大数据的价值发现方法

杜小勇[1,2]，陈跃国[1,2]
1. 数据工程与知识工程教育部重点实验室（中国人民大学），北京 100872；
2. 中国人民大学信息学院，北京 100872

摘要
大数据的价值有不同的体现形式和发现价值的途径。总结了从大数据中发现价值的3种基本途径：数据服务、数据分析与数据探索，并对它们的特点进行了分析和对比。数据服务通过提供高性能和高并发的数据访问从微观层面体现数据价值；数据分析侧重利用统计模型的方法在宏观层面上对大数据进行处理，通过产生数据洞察的形式体现数据价值；数据探索侧重通过交互模型在微观和宏观的不断变换，引导用户浏览和发现数据的价值。

关键词
大数据；价值发现；数据服务；数据分析；数据探索

中图分类号：TP391.1 文献标识码：A doi: 10.11959/j.issn.2096-6271.2017015

Approaches for value extraction on big data

DU Xiaoyong[1,2], CHEN Yueguo[1,2]
1. MOE Key Laboratory of Data Engineering and Knowledge Engineering (Renmin University of China), Beijing 100872, China
2. School of Information, Renmin University of China, Beijing 100872, China

Abstract
The value of big data can be presented in different means, and therefore it has different ways to extract the value out of big data. Three approaches of value extraction on big data: data service, data analytics, and data exploration were summarized. The characteristics of these approaches were analyzed and compared. In summary, data service reflects the value of data from the micro-level by supporting high-performance and high-throughput read and write operations. Data analysis focuses on the usage of statistical models to generalize data distribution at macro-level, and it extracts values by generating insights from data. Data exploration focuses on interactive models in the constant interchange of micro-level and macro-level to guide the users browse and discover values out of the data.

Key words
big data, value extraction, data service, data analytics, data exploration

1 引言

关于大数据的讨论，一方面人们需要厘清大数据的概念，开发适用的大数据系统和工具，探索大数据的应用模式等，另一方面人们更关心如何将大数据的价值变现。这对于一个企业来说尤其重要，否则，收集和存储了大量的数据，消耗了大量的钱财，如果大数据不能被很好地利用，从经济上讲就是不合算的，这样的事情也不会长远。本文对大数据价值发现的主要方法进行综述，并结合一些例子展示其特点。

数据的价值在信息系统作为计算机科学与技术的一个非常重要方向之初，就已经得到了人们足够的认可和重视。那么，在大数据时代，数据的价值有什么独特之处吗？很明显，大数据特有的价值源于其规模效应，当数据量足够大时，其价值能够随之得到足够的放大。这一点在当前十分火热的深度学习领域得到了很好的验证。最近几年，在机器学习和人工智能领域，研究者们利用大数据能够更好地训练分类器（深度神经网络），在很多应用领域（如语音和图像识别等）取得了巨大的进步。包括一直以来被认为人工智能难以攻破的围棋领域，新的人工智能技术已经在很多方面超过了人类。究其原因，就是大规模数据，尤其是具有标注的大数据，使得人们构建更精细化的分类器，让一度不被重视的神经网络分类器，以至整个人工智能领域，重新焕发了青春。这是大数据在深度分析领域体现出价值的一个例子。

大数据价值的发现与其所处的应用场景密切相关。概括起来，笔者认为大数据价值发现可以划分为三大类：数据服务、数据分析和数据探索。数据服务是面向大规模用户，提供高性能的数据查询、检索、预测等服务，通过直接满足用户需求而将数据价值变现的形式；数据分析是分析人员利用经验，通过对大规模数据使用特定的计算模型进行较为复杂的运算，从而发现易于人们理解的数据模式或规律所进行的数据价值变现的一种运算形式；数据探索是一种利用数据分析和人机交互的结合，通过不断揭示数据的规律和数据间的关联，引导分析人员发现并认识其所未知的数据模式或规律，其价值更多地体现在对未知途径的数据模式和规律的探索。下面分别就这3类数据价值发现方法加以细致的介绍。

2 数据服务

数据服务针对用户非常明确的数据查询和处理任务，以高性能和高吞吐量的方式实现大众化的服务，是数据价值最重要也是最直接的发现方式。由于要处理大众化的服务请求，每个服务任务必须能够被快速地处理掉，因此，数据服务的单个任务/负载不能过于复杂，单任务直接处理的数据不能太大，任务对应的用户需求和采用的数据处理方法必须是明确的。一些典型的数据服务包括事务处理、数据查询、信息检索、数据预测。

