB)，相当于2016年所产生16.1ZB数据的10倍(Reinsel et al.，2019)。

可以看到，基于数据的决策逐渐成为研究和应用的主流(徐宗本等，2014)，变成触及产业与经济发展的基础性机制以及经济与管理决策的基本形式，也引发了各国政府推出不同的研究规划以应对随之而来的深层次挑战。例如，欧盟在2014年发布的《数字驱动经济战略》；美国在2016年推出的《联邦大数据研究与开发战略计划》；我国2015年党的十八届五中全会提出实施“国家大数据战略”，国务院印发《促进大数据发展行动纲要》(国发〔2015〕50

*本研究得到国家自然科学基金重大研究计划项目“大数据驱动的管理与决策研究重大研究计划战略研究项目”(基金号：91846000)、国家自然科学基金面上项目“面向在线知识付费的智能知识服务理论与方法研究”(基金号：72072177)、国家自然科学基金面上项目“基于在线问答社区的智能信息服务方法及其用户决策影响研究”(基金号：71772177)以及国家自然科学基金重大研究计划项目“面向管理决策大数据分析的理论与方法”(基金号：92046021)的资助。张瑾为本文通讯作者。

号）；2019年第十三届全国人民代表大会第二次会议强调深化大数据、人工智能等研发应用，为制造业转型升级赋能，壮大数字经济；2020年中共中央和国务院发布的《中共中央、国务院关于构建更加完善的要素市场化配置体制机制的意见》中，将数据与土地、劳动力、资本、技术等传统要素并列为生产要素之一。

数字经济的发展使得具有超规模、富媒体、低密度、流信息特征的大数据（冯芷艳等，2013；Hilbert and Lopez，2011）及其应用成为了赋能和创新的重要源动力。在此环境下，领域情境、决策主体、理念假设、方法流程等决策要素受到冲击，导致决策范式正在发生着深刻转变，催生了新型决策范式（即大数据决策范式）（陈国青等，2020）。从决策诉求的角度看，大数据驱动的管理决策寻求对于多维因素的关联模式和因果关系的揭示，以期获得决策情境的全局视图（陈国青等，2018）。这就要求决策者能够获得对于大数据决策场景全貌的洞察。然而，在现实中虽然获得大数据（数据集合全体）成为可能，但是在很多应用场景中，出于数据的可获性及其成本、时间、乃至人们的认知能力、阅读心理等相关因素影响，人们面对或者能够直接处理的数据往往是有限的、部分的。也就是说，人们的许多决策是基于小数据（数据集合全体的子集）的。例如，消费者只有有限的时间和耐心阅读全部产品评论中的一小部分；关键词搜索者只可能浏览海量查询结果的前两页条目；企业管理者只能利用有限的时间和精力从所有企业博客或微信群发帖中看到部分的内容；财务审计师囿于时间和成本只能从海量的内外部数据中阅读有限的报表和文本信息；政府决策者限于能力和时间可能只了解到所有受众诉求和舆情中的局部细节；等等。这里，值得一提的是，虽然通过机器学习工具的广泛应用，生成数据集合全体的概括汇总和特征表示等信息成为可能（如文本概要、评论极性、统计均值、话题标签等）并发挥着积极的决策支持作用，但是人们对于心理和人格的刻画、对于个体和组织的了解、对于事件和活动的诠释、对于模式和因果的解构等通常需要具象的、丰富的、细节的、情景化的体验和感知。换句话说，实例子集是认识和反映全集的不可忽视的重要方面，在决策中可以通过实例子集帮助人们以局部看整体，达到见微知著的效果。

上述讨论引出了一个重要问题，即基于小数据的决策与基于大数据的决策在效果上取决于小数据与大数据之间的信息不对称程度。在此，我们将该问题称为“大数据—小数据”问题。“大数据”是指相关数据全体，“小数据”是相关数据全体的一个子集，小数据通过部分数据反映大数据在特定方面的语义（semantics）内容。从集合概念的角度出发，作为相关数据全体的“大数据”对应着“大集合”，而作为相关数据全体之子集的“小数据”对应着“小集合”。在这个意义上讲，“大数据—小数据”问题也可以表达为“大集合—小集合”问题。进而，“大数据—小数据”问题可以表示为小数据集合反映大数据集合的问题。这里的“反映”是指语义反映，即小数据所携载的语义与大数据所携载的语义之间的异同关系（如距离或相近性）。如果给定大数据集合（即大集合），对于“大数据—小数据”问题的求解则是寻求一个小数据集合（即大数据集合的子集—小集合），使得小数据集合的语义与大数据集合的语义尽可能相近。这里，根据应用情境的不同，对于小数据集合的规模通常有特定约束。一般说来，小数据集合的规模远远小于大数据集合的规模。为方便起见，在本文后面的讨论中，“大数据—小数据”问题也简称为B-S问题（The Bigdata-Smalldata Problem）。

二、“大数据—小数据”问题的形式化定义

在“大数据—小数据”问题中，大数据集合（大集合）和小数据集合（小集合）分别用D和D'表示，D'是D的子集。具体说来，设$\mathbb{A}=\{A_1,A_2,\cdots\cdots,A_q\}$是$D$的属性集合，$U_j$是属性$A_j$的论域（$j=1,\cdots\cdots,q$）。对于属性向量（$A_1,A_2,\cdots\cdots,A_q$），其对应的论域空间为$\mathbb{U}=U_1\times U_2\times\cdots\cdots\times U_q$。给定大集合$D=\{d_1,d_2,\cdots\cdots,d_n\}$和整数k（$k\ll n$），“大数据—小数据”问题（B-S问题）旨在寻求获得一个小集合$D'=\{d'_1,d'_2,\cdots\cdots,d'_k\}$，$D'\subset D$，使得小集合的语义$s(D')$尽可能接近大集合的语义$s(D)$，即：

$$\max_{D'\subset D}(1-(s(D)\ominus s(D'))) \tag{1}$$

其中，$s(D)$是从$\mathbb{U}$的高阶论域空间$\mathbb{U}^p$到语义空间$\mathbb{V}$的映射（即：$\mathbb{U}^{p'}\rightarrow\mathbb{V}$），$p\geq 1$；$s(D')$是从$\mathbb{U}$的高阶论域空间$\mathbb{U}^{p'}$到语义空间$\mathbb{V}$的映射（即：$\mathbb{U}^{p'}\rightarrow\mathbb{V}$），$p\geq p'\geq 1$；$\ominus$为超减法运算，是从$\mathbb{V}\times\mathbb{V}$到[0,1]的映射（即：$\mathbb{V}\times\mathbb{V}\rightarrow[0,1]$）。也就是说，语义是通过映射关系及其映射到$\mathbb{V}$上的结果（像）表示的。超减法运算优化的核心是度量