事务处理是传统数据库范畴的价值发现形式，它针对的主要是任务关键型（mission-critical）的数据服务，如银行记账、商业交易等。几十年来，随着数据库技术日趋成熟，此类数据价值得到了较为充分的挖掘。然而，随着移动互联网时代的到来，很多互联网特有的现象级应用给事务处理带来了不小的挑战。两个较为典型的例子就是“双十一”和12306春节购票。高并发的事务处理请求给现有数据库系统带来了巨大的挑战，研究人员尝

试使用新硬件提升数据库系统性能，甚至在软件层面重新构建数据库系统，形成以NewSQL数据管理技术为主的新型事务处理系统[1]。

数据查询主要是面向快速查找或修改数据的服务需求，它比事务处理更简单，对数据一致性要求没那么强，但对服务的吞吐量要求非常高。典型的例子就是采用键值对模型的NoSQL数据库[2]。查询服务通常以行键、列名、版本号等确定数据的逻辑单元，并通过行键、列名和版本等信息进行基于键值的数据查询。对于NoSQL数据库，由于弱化了数据事务一致性准则（采用最终一致性），数据索引相对简单，事务类型单一，适用于并行化处理，其在一定规模的集群下能够达到较高的数据读写吞吐率（每秒百万级）。例如，用户画像的使用。互联网行业通过为用户提供服务，收集了很多用户行为数据，利用数据分析手段构建了用户画像，这些画像为了精细地刻画用户的特征，经常使用成千上万个属性。数据服务就是要针对具体用户，快速读取或者写入一些属性。比如，很多门户网站有个性化的用户登录界面。当用户登录后，立即对NoSQL数据库发起一个服务请求，获得用户某些方面的属性，利用这些属性和一些特有的业务规则，网站引擎对用户进行个性化的界面显示。要求整个服务过程能够在几毫秒到几十毫秒的时间处理完，是一种高标准性能要求的数据服务请求。

信息检索是指从大规模的数据集中快速查找满足用户需求的资料或数据片段的过程。此类服务中用户通过关键词（或自然语言语句）表达信息需求，因此用户需求具有一定的模糊性。为了快速得到查询结果，信息检索系统必须预先构建数据索引（如倒排表）。完成检索后，检索结果要根据查询的相关度进行排序。搜索引擎的成功正是得益于其对大规模互联网用户提供在海量Web数据基础上的信息检索服务，并从中获得大规模的用户行为数据，为其提升信息检索服务质量、拓宽广告服务等增值业务奠定基础。可以说，大数据的价值正是被搜索引擎巨头率先发掘出来的。之前人们一直认为信息检索的核心是排序模型，并投入了大量精力改进排序模型，以求提升信息检索的精度。然而，随着越来越多的用户使用搜索引擎，搜索引擎公司逐渐意识到用户对结果的点击行为是一种非常好的反馈，利用海量用户的点击数据，研究人员使用排序学习的方法[3]，可以大幅度提升信息检索的精度。这是搜索引擎公司对其收集和索引的大数据的一种重要价值发现形式。

数据预测和数据分类被很多人认为是一种数据分析任务，其实，很多针对个体的数据预测和分类任务实际上是一种数据服务，它使用数据分析得来的预测模型，对个体数据实例进行预测，从而能够高并发地为大规模用户提供分类和预测服务，进而更好地体现出数据的价值。推荐系统是一种典型的数据预测服务。根据大规模用户的购买或浏览行为，使用数据分析技术学习得到推荐模型。当用户在电商网站选购商品时，这些网站会利用先前学习到的推荐模型，结合用户当前的行为，为用户实时推荐商品。推荐实际上就是一种数据预测服务，其发现的数据价值也非常高。

3 数据分析

数据分析是指用适当的统计分析方法对大量数据进行分析或建模，提取有用信息并形成结论，进而辅助人们决策的过程。在这个过程中，用户会有一个明确的目标，通过

“数据清理、转换、建模、统计”等一系列复杂的操作，获得对数据的洞察，从而协助用户进行决策。常见的数据分析任务又可以被进一步划分为描述型分析、诊断型分析、预测型分析、策略型分析①。