小集合D'与大集合D之间的语义偏差，语义偏差越小则说明小集合的语义能够“反映”大集合的语义。

这里，对于任一数据集合X，其语义总体$S(X)$是X中元素的属性特征或X元素关系的含义集合表示。在不同的情景、视角和认识条件下，X的语义有着不同的体现，反映X的数据在相关属性上的取值及其模式（如结构、类别、关系等）。如$s(X)$就是语义的某一特定体现，即$s(X)\in S(X)$。在B-S问题中，$X=D$时，有$s(D)\in S(D)$。同样，$X=D'$时，有$s(D')\in S(D')$。以$X=D$为例，D的元素间相似关系体现了D的一个语义，是$\mathbb{U}\times\mathbb{U}$到[0,1]的映射，即$s(D):\mathbb{U}^p\to[0,1]$，$p=2$。再如，$D$的均值也体现了$D$的一个语义，即$s(D):\mathbb{U}^p\to\mathbb{U}$，$p=|D|=n$；类似地，$D'$的均值亦体现$D'$的一个语义$s(D'):\mathbb{U}^{p'}\to\mathbb{U}$，$p'=|D'|=k$。

从表达的层次来看，语义可以分为显式语义和隐式语义。显式的语义比较直接，通常可以直接观察到，而隐式的语义则可能需要进一步的揭示或者表达。例如，最直观的一种显式语义表述的就是D的一种存在或者是D中元素d的一种存在。换句话说，d在属性$A_j(j=1,2,\cdots\cdots,q)$上的取值$(d(A_1),d(A_2),\cdots\cdots,d(A_q))$刻画了$d$的一种存在，即：

$$d=(d(A_1),d(A_2),\cdots\cdots,d(A_q)) \tag{2}$$

$$D=\left\{(d_i(A_1),d_i(A_2),\cdots\cdots,d_i(A_q)) \mid d_i(A_j)\in U_j;i=1,2,\cdots\cdots,n;j=1,2,\cdots\cdots,q\right\} \tag{3}$$

举例说来，假设公司的客户关系管理（CRM）系统中记录的数据如表1所示，其中包含所有客户的基本注册信息以及购物信息。表1中的每一条记录通过在属性上的取值构成了一个存在实例，体现出其关于这些属性特征的语义内涵。所有这些记录数据则反映了客户全体的存在以及相关属性上的语义。给定了表1，其中的记录在属性上的取值就是语义的一种显性表示。

进一步，如果决策过程需要了解所有客户的年龄构成，那么则需要对表1中的所有记录在年龄属性上的取值情况进行分析，得到如图1所示的客户年龄分布模式。这里，客户年龄分布是客户记录在“年龄”属性上的取值模式，其作为一种语义并没有显式地呈现在表1中，而是隐式地体现在年龄取值关系中。其它隐式语义还包括属性取值的统计特征、顺序特征、类别特征、结构特征等，主要通过属性取值关系模式体现（包括上面相似关系和均值的例子）。隐式语义特别是复杂模式常常不易直接观察得到，需要通过数据分析手段进行挖掘予以揭示。

此外，在大数据环境下，数据呈现多模态、富媒体特点。表1作为CRM的原始记录集合，属于结构化数据，而现实中还存在着更大规模的非结构化数据，如文本、图像等。对于这类非结构化数据，当前的主流技术应用是通过表示学习来获得非结构化数据的特征向量集合，即通过数据在这些属性特征上的取值来获得基本的语义内容。例如，产品评论是一类非结构化的自由格式的文本，通过特征提取方法可以将其转换成在特征向量上的取值，以结构化的形式来表示这些文本的基本语义内容。这里，产品评论是观测到的、作为显式存在的原始数据实例，而其在相关属性（特征）上的取值则是揭示出的一种隐式语义内容。进一步，这些评论数据在属性特征上的取值模式也是一种需要深层次揭示的语义内容。通常，文本、图像等富媒体数据的语义表示相对复杂，需要借助数据挖掘和机器学习技术与应用的发展成果。

表1 CRM系统的客户数据示例

ID	姓名	手机号	性别	年龄	上次购物时间	……
1	Tom	138####1235	Male	24	2020-12-01 17:34	……
2	Jane	138####3336	Female	40	2020-11-30 12:10	……
3	Marry	184####8789	Female	45	2020-11-27 08:45	……
4	Claudia	139####0109	Female	32	2020-10-19 21:13	……
5	Justin	135####4662	Male	19	2020-12-02 19:34	……
……	……	……	……	……	……	……

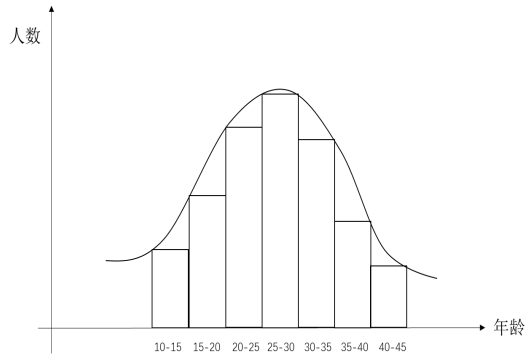

图1 CRM系统中的客户年龄分布情况

综上所述，大数据集合的语义在不同的情境、视角和认识情形下存在不同的表示，并具有显式和隐式特点。下面我们将从语义反映的角度出发，讨论“大数据—小数据”问题的3种典型类型（即代表性语义反映、一致性语义反映和多样性语义反映）及其可能的组合形式，并结合作者团队近年来的系列研究创新，阐述相关的领域情境、概念内涵、问题建模、求解路径

以及管理决策意义。

三、代表性语义反映

在决策场景中，人们常常需要了解大数据中的不同内容或者不同数据元素的存在，即希望通过数据实例这种显式语义来形成对于数据内容的具体印象和直观认识。代表性语义反映是“大数据—小数据”问题的一种类型，旨在从上述显式语义的视角获得一个数据实例的小集合，以求尽可能地反映数据实例全体的内容语义。例如，当需要从所有搜索结果中浏览一小部分条目时，当需要从所有企业博文中读取一小部分文章时，当需要从所有客户反馈中阅看一小部分评论时，当需要从所有舆情专报中审视一小部分报告时，……，林林总总，人们遇到了依据小数据（子集）认识全局进行决策的情形。此时，小数据通过部分具体的数据实例内容来反映大数据的数据实例内容，这种“反映”称为代表性语义反映。

代表性语义反映的概念内涵主要体现的是大数据D与小数据D'在元素内容上的对应关系，这种对应关系可以通过元素实例之间的相似关系来度量，例如，对于$d\in D$，$d'\in D'$，二者相似度$Sim(d,d')$测量了二者之间的内容异同，反映了d代表（或覆盖）d'内容的程度，也反映了d'代表（或覆盖）d内容的程度。这里，相似度测度$Sim()$通常设定具有自反性和对称性，且$Sim(d,d')\in[0,1]$。换句话说，代表性的含义是通过元素内容间的相似关系体现的，也体现了小数据D'在内容上对大数据D的覆盖情况，即代表性具有内容覆盖（content coverage）的意味。

需要指出的是，由于代表性是通过显式数据实例之间的相似关系来刻画的，那么数据相似度（或差异性）测量决定着大数据与小数据间的“反映”关系，也影响着生成小数据集合的思路逻辑。

如图2所示，首先，D中d被D'中d'内容上代表的程度通过相似度$Sim(d',d)$表示。则从子集D'的角度，D'中各元素与D中d相似度最大的那个元素d'^{*}与d的相似程度视为整个D'代表d的程度，即：$Sim(D',d)=\max_{d'\in D'}Sim(d',d)$，这也是对整个$D'$从内容上代表$d$的一个表达。进而，$D$被$D'$代表的势（cardinality）是所有$D$中$d$被$D'$代表的程度之和，即：$\sum_{d\in D}Sim(D',d)$，为$D'$在内容上对于$D$的覆盖“量”，可视作$s(D')$。即：

$$s(D')=\sum_{d\in D}Sim(D',d) \tag{4}$$

而D的势（D的总量$|D|$）可视作$s(D)$。若将语义转换为0-1之间的量（即占总量的比例，或内容覆盖程度），则$s(D)$为1，且：

$$s(D')=\frac{1}{|D|}\sum_{d\in D}Sim(D',d) \tag{5}$$

这样，给定D和k，代表性语义反映的求解问题可以表达为：

$$max_{D'\subset D,|D'|=k}\left[1-\left(\left|s(D')\ominus s(D)\right|\right)\right]=max_{D'\subset D,|D'|=k}\frac{1}{|D|}\sum_{d\in D}Sim(D',d) \tag{6}$$

虽然从前面的讨论中可以看到该问题的提出和求解对于决策支持具有重要意义，但是其求解难度是一个挑战，因为此问题是一个NP-难问题（Ma et al.，2017）。这对于方法创新提出了高要求，特别是需要设计出新颖有效的优化策略和启发式方法。再者，如何确保启发式方法的寻优效果是此问题求解的另一个难点。鉴于此，相关研究证明了该问题具有一个重要性质，即子模性（submodularity）。子模性表明：对于该问题，如果D有两个子集D'_1和D'_2，存在$D'_1\subset D'_2$关系，那么可以得到：

$$s(D'_1+d)-s(D'_1)\geq s(D'_2+d)-s(D'_2),\ d\in D-D'_2 \tag{7}$$

子模性的一个直观经济学含义是其满足边际收益递减的原则。也就是说，对于D的一个相对较大的子集D'_2而言，增加一个新的元素d所带来的D'_2对D的语义反映增益要小于D'_1对D的语义反映增益。重要的是，子模性保证了即使采用直接的贪心启发式方法，得到的代表性语义反映的子集会以$(1-1/e)$的近似程度逼近最优解（Ma et al.，2017）。

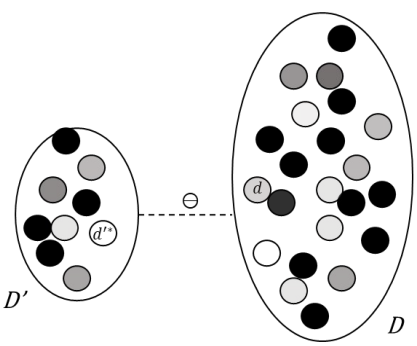

图2 代表性语义反映示意

此外，有一类数据情形值得关注，体现在数据元素之间的差异性方面。如果数据本身具有类别标签，或者数据集合中存在较多相似的数据实例可以进行归聚或类别划分，则可以采用一个不同的求解策略，即构造式策略。基本思路是从大数据集合D的每一个类别中提取一个代表元素，以此来构造生成小数据集合D'（如图3所示）。

具体来说，设H为D的一个划分，即$H=\{D_1, D_2, \cdots\cdots, D_k\}$，且$D=\bigcup_{j=1}^{k} D_j$，$D_j$为划分的一个群簇 。群簇内的数据元素是相似的（或具有超过阈值标准的强相似程度），而群簇间的数据元素是不相似的（或具有低于阈值标准的弱相似程度）。因此，可以通过选取群簇内的某一个元素来代表该群簇的其它相似元素，这个元素就被视为该群簇的代表元素。记d^*_j为D_j的代表元素，则D_j被d^*_j代表的势为：$\sum_{d^\circ \in D_j} Sim(d^*_j, d^\circ)$。进而$D=\bigcup_{j=1}^{k} D_j$，被$D'=\{d^*_1, d^*_2, \cdots\cdots, d^*_k\}$代表的势（内容覆盖“量”）为：$\sum_{j=1}^{k}\sum_{d^\circ \in D_j} Sim(d^*_j, d^\circ)$，可视作$s(D')$。即：

$$s(D')=\sum_{j=1}^{k}\sum_{d^\circ \in D_j} Sim(d^*_j, d^\circ) \tag{8}$$

而D的势（D的总量$|D|$）可视为$s(D)$。同样，若将语义转换为0-1之间的量（即占总量的比例，或内容覆盖程度），则：

$$s(D')=\frac{1}{|D|}\sum_{j=1}^{k}\sum_{d^\circ \in D_j} Sim(d^*_j, d^\circ) \tag{9}$$

而$s(D)$为1。此时，给定D和k，代表性语义反映问题可表示为：

$$max_{D'\subset D}\left[1-\left(\left|s(D')\ominus s(D)\right|\right)\right]=max_{D'=\{d^*_j \mid d^*_j\in D_j,\ j=1,\cdots,k\}\subset D}\frac{1}{|D|}\sum_{j=1}^{k}\sum_{d^\circ \in D_j} Sim(d^*_j, d^\circ) \tag{10}$$

如果原始数据本身没有类别划分，则该问题求解的一个核心难点是如何获得相应的类别划分，以保证形成的群簇内部的数据元素之间相似度较高，同时群簇之间的数据元素差异较大。惯常的类别划分方法可以通过聚类实现，如划分式聚类、凝聚式聚类、基于密度的聚类、基于网格的聚类等不同形式（Han et al.，2011）。不同的聚类方法适用于不同类型的数据，如划分式聚类和凝聚式聚类就比较常应用于文本数据（Steinbach et al.，2000）。

进一步，虽然代表性语义反映问题一般具有小数据集合的规模约束，如$|D'|=k$，但是，在实际应用中，k的设置经常是变化的，会随着决策者的需要或偏好不断调整。比如，在购物过程中，先看10条产品评论（即$k_1=10$），之后又想看10条评论（即$k_2=10$）；在审阅舆情报告时，先看8份报告（即$k_1=8$），随后又看7份（即$k_2=7$），然后再看5份（即$k_3=5$）。一般说来，每次k值的设置都意味着重新进行一次问题求解，比如先用K-means方法在大数据集合D 中进行聚类，再从各群簇中进行代表元素提取以构成小数据集合D'。一个可行的思路是采用凝聚式聚类方法，通过构建多个层次不同粒度的聚类结果来应对类别数变化的情况。然而，传统凝聚式聚类存在一个局限，即在凝聚过程中只考虑局部信息（任何两个数据或者两个簇之间的相似性），忽略了全局信息，因此容易在边缘点处产生错误划分进而影响聚类质量。鉴于此，相关研究提出了一种新颖有效的代表元素生成方法REPSET，其中包括新型层次聚类模块，设计了一个回溯机制来动态修正类别凝聚过程中的误分，同时通过一次聚类迭代生成多个层次不同粒度的聚类结果（Guo et al.，2017）。该聚类模块的基本思路流程如图4所示。

在此基础上，群簇中元素的平均相似度最高的元素为该群簇的代表元素，即：

$$d^*_j=max_{d\in D_j}\left\{\frac{1}{|D_j|}\sum_{d^\circ \in D_j} Sim(d, d^\circ)\right\} \tag{11}$$