描述型分析的主要特点是对数据代表的含义进行描述性的揭示，通过数据统计分析揭示数据隐含的现象，从而帮助人们更好地进行决策。常见的描述型分析有对历史数据的在线联机分析处理（online analytical processing，OLAP）和对新产生数据的流式分析。OLAP一般采用SQL查询语句对结构化数据进行多维度的聚集查询处理。采用立方体模型组织和管理数据，即选取一些属性作为分析的维度（称为维属性或者维表），另外一个或者若干个属性作为分析的对象（称为事实属性或者事实表），在分析对象上，可以进行常规的统计操作，包括计数、求和、求平均、求最大/最小值等。OLAP的操作包括切片、切块、上钻、下钻等。为了更容易地理解分析结果，OLAP通常都有一个图形化的数据展示接口。例如，某连锁企业汇集了全部门店的商品销售明细数据，包括时间、门店、商品种类、销售额等。可以构建以时间、门店、商品种类等为维度，以销售额为目标的数据仓库。通过这个仓库，可以展示本月份按照门店统计的销售额（采用切片、切块操作），也可以在上述分析结果的基础上，进一步展示销售额最低的门店的各商品销售数据和本月份每周的销售数据（采用下钻操作），试图找到销售不佳的原因。

流式分析与OLAP的主要差别是要持续不断地对新产生的数据进行统计分析，在实时性要求比较强的应用中能发挥出重要作用。比如，通过传感器采集到的设备传感数据分析设备的运行状况，持续监控设备的状态，就属于典型的流式分析，其核心还是在于对最新的数据产生洞察。

除了流式分析，还有一类描述性分析也越来越受到人们的重视，即图数据分析。社交网络的发展让图分析能够发挥出越来越重要的作用，其主要目的在于分析图上节点（包括边）的影响关系、发现图的模式等。例如，在论文数据库DBLP的基础上，可以获得作者之间的合作关系，评估作者之间的相互影响程度。对于一个特定作者来说，可以分析哪些作者对其影响最大，还可以分领域（用关键字表达）、按时间等展示这种影响。

诊断型分析（diagnostic analysis）主要用来揭示一些现象背后的成因，因此，它比描述型分析更深入。很多数据挖掘方法与诊断型分析密切相关。比如相关性分析和因果关系的分析等，都是想通过对数据的深度分析揭示描述型分析所发现的某些现象背后的成因。例如，一个互联网金融公司使用描述型分析的商务智能（business intelligence，BI）技术，发现其某类产品在过去一个月的销售出现明显的下滑，分析人员需要借助一些诊断型数据分析方法，发现造成销售下滑的主要原因（比如，是否因为产品定价原因而影响了销售），分析人员一般在基于某些假设的前提下，对不同维度的数据进行相关性分析和因果性分析。复杂的甚至还会引入一些图分析技术，比如，通过对用户行为图谱的分析，辨别是否存在群体性恶意抵制的行为，影响商品在某些方面的销售。

预测型分析（predictive analysis）主要是使用机器学习技术，对现有的大数据进行深度分析，构建数据预测和分类的模型，从而更好地支持数据预测和分类服务。机器学习是一个比较宽泛的研究领域，近些年随着大数据的迅速发展，取得了很多令人振奋的进展。最为典型的就是深度学习（深度神经网络）技术[4]等，使得很多基于大数据的人工智能（artificial intelligence，AI）应用走上前台。深度学习是机器学习的子领域，已经在计算机视觉、语音、自然语言处理、游戏博弈等领域取得了巨大的突破。深度

① http://www.informationweek.com/big-data/big-data-analytics/big-data-analytics-descriptive-vs-predictive-vs-prescriptive/d/d-id/1113279

学习可以处理结构化数据，甚至是非结构化数据。深度学习模型最基本的单元是神经元。一个神经元接收输入X，输出$f(X \cdot W)$。其中X和W是向量，W是神经元的参数，$\cdot$是内积符号，f是一个非线性函数。神经元其实是一个非常简单的函数映射。深度学习模型就是这些众多简单神经元的组合。例如，数据首先传递给直接接收原始输入的神经元，这些神经元的输出再作为后面神经元的输入，以此类推。

目前主流的深度学习模型都是有监督学习，其模型的训练过程如下。

第一步，准备大量的带标签的样本$\{<x, y>\}$，其中x是数据（如一张图片），y是其标签（如图片的类别）。

第二步，定义从x到y的映射。这一步就是在设计深度学习模型的结构（或者说神经元的组合方式）。有人把深度学习模型的设计比喻成搭积木，每个积木就是一层神经元（神经网络层）。选择不同的神经网络层，每一层的输出作为下一层的输入，直到输出对标签y的预测。