每个群簇的代表元素合并生成小数据集合$D'=\{d^*_1, d^*_2, \cdots\cdots, d^*_k\}$。REPSET方法与其他相关方法比较存在显著优势。同时，REPSET方法也在大型企业的员工博客平台上进行了检验应用，获得了良好的管理决策支持效果（Guo et al.，2017）。

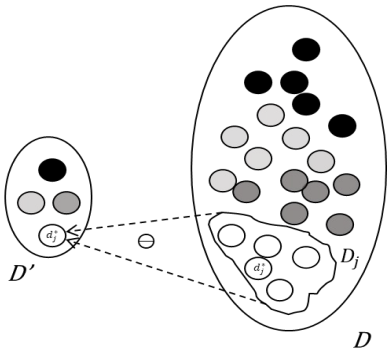

图3 代表性语义反映示意
（类别划分情形）

概括说来，代表性语义反映作为“大数据—小数据”问题的一类情形，体现了大数据环境下管理决策的新挑战，其应对策略对于在数据实例和内容覆盖层面上的“以小见大”洞察具有重要决策支持意义。相关求解方法在丰富场景中不断创新和应用，可以帮助有效解决在基于数据决策中的大小数据间内容语义的信息不对称性，进而更好地为各领域的价值创造进行大数据驱动的决策赋能。

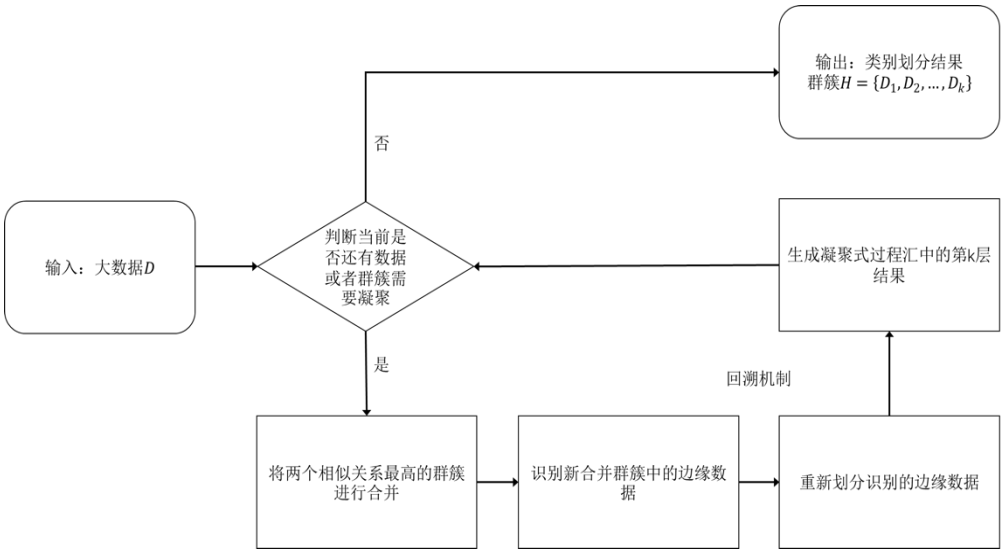

图4 新型层次聚类模块的思路流程

四、一致性语义反映

在决策场景中，人们常常需要在了解大数据全貌的同时获得有温度的具象感知，即希望通过大数据集合的概括性语义（如属性特征的统计汇总）和小数据集合的实例语义来形成对于数据内容及其含义的认识。一致性语义反映是“大数据—小数据”问题的一种类型，旨在从隐式语义的视角获得一个数据实例的小集合，以求小集合在特定属性特征下反映的概括性语义与大数据集合的概括性语义尽可能地一致。例如，在线上购物环境中，消费者可以方便地看到平台提供的每一产品所有评论在相关属性特征上的极性分布情况（如正面负面评论分别在价格、质量、颜色、服务等特征上的占比），同时也可以浏览每一产品的具体评论内容。前者是汇总信息，概括性强且相对抽象；后者是实例型的具体数据内容，临场感强且相对具象。由于消费者通常只是阅看所有评论中的一小部分评论（如显示在第一页的评论），所以，阅看到的评论在特定属性特征上的极性分布（如10条评论中“服务”的正面评论有3条，负面评论有7条）就可能与汇总信息显示的极性分布情况不同（如“服务”的正负评论比例是80%∶20%）。无疑，这会使消费者产生困惑，从而产生有偏的决策。类似的场景很常见，如企业口碑的详略画像、受众声音的宏微聆听、媒体报道的点面呈现、政策分析的繁简要义等等。解决这种在大数据集合语义与小数据集合语义之间的信息不对称性具有重要意义。此时，语义反映强调小数据集合在相关属性特征（如“服务”）上的取值模式是否与大数据集合的一致性，这种“反映”称为一致性语义反映。

一致性语义反映的概念内涵是小数据集合D'在相关属性上的取值模式与大数据集合D在这些属性上的取值模式相一致。取值模式一般是对于数据实例属性取值的深层次刻画，所以一致性语义反映属于隐式语义反映的范畴。一致性语义反映如图5所示，大数据集合D在属性特征上的取值模式为其在各属性特征(A_1,A_2,A_3,A_4)上的取值分布，即语义$s(D)$。一致性是指对于子集D'，其元素在部分属性特征（如A_1,A_3）上的取值模式$s(D')$与D在这些属性上的取值模式$s(D)$相一致。通常，子集D'的非空取值属性特征个数远小于全集D的属性特征个数。以产品评论为例，所有产品

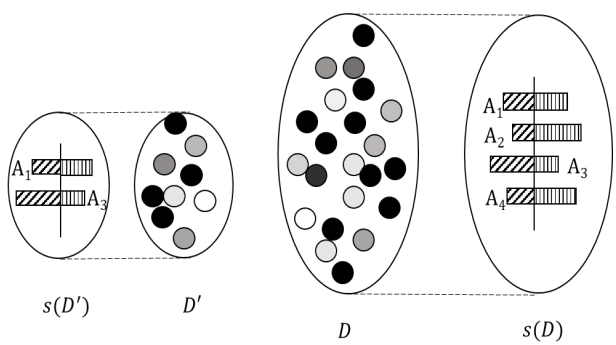

图5 一致性语义反映示意

评论涉及到的属性特征可能很多，涵盖产品从内在到外延属性特征的方方面面。然而，就一个评论而言，可能只涉及到客户购买/使用体验的个别方面，如价格和物流，而没有提及其它方面。对于消费者浏览的小集合评论而言，通常提及的属性特征个数也会很有限。

由于取值模式作为隐式语义存在多种形态，这里围绕取值分布形态讨论一致性语义反映问题的形式化表示。令 D 表示大数据集合，D' 表示小数据集合，$\mathbb{A}_{D'}$ 是 D' 上所有特征 A 的集合（即 $A\in\mathbb{A}_{D'}$），D_A 表示 D 中包含属性 A 的数据所形成的集合。不失一般性，假设每个属性特征的值域为二值集合 $\{x,y\}$（如评论极性取值）。对于属性特征 A，取值为 x 的数据在 D_A 上的占比（取值分布）为 $|D'^{x}_A|/|D'_A|$，类似地，其在 D 上的占比为 $|D^x_A|/|D_A|$。这里，就分别对应着 D'_A 和 D_A 的语义：$s(D'_A)=|D'^{x}_A|/|D'_A|$，$s(D_A)=|D^x_A|/|D_A|$。进一步，考虑所有属性特征，可以得到 D' 和 D 的语义：$s(D')=\sum_{A\in\mathbb{A}_{D'}}s(D'_A)$，$s(D)=\sum_{A\in\mathbb{A}_{D'}}s(D_A)$。则给定 D 和 k，一致性语义反映问题可表示为：

$$\max_{D'\subset D,\,|D'|=k}\left[1-\left(\left|s(D')\ominus s(D)\right|\right)\right]=\max_{D'\subset D,\,|D'|=k}\sum_{A\in\mathbb{A}_{D'}}\left(1-\left|\frac{|D'^{x}_A|}{|D'_A|}-\frac{|D^x_A|}{|D_A|}\right|\right) \tag{12}$$

如果希望体现不同属性特征的重要性，可以引入属性特征权重 $w_A=|D_A|/|D|$。此时，求解问题可以为：

$$\max_{D'\subset D,\,|D'|=k}\sum_{A\in\mathbb{A}_{D'}}w_A\times\left(1-\left|\frac{|D'^{x}_A|}{|D'_A|}-\frac{|D^x_A|}{|D_A|}\right|\right) \tag{13}$$

值得一提的是，因为这里表述的为二值属性的占比，所以大小数据集合在一致性上的语义差异最终体现在属性的一个取值上面。当然，如果属性的取值更多，则可以相应进一步扩展。

对于上述一致性语义反映问题的求解存在难点。首先，该问题具有高的复杂度，因此对于求解带来挑战。相关研究证明，该问题是一个 NP-难问题（Zhang et al.，2016）。这就使得问题求解难以采用传统优化方法，而需要进行方法创新以研发有效的启发式策略。进而，带来了的挑战就是如何提出有效的启发式方法。相关研究提出了一种新颖的求解方法 eSOP，通过设计增强型逐步寻优策略（思路流程见图 6），使得在求解精度和效率上具有优势（Zhang et al.，2016）。

具体而言，在求解方法逐次迭代进行小数据集合扩展时，不仅保留具有最高一致性得分的数据，而是通过引入一个参数 $\alpha\in[0,1]$ 来控制一致性得分的阈值，以生成一个一致性得分在 $[\min Value+\alpha\times(\max Value-\min Value),\max Value]$ 区间的候选集合供后续一致性子集使用。当 $\alpha=1$ 时，相当于仅保留具有最大一致性得分的数据进入后续迭代中，即求解方法退化为贪心方法；而当 $\alpha=0$ 时，相当于保留一致性得分在 $[\min Value,\max Value]$ 之间的数据作为候选集合，相当于方法退化为穷举求解方法。因而，通过控制 α 的取值，可以调节方法的寻优空间及求解效率。

围绕产品评论情景，通过大量的数据实验表明，eSOP 方法与其他相关方法比较存在显著优势。同时，eSOP 方法也在大型线上购物平台上进行检验应用，获得了良好的决策支持效果（Zhang et al.，2016）。此外，近年相关研究针对问题的不同视角和情境进行了扩展。扩展方面包括：Zhang 等（2016）从消费者行为学角度，刻画消费者可能在不同位置的评论处停止下来，引入消费者在第 i 条评论处阅读停止概率 p_i，进而将求解问题转化为优化期望意义下的一致性，即 $\max\sum p_i\sum_{A\in\mathbb{A}_{D'}}w_A\times(1-\left|\frac{|D'^{x}_A|}{|D'_A|}-\frac{|D^x_A|}{|D_A|}\right|)$。Wang 等（2018）针对消费者对在线评论发布时间的关注，引入时间衰减函数对在线评论进行赋权，进而形成考虑评论时效性的一致性语义反映问题，并在寻优方法中引入优

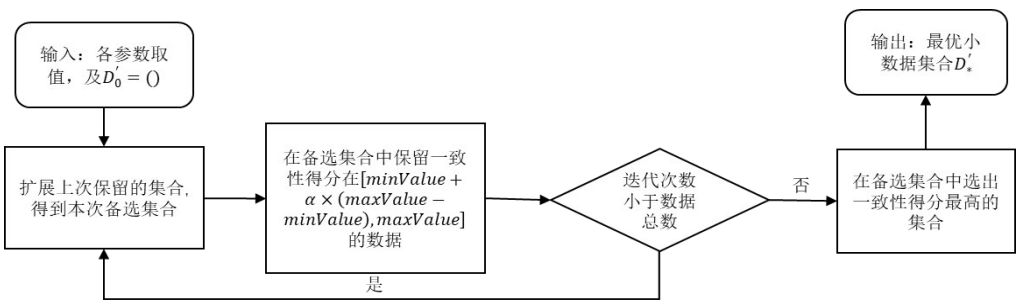

图 6 增强型逐步寻优策略思路流程

先队列结构和剪枝策略，在保证优化效果同时进一步优化求解效率。Chen等（2018）进一步考虑了在线评论质量因素，设计了结合评论质量评估框架的语义衡量方法，进而形成考虑评论质量的一致性语义反映问题，在寻优方法中结合了深度优先搜索方法和贪心策略，从而兼顾求解效率及精度。

概括说来，一致性语义反映作为“大数据—小数据”问题的一类情形，体现了大数据环境下管理决策的新挑战，即如何兼顾和协同大数据与小数据传递语义信息的一致性，使大小数据传递同频声音，进而帮助管理者做出精准决策。建模和求解方法的不断创新，可以为大数据驱动的决策赋能，对于在取值模式（如分布一致性）层面上的“以小见大”洞察具有重要决策支持意义。

五、多样性语义反映

在决策场景中，人们常常需要从多样性的角度来观察世界，即希望通过小数据集合来体现大数据集合的多样性特点。多样性语义反映是“大数据—小数据”问题的一种类型，旨在从隐式语义的视角获得一个数据实例的小集合，以求小集合在特定属性下反映的结构性语义与大数据集合尽可能地相近。例如，浏览新闻时，人们期待多角度的报道；了解舆情时，人们期待不同的诉求；竞品搜索时，人们期待提供丰富的选择；政策制定时，人们期待面向各类人群，等等。此时，语义反映强调小数据集合反映大数据集合的类别多样性。

多样性语义反映的概念内涵是指小数据集合D'对于大数据集合D的相应类别结构$H=\{D_1,D_2\cdots\cdots D_k\}$的覆盖程度。相对于$D'$，$D$的元素在特定属性（如形状）上的取值存在某种模式（如类别：圆形、矩形、多边形、梯形），即构成了D上的一个语义$s(D)$。取值模式一般是语义的深层次刻画，所以多样性语义反映属于隐式语义反映的范畴。多样性语义反映如图7所示，显示出大小数据集合语义在类别上的对应。