第三步，定义损失函数。需要确定一个目标函数，衡量模型预测的标签和正确标签之间的差距。

第四步，训练参数。利用收集的训练样本集$\{<x,y>\}$训练模型中的参数。整个训练过程是以下3个步骤的不断迭代：正向传播，将x作为输入，通过事先确定的神经网络，得到预测的标签y'；反向传递，将y（正确的标签）和y'（预测的标签）代入损失函数，计算误差，通过反向传递，不断地将误差值从后向前传递，可以得到每个神经元以及神经元中参数的导数；调整参数，有了参数的导数，就可以使参数向负梯度的方向更新迭代，以求在下一次的迭代中减少损失函数的值。

例如，自动驾驶场景中的交通标志识别，这是2016年中国计算机学会大数据与计算智能大赛中的一个题目。交通标志检测是无人驾驶技术中一项非常有挑战性的任务，交通标志的正确识别对辅助定位和导航有决定性的作用。交通标志的种类繁多，其大小和角度不一致，受天气、光照等环境因素影响大，这使得对交通标志的检测非常困难，但好处是很容易获得大量真实场景下的图像数据用于训练和测试。获得该赛题一等奖的团队采用的正是神经网络模型，特别是通过将关键部位的图像进行放大增强等措施，提高了识别的准确率。

策略型分析（prescriptive analysis）也称指导型分析，是在分析过程中减少甚至排除人的参与，在给定目标的驱动下，直接帮助人们找到好的策略，作用于大数据应用，使得未来数据指标能够按照设想的某些趋势发展。它是数据分析的高级阶段，更能发挥出大数据的价值。然而，它也是最为复杂最难的分析任务，因为影响实际数据指标变化的因素可能很多，片面地分析往往达不到想要的效果。例如，交通拥堵问题是一个困扰很多大城市的难题。策略型分析就是要通过对城市交通数据、气象数据、就业数据、地理信息等进行综合的数据分析和建模，从微观层面到宏观层面制定合理有效的交通疏导策略，以缓解城市的交通拥堵问题。在很多社会大数据治理的问题上，策略型分析能发挥出很大的作用，它使得数据分析系统发现的策略在被执行后，能够让数据指标按事先预定的趋势发展，目前面临的研究挑战还非常大。

总之，数据分析一般基于大量数据和较为复杂的运算模型，其结果信息量通常很大，适用于宏观决策。而对于细节层面信息的获取，数据分析缺乏如索引和访问控制等方面的技术支持。如何在一个平台上，既支持宏观的分析，也支持细节的分析，是当今一个挑战的技术难题。

4 数据探索

数据探索是指针对目标可变、持续、多角度的搜索或分析任务，其搜索过程是有选择、有策略和反复进行的。它将以找到信息为目的的传统信息检索模式变为以发现、学习和决策为目的的信息搜寻模式。这样的搜索模式结合了大量的数据分析与人机交互过程，适合于人们从数据中发现和学习更多的内容和价值。

对于数据探索，用户可以在微观层面（数据搜索）和宏观层面（数据分析）之间进行自由切换，用交互式的方式探索并发现数据的价值。例如，某家环保研究机构收集了很多数据（如企业排放数据、城市机动车辆数据、天气数据、地理信息数据等），分析人员想利用这些数据分析北京地区雾霾的成因，这看上去更像是诊断型分析。但由于分析人员对雾霾成因还缺少深刻的认识，只能通过很多尝试性的分析策略分析数据，对比了北京雾霾天气和气温的关系，发现冬天雾霾更多，然而进一步分析发现气温和雾霾存在相关性，并不是决定性。因为可以从不同季节、不同年份的分析结果否定气温是主因的结论。分析系统可以提示分析人员对不同地域、不同时段、不同类型数据进行进一步细致的分析，这种交互式的带有引导式的分析过程，能够不断帮助分析人员发现一些重要的规律，进而缩小数据分析的范围。比如，当分析人员认为煤燃烧很可能是主因时，系统会自动推荐进一步的分析策略，如不同行业的煤燃烧情况、地域分布、时域分布等，引导用户进行深入数据探索。用户也可以随时切换到别的分析场景，比如汽车尾气等，进行更有广度的数据探索和分析。

在这个例子中，尽管分析人员想找到雾霾成因，但分析任务和模型一开始并不能确定下来，具体目标也比较模糊，分析人员需要在不断获取信息的过程中调整搜索和分析的目标。此外，分析人员还需要系统提供额外的信息进行引导，在引导的过程中，目标随时可能改变，这种改变的动机可能出于获取必要信息，也可能出于好奇心。