给定大数据集合D，多样性语义反映问题是寻找规模为k的子集D'，使得D'在H的框架下（即D的类别结构下）对于D的覆盖最大。换句话说，多样性语义反映具有结构覆盖的意味。计算结构覆盖程度的一个思路是通过信息熵（Shannon，1948）来计算H的“信息载量”。具体而言，对于D的子集D'（$|D'|=k$），假设D'中的每一个元素d'_j对应一个D的子类D_j，$j=1,2,\cdots\cdots,k$（或将d'_j看作D_j的类别标签）。进而，对于D中的任一元素d，按照与D'中每一个元素d'_j的相似程度高低，确定其类别归属。即元素d属于的D_j程度（隶属度）为d与d'_j的相似度$sim(d,d'_j)$，当d与d'_j相似度最大时，d的类别划分为D_j。基于此，可以得到集合D的类别划分$D_1,D_2,\cdots\cdots,D_k$。相应地，D_j的势$n^v_j=\sum_{d\in D_j}sim(d'_j,d)$表示以$d'_j$为类标签的类别$D_j$中元素的隶属度的和，构成了$d'_j$与$D_j$的对应。$\sum^k_{j=1}n^v_j=n^v$则是$D'$关于$D$的对应。通过$D'$中元素与$D$中不同类别的对应，可以进一步计算$D$的“信息熵”以体现$D'$在类别划分$D_1,D_2,\cdots\cdots,D_k$的框架下对于$D$的多样性反映。这样，$D'$对于$D$的多样性语义为（$k>1$）：

$$s(D')=-\frac{1}{log_2 k}\sum_{j=1}^{k}\frac{n^v_j}{n^v}log_2\left(\frac{n^v_j}{n^v}\right) \tag{14}$$

当$k=1$时，令$s(D')=1$。类似，有$s(D)=1$。则多样性语义反映问题可以表示为：

$$\max_{D'\subset D,|D'|=k}\left[1-\left(\left|s(D')\ominus s(D)\right|\right)\right]=-\frac{1}{log_2 k}\sum_{j=1}^{k}\frac{n^v_j}{n^v}log_2\left(\frac{n^v_j}{n^v}\right) \tag{15}$$

多样性语义反映问题同样具有NP-难的性质（Ma et al.，2017），通常也需要采用启发式方法进行求解。然而单一的启发式方法容易陷入局部最优，导致影响问题的求解精度，因而可采用多种启发式方法进行组合寻优的方式来提升效果。相关研究提出一个融合方法，通过贪心算法和模拟退火算法的组合策略，以获得良好的寻优效果和算法效率（Ma et al.，2017）。图8给出了模拟退火随机搜索

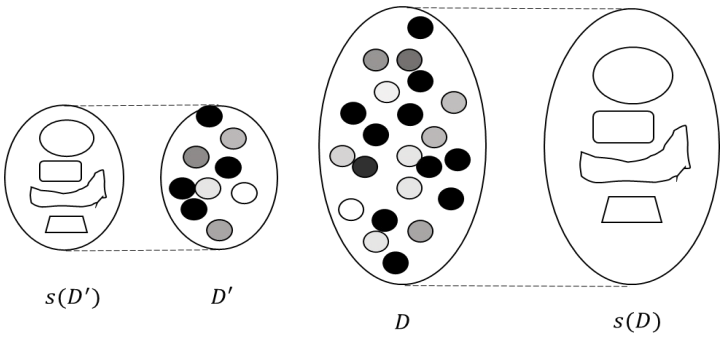

图7 多样性语义反映示意

策略的思路流程。

图8思路是一个迭代的过程，其中的两个主要步骤是：（1）计算新解与当前解的多样性语义反映的差值。这一步需要计算当前结果集合和可能的新结果集合的多样性语义反映，并计算新解带来的增益（负值表示减少值）。这里计算的增益，一方面决定是否使用新解代替当前解，另一方面对后续“接受概率”的产生起作用。（2）判断新解是否被接受。这一步是引入模拟退火的“随机搜索”思路的核心内容。如果出现新解没能改进多样性语义反映的效果，并不一定放弃新解，而是以一定的“接受概率”接受无改进的新解。这种“接受概率”会受到新解的多样性语义反映增益以及当前冷却温度的双重影响。如果当前“接受概率”大于随机概率，则接受新解代替当前解，否则仍保留当前解。当新解与当前解测度值相同时，仍然坚持使用新解代替当前解，这也是为了鼓励算法进行更多寻优，避免陷入某个局部最优值附近。围绕信息搜索服务场景，通过大量的数据实验表明，与其他相关方法比较，图8所示的融合方法能够获得更好的多样性反映效果。此外，所提出的融合方法也在大型搜索平台上进行了检验应用，显著提升了平台的信息服务效果（Ma et al.，2017）。

概括说来，多样性语义反映作为“大数据—小数据”问题的一类情形，体现了大数据环境下管理决策的新挑战，即如何缩减大数据与小数据在特定属性下的结构性语义差异，让小数据以棱镜的方式折射出大数据中的不同颜色，进而帮助决策者掌握大数据中的结构语义，做出全面性判断。建模和求解方法的不断创新，可以为大数据驱动的决策赋能，对于在取值模式（如结构覆盖）层面上的“以小见大”洞察具有重要决策支持意义。

六、其他相关工作

代表性、一致性、多样性是“大数据—小数据”问题的3类典型情形，体现了语义反映的不同视角和对小数据集合的不同获取策略。这里，有两个相关的问题值得一提。一个相关问题是Top-k问题（Fagin，1999；Fagin et al.，2003；Zhang et al.，2020）。Top-k问题亦是旨在从大数据集合中提取一个规模为k的子集。但是，它与“大数据—小数据”问题存在着显著不同。Top-k问题寻求的子集具有高偏序特征（根据排序策略涵义），而“大数据—小数据”问题寻求的子集具有高语义相近性（根据语义反映涵义）。例如，对于数列整数数据集D={1，2，……，80，81}，如果k=9，那么对于两个子集D_1={81，80，79，78，77，76，75，74，73}和D_2={5，14，23，32，41，50，59，68，77}而言，Top-k将获得D_1。显然，以中值、均值、标准差来衡量，和D_2相比，D_1对D的语义反映较差，而采用文中的代表性方法思路可以获得与D_2类似的结果。此外，另一个相关问题是随机抽样，其属于代表性内容覆盖的范畴。给定k，可以在D上抽取一个规模为k的随机样本来构成D'。然而，由于随机偏差（Heckman，2010），一次随机抽样获得的D'通常难以保证在内容覆盖意义上的寻优，即D中d被D'中d'代表的程度一般低于d被d'^{*}代表的程度（$Sim(d'^{*},d)>Sim(d',d)$）（参见图2）。而且，k越小（相对于$|D|=n$，$k\ll n$），则随机偏差导致的内容覆盖语义偏离越严重。

此外，在复杂的决策场景中，视角融合以及特定目标诉求等需要对于问题建模和求解路径进行新的探索，进而引出了组合式语义反映、小数据质量优化等相关问题。

如果关注代表性与多样性的视角融合，可以考虑在前面提到过的内容覆盖和结构覆盖含义下的组合式语义反映。相关研究通过小数

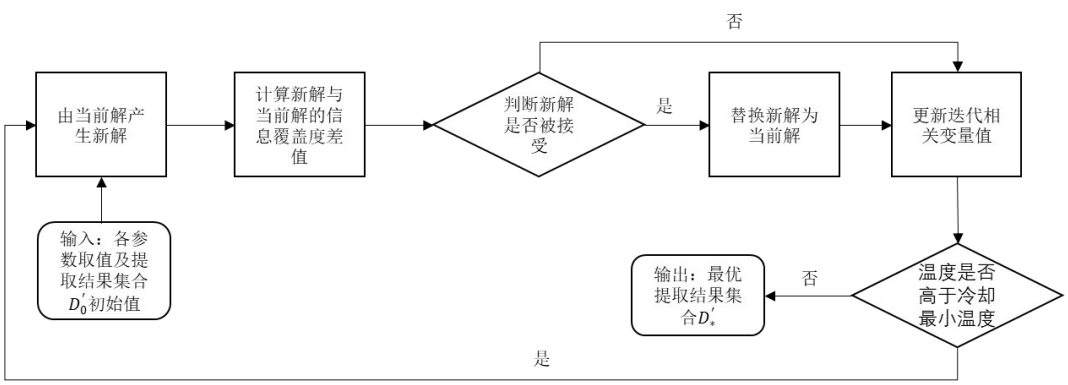

图8 模拟退火随机搜索策略思路流程

据集合同时在内容实例层面和类别结构层面反映大数据集合相关语义的形式，将B-S问题（$\max_{D'\subset D,|D'|=k}[1-(|s(D')\ominus s(D)|)]$）表示为：

$$max_{D'\subset D,|D'|=k}\left(\frac{1}{|D|}\sum_{d\in D}Sim(D',d)\right)\times\left(-\frac{1}{\log_2 k}\sum_{j=1}^{k}\frac{n_j^v}{n^v}\log_2\left(\frac{n_j^v}{n^v}\right)\right) \tag{16}$$

即为内容覆盖度（$Cov_C(D',D)$）与结构覆盖度（$Cov_s(D',D)$）的乘积（Ma et al.，2017）。

从前面讨论可以得到，这一组合式语义反映问题同样具有NP-难的复杂度。在求解过程中除考虑多种启发式方法的组合外，还可以结合问题的特点设计特定的剪枝策略，以对求解空间进行合理的收缩，在不牺牲求解精度的情况下尽量提升求解效率。相关研究提出了有效的启发式方法 COV_{C+S}-*Select* 及其近似策略 *Fast-*COV_{C+S}-*Select*，比如候选集合生成策略、精简寻优空间迭代策略等（Ma et al.，2017）。

再者，“大数据—小数据”问题可以根据特定目标的需求并结合其它语义表述和测度进一步优化小数据提取结果。例如，一类语义约束测度可以作用于小数据集合以提升集合本身的质量，如紧凑度、冗余度等（Zhang et al.，2012；Ma et al.，2012；Ma et al.，2017）。这样，可以在求解B-S问题前/后进一步凝炼小数据，增强以小见大的洞察质量。例如，在小数据集合 D' 中，任意元素 d' 对集合 D' 的冗余度定义如下：

$$Red(d',D')=1-\frac{1}{\sum_{d\in D'}sim(d',d)} \tag{17}$$

其中，$\sum_{d\in D}sim(d',d)$ 表示 d' 与 D' 中元素的总相似度之和。进而集合 D' 的冗余度可表示如下：

$$Red(D')=\frac{\sum_{d'\in D'}Red(d',D')}{|D'|}=\frac{1}{|D'|}\times\sum_{d'\in D'}\left(1-\frac{1}{\sum_{d\in D}sim(d',d)}\right) \tag{18}$$

该测度可以通过前剪枝或后剪枝的形式作用于B-S问题的求解，将冗余度较高的元素滤除，使得获取的小数据集合具有简明高效的特点，更好地支持基于数据的管理决策。

七、总结

现代科学技术正深刻改变着人类的思维、生产、生活和学习方式，也正在以数字的方式重构着个人、组织、社会与政府的管理决策，催生出大数据驱动的新型决策范式。在此背景下，一方面，大数据为具有全局视图的管理决策提供了可能；另一方面，数据可获性、成本、时间以及人们对于大量数据的接受和消化能力使得在许多应用场景中，人们只能基于有限的小数据进行决策。这种决策信息的不对称性，使得提出“大数据—小数据”问题以寻求小数据更好地反映大数据语义对于科学决策来讲变得尤为关键，具有重要的理论和实践意义。鉴于此，本文提出了“大数据—小数据”问题及其概念内涵，并从代表性、一致性、多样性的视角出发，讨论了小数据“反映”大数据语义的形式和问题求解路径，呈现了大数据—小数据问题的多种应用场景和方法创新。

在数字经济中，随着数据要素和数智化作用的日益显现，也将出现更多的“大数据—小数据”问题的应用场景和有效实践。进一步的研究可在本文的基础上，继续探索在新场景下“大数据—小数据”问题建模、求解及其赋能的不同形式，洞察和解构大数据中的深层次语义，提升大数据驱动的管理决策和价值创造水平。

（作者单位：陈国青、卫强、郭迅华，清华大学经济管理学院；张瑾，中国人民大学商学院；王聪，北京大学光华管理学院）

参考文献

（1）陈国青、吴刚、顾远东、陆本江、卫强：《管理决策情境下大数据驱动的研究和应用挑战——范式转变与研究方向》，《管理科学学报》，2018年第7期。

（2）陈国青、曾大军、卫强、张明月、郭迅华：《大数据环境下的决策范式转变与使能创新》，《管理世界》，2020年第2期。

（3）冯芷艳、郭迅华、曾大军、陈煜波、陈国青：《大数据背景下的商务管理若干前沿课题》，《管理科学学报》，2013年第1期。

（4）徐宗本、冯芷艳、郭迅华、曾大军、陈国青：《大数据驱动的管理与决策前沿课题》，《管理世界》，2014年第11期。

(5)Chen, G. Q., Wang, C., Zhang, M. Y., Wei, Q. and Ma, B. J., 2018, "How 'Small' Reflects 'Large'?: Representative Information Measurement and Extraction", *Information Sciences*, vol. 460, pp. 519~540.

(6)Fagin, R., 1999, "Combining Fuzzy Information from Multiple Systems", *Journa of Computer and System Sciences*, vol. 58, pp. 83~99.

(7)Fagin, R., Lotem, A. and Naor, M., 2003, "Optimal Aggregation Algorithms for Middleware", *Journal of Computer and System Sciences*, vol. 66, pp. 614~656.

(8)Guo, X. H., Wei, Q., Chen, G., Zhang, J. and Qiao, D. D., 2017, "Extracting representative Information on Intra-organizational Blogging Platforms", *MIS Quarterly*, 41(4), pp.1105~1127.

(9)Han, J. W., Kamber, M. and Pei, J., 2011, *Data Mining Concepts and Techniques Third Edition*, Elsevier.

(10)Heckman, J. J., 2010, *Selection Bias and Self-selection*, Palgrave Macmillan, London.

(11)Hilbert, M. and Lopez, P., 2011, "The World's Technological Capacity To Store, Communicate, and Compute Information", *Science*, 332(6025), pp. 60~65.

(12)Ma, B. and Wei, Q., 2012, "Measuring the Coverage and Redundancy of Information Search Services on E-commerce Platforms", *Electronic Commerce Research and Applications*, 11(6), pp.560~569.

(13)Ma, B., Wei, Q., Chen, G., Zhang, J. and Guo, X., 2017, "Content and Structure Coverage: Extracting a Diverse Information Subset", *INFORMS Journal on Computing*, 29(4), pp. 660~675.