目前，随着大数据研究的兴起，探索式搜索这种交互式分析和探索数据价值的方式，逐渐引起人们的重视，还有很多问题等待研究者们进行深入的研究。

5 结束语

数据服务强调从微观层面获取满足用户需求的精准信息，数据分析强调从宏观层面为用户提供数据洞察，进而提供决策支持，而数据探索则需要在宏观和微观两个层面进行自由切换。**表1**总结归纳了数据服务、数据分析和数据探索3种数据价值发现方式的一些特点。

- 在用户信息需求层面，数据服务和数据分析都要求用户有明确的信息需求，相比数据分析，数据服务的信息需求更加单一，数据探索的信息需求则是不断变化的。
- 在搜索对象层面，数据服务的对象是数据集合内的某些元素，而数据分析的对象是整个数据集或其子集，数据探索则经常需要在整体数据和局部数据之间进行切换。
- 在观察角度层面，数据服务的角度是微观的，数据分析的角度是宏观的，数据探索以微观为主，也附带着一些宏观的数据分析任务。
- 在用户目的层面，数据服务侧重于查询资料和数据片段，数据分析的目的侧重于决策支持，数据探索侧重于对数据的学习和研究，以加深对数据的理解。
- 在交互模式层面，数据服务与数据分析主要是一次性的交互模式。但在交互式

场景中，它们也会遇到查询调整的问题，用户通过多轮交互满足信息需求，而各轮之间却是独立地查询或者分析任务。数据探索则必须是多轮迭代并存在关联的过程。

大数据蕴含大价值，数据服务、数据分析和数据探索是3个层次的数据价值发现方法。在很多应用下，这3类方法需要混合使用，才能更好地发现大数据的价值。人们在实际工作中，可以根据具体的需要进行选用。

表1 各类大数据价值发现方式比较[5]

比较项	数据服务	数据分析	数据探索
用户信息需求	单一、明确	多样、明确	多样、变化
搜索对象	点	面	点面结合
观察角度	微观	宏观	微观为主
用户目的	查阅资料	决策支持	学习、调研
交互模式	一轮	一轮或多轮	多轮

致谢：

博士生陈峻、赵哲对部分内容也有贡献。

参考文献：

[1] PAVLO A, ASLETT M. What's really new with newSQL[J]. ACM Sigmod Record, 2016, 45(2): 45-55.

[2] STONEBRAKER M. SQL databases v. NoSQL databases[J]. Communications of the ACM, 2010, 53(4): 10-11.

[3] LIU T Y. Learning to rank for information retrieval[M]. Berlin: Springer Press, 2011.

[4] GOOFELLOW I, BENGIO Y, COURVILLE A. Deep learning[M]. Cambridge: MIT Press, 2011.

[5] 杜小勇，陈峻，陈跃国. 大数据探索式搜索研究[J]. 通信学报，2015，36(12)：77-88.
DU X Y, CHEN J, CHEN Y G. Exploratory search on big data[J]. Journal on Communications, 2015, 36(12): 77-88.

作者简介

杜小勇（1963-），男，中国人民大学信息学院教授，博士生导师，教育部数据工程与知识工程重点实验室主任，中国计算机学会会士，《大数据》期刊编委会副主任。主要研究方向为智能信息检索、高性能数据库、知识工程。主持和参与多项国家核高基（核心电子器件、高端通用芯片及基础软件产品）、“973”计划、“863”计划、国家自然科学基金项目，近年来在SIGMOD、VLDB、AAAI、IEEE TKDE等国际重要期刊和会议上发表论文百余篇。

陈跃国（1978-），男，中国人民大学信息学院副教授，博士生导师，中国计算机学会数据库专家委员会委员、大数据专家委员会通讯委员。主要研究方向为大数据分析系统和语义搜索。主持国家自然科学基金项目2项，参与多项国家核高基（核心电子器件、高端通用芯片及基础软件产品）、“973”计划、“863”计划项目，近年来在SIGMOD、ICDE、AAAI、WWW、IEEE TKDE等国际重要期刊和会议上发表论文30余篇。

收稿日期： 2015-10-09

基金项目： 国家自然科学基金资助项目（No.61472426）；国家高技术研究发展计划（“863”计划）基金资助项目（No.2015AA015307）

Foundation Items: The National Natural Science Foundation of China (No. 61472426), The National High Technology Research and Development Program of China (863 Program) (No. 2015AA015307)