(14)Reinsel, D., Gantz, J. and Pydning, J., 2019, *Data Age 2025*, IDC.

(15)Shannon, C., 1948, "A Mathematical Theory of Communication", *Bell System Technical Journal*, 27(3), pp. 379~423.

(16)Steinbach, M., Karypis, G. and Kumar, V. A., 2000, "Comparison of Document Clustering Techniques", in Grobelnik, M., D. Mladenic and N. Milic-Frayling, eds: Proceedings of the KDD Workshop on Text Mining, ACM, New York, NY, United States.

(17)Wang, C., Chen, G. and Wei, Q., 2018, "A Temporal Consistency Method for Online Review Ranking", *Knowledge-Based Systems*, vol. 143, pp. 259~270.

(18)Zhang, Z., Chen, G., Zhang, J., Guo, X. and Wei, Q., 2016, "Providing Consistent Opinions from Online Reviews: A Heuristic Stepwise Optimization Approach", *INFORMS Journal on Computing*, 28(2), pp. 236~250.

(19)Zhang, J., Chen, G. and Tang, X., 2012, "Extracting Representative Information to Enhance Flexible Data Queries", *IEEE Transactions on Neural Networks and Learning Systems*, 23(6), pp. 928~941.

(20)Zhang, W. X., Deng, Y. and Lam, W., 2020, "Answer Ranking for Product-related Questions Via Multiple Semantic Relations Modeling", in Huang, J., Y. Chang and X. Q. Cheng, eds: SIGIR'20: Proceedings of the 43rd International ACM SIGIR Conference on Research & Development in Information Retrieval, ACM, New York, NY, United States.

(上接第202页)

参考文献

(1)陈锡康:《投入产出方法与国民经济综合平衡》,《数学的实践与认识》,1983年第2期。

(2)邓小平:《邓小平文选:第3卷》,人民出版社,1993年。

(3)黄海军、姚忠、张人千、刘作仪、赵秋红、吴俊杰:《管理科学与工程学科——"十二五"发展战略与优先资助领域研究》,科学出版社,2012年。

(4)黄海军、刘作仪、姚忠、赵秋红、吴俊杰、单伟、部慧、田琼:《管理科学与工程学科——"十三五"发展战略与优先资助领域研究报告》,科学出版社,2016年。

(5)金佳绪:《习近平为我国"新发展阶段"定向》,《理论导报》,2020年第8期。

(6)蔺亚琼:《管理学门类的诞生:知识划界与学科体系》,《北京大学教育评论》,2011年第2期。

(7)骆茹敏:《奋进的中国管理科学》,科学出版社,2010年。

(8)钱学森、许国志、王寿云:《组织管理的技术——系统工程》,《上海理工大学学报报》,2011年第6期。

(9)盛昭瀚、刘慧敏、燕雪、金帅、邱聿旻、董梁:《重大工程决策"中国之治"的现代化道路——我国重大工程决策治理70年》,《管理世界》,2020年第10期。

(10)盛昭瀚:《问题导向:管理理论发展的推动力》,《管理科学学报》,2019年第5期。

(11)王双梅:《邓小平与20世纪60年代的国民经济调整》,《党的文献》,2011年第5期。

(12)习近平:《在经济社会领域专家座谈会上的讲话》,《人民日报》,2020年8月25日。

(13)徐伟宣、张玲玲、李建平、林则夫、郑秀榆:《管理科学与工程学科发展现状与前景展望》,中国优选法统筹法与经济数学研究会,《2007~2008年管理科学与工程学科发展报告》,中国优选法统筹法与经济数学研究会,2008年。

(14)徐伟宣:《华罗庚与优选法统筹法》,《高等数学研究》,2006年第6期。

(15)中共中央文献研究室:《邓小平思想年谱》,中央文献出版社,1998年。

(16)中国运筹学会:《中国运筹学发展研究报告》,《运筹学学报》,2012年第3期。

The "Big Data- Small Data" Problem: Insights for the Big through the Small

Chen Guoqing[a], *Zhang Jin*[b], *Wang Cong*[c], *Wei Qiang*[a] *and Guo Xunhua*[a]

(a. School of Economics and Management, Tsinghua University; b. School of Business, Renmin University of China; c. Guanghua School of Management, Peking University)

Summary: Recent years have witnessed a paradigm shift in managerial decision-making, especially in the context of digital economy, where big data becomes a strategic asset and innovation enabler. Big data extends the possibility for people to see the whole picture of the reality in a panoramic and fine-grained manner, so as for them to have a comprehensive understanding of the things at hand in their decision processes. However, in many cases, due to various reasons for data availability, cost, time, capability and psychological factors, people often face the data that is limited and partial (i.e., small data). In other words, people wish to have a global view of what big data is, but may often have to rely upon small data that they can reach or handle. Then a concern arises for the discrepancy between big data and small data, giving rise to an important issue of information asymmetry, which will affect decision effectiveness. Apparently, addressing the issue is of high significance for both academia and practitioners in individual, organizational and governmental levels of decision-making.

In that regard, this article introduces the notion of the Big data-Small data problem (in short, the B-S problem) in light of semantic reflection, and articulates the problem as a pursuit of insights for the Big data through the Small data that decision makers face. Concretely, given the set of big data, the problem is to find a subset as small data with a pre-specified size, such that the semantics of small data reflect, as closely as possible, the semantics of big data. Formally, the B-S problem is formulated as follows: $\max_{D' \subset D}(1-(s(D) \ominus s(D'))$, where D is the set of big data, D' is a subset of D as small data with a pre-specified size, $s(D)$ is semantics of D, $s(D')$ is semantics of D', and $\ominus$ is a super-subtraction operator mapping the distance between semantic spaces to $[0,1]$. Here, the semantics of a dataset can be explicit or implicit, depending on data values of directly observed instances or indirect patterns.

Subsequently, the article discusses the B-S problem in three types mainly from the perspectives of representativeness, consistency and diversity in terms of how small data can representatively reflect, consistently reflect and diversely reflect the semantics of big data, respectively.

For semantic reflection on representativeness, it is the B-S problem that is of explicit semantics nature. The reflection is generally at the level of raw facts, represented with similarity measures between data instances in forms of content coverage. For semantic reflection on consistency, it is the B-S problem that is of implicit semantics nature. The reflection is generally at the level of indirect patterns, represented with distributions of data values with respect to related attributes. For semantic reflection on diversity, it is the B-S problem that is of explicit semantics nature. The reflection is generally at the level of indirect patterns, represented with entropy measures for groupings in forms of structure coverage. The respective solutions to different types of the B-S Problem in various application scenarios show their superiority over other baselines, as well as their methodological merits and practical effectiveness in managerial decision support.

Notably, introducing the notion of the B-S problem is deemed important and meaningful in threefold. First, the information asymmetry may mislead decision makers if the insights of big data cannot be well captured in small data. Second, the B-S problem is generally of high complexity, which needs innovative efforts to develop solutions in effective and efficient ways. Third, the B-S problem may have other forms/types in rich and complex contexts, which further broadens the realm of explorations for big data - driven decision sciences.

Keywords: The "Big Data-Small Data" problem; semantic reflection; managerial decision-making

JEL Classification: C8