从大数据到大知识: HACE + BigKE

吴信东 [1, 2] 何 进 [1] 陆汝钤 [3] 郑南宁 [4]

摘 要 大数据面向异构自治的多源海量数据, 旨在挖掘数据间复杂且演化的关联. 随着数据采集存储和互联网技术的发展, 大数据分析和应用已成为各行各业的研发热点. 本文从大数据的本质特征开始, 评述现有的几种大数据模型, 包括 5V, 5R, 4P 和 HACE 定理, 同时从知识建模的角度, 介绍一种大数据知识工程模型 BigKE 来生成大知识, 并对大知识的前景进行展望.

关键词 大数据, 知识挖掘, 异构, 碎片化知识, 在线学习

引用格式 吴信东, 何进, 陆汝钤, 郑南宁. 从大数据到大知识: HACE + BigKE. 自动化学报, 2016, **42**(7): 965−982

DOI 10.16383/j.aas.2016.c160239

From Big Data to Big Knowledge: HACE + BigKE

WU Xin-Dong[1, 2] HE Jin[1] LU Ru-Qian[3] ZHENG Nan-Ning[4]

Abstract Big data deals with heterogeneous and autonomous multi-sources, and aims at mining complex and evolving relationships among data. With the fast development of data collection, data storage and networking technologies, big data analytics has become a hot topic for research and development in various fields. This paper starts with the essential characteristics of big data, reviews existing popular models for big data, including 5V, 5R, 4P and the HACE theorem. Also, from the viewpoint of knowledge modeling, this paper introduces BigKE, a big data knowledge engineering model for big knowldedge, and discusses the challenges and opportunities of big knowledge research and development.

Key words Big data, knowledge mining, heterogeneity, fragmented knowledge, online learning

Citation Wu Xin-Dong, He Jin, Lu Ru-Qian, Zheng Nan-Ning. From big data to big knowledge: HACE + BigKE. *Acta Automatica Sinica*, 2016, **42**(7): 965−982

随着互联网的不断发展, 我们可以收集和获取的数据以不可预计的速度增长. 尽管数据的收集、存储和处理技术还在不断进步并日趋成熟, 但基于如此复杂的数据背景, 我们仍然面临着许多分析和处理数据的问题与挑战. 因此, 大数据的分析及其应用成为了一大科研热点. 对大数据的本质特征的概括始于 2001 年美国高德纳公司 (Gartner Group) 的分析师 Laney 等提出的 3V 特征[1]. 之后 IT 业界的科技大厂 IBM 对其进行了应用并加以扩充, 获得了 4V 或 5V: 包括了大数据巨大的数据量 (Volume)、快速的分析和处理速度 (Velocity)、多样化的数据种类和数据来源 (Variety)、对商业领域巨大的价值 (Value) 和其隐藏知识的真实性 (Veracity)[2]. 大数据广阔的应用背景, 使其不仅在科研领域, 乃至于在商业、政治、经济、医疗和文化等多领域内, 都在引发和领导一场变革.

在网络 2.0 时代, 用户已经从被动的信息接受者转变为主动的创造者. 一些数字可以说明这个事实: 美国每年的线上零售交易记录数量、推特网的发帖数量、各大物理实验室和天文望远镜观测记录值, 就足以产生大约 1.2 ZB 的电子数据, 由此, 美国国家科学基金会 (National Science Foundation, NSF) 在大数据领域的投入也日益增多[3]. 我们再从数据产生速度来看: 全球范围内, 每一秒产生约 2.9 百万封电子邮件, 同时, Youtube 网上可以上传 2.88 万小时的视频数据. 这些数据信息, 足够一个用户昼夜不息地看上几年.

这些来自商业、天文、科学和工程等多领域的可用数据规模不断扩大, 数据从数兆兆字节 (Tera-byte, TB) 到数千兆字节 (Peta-byte, PB) 的爆炸式增长, 对数据和信息的获取、存储和处理提出了新

收稿日期 2016-03-03 录用日期 2016-05-31
Manuscript received March 3, 2016; accepted May 31, 2016
国家重点基础研究发展计划 (973 计划) (2013CB329604), 国家自然科学基金 (61229301), 教育部长江学者和创新团队发展计划 "多源海量动态信息处理" (IRT13059) 资助
Supported by National Basic Research Program of China (973 Program) (2013CB329604), National Natural Science Foundation of China (61229301), and the Program for Changjiang Scholars and Innovative Research Team in University (PCSIRT) of the Ministry of Education of China (IRT13059)

1. 合肥工业大学计算机与信息学院 合肥 230009 中国 2. 佛蒙特大学计算机科学系 伯灵顿 VT05405 美国 3. 中国科学院数学与系统科学研究院 北京 100190 中国 4. 西安交通大学人工智能与机器人研究所 西安 710049 中国
1. School of Computer Science and Information Engineering, Hefei University of Technology, Hefei 230009, China 2. Department of Computer Science, University of Vermont, Burlington VT05405, USA 3. Institute of Mathematics, Academy of Mathematics and Systems Science, Chinese Academy of Sciences, Beijing 100190, China 4. Institute of the Artificial Intelligence and Robotics, Xi′an Jiaotong University, Xi′an 710049, China

的要求. 在网络 2.0 和工业 5.0 时代的共同作用下, 我们应当注意到, 这个庞大的数据量有很大一部分是数据和信息在向知识的转化过程中生成的, 这实际上就是我们主张的大数据知识工程的基本思路. 文献 [4] 中所说的 "知识自动化" 这一词源于 Fish 于 2012 年出版的 *Knowledge Automation* 一书[5], 这和我们的大数据知识工程的基本思路是一致的. 人类直接生产的数据形成的网络流量不足大部分网站流量的 37 %, 大部分的网络数据流量是数据和信息在向知识转化过程中生成的二次数据. 这种二次数据形成的过程可以理解为基于知识的服务 (Knowledge-based services, KBS), 这与基于位置的服务 (Location-based services, LBS)、基于信息的服务 (Information-based services)、基于情报的服务 (Intelligence-based services), 以及基于任务的服务 (Task-based services) 相类似[4]. 大数据的自动化产生, 大数据技术的广泛应用对有用知识的自动产生和获取提出了进一步的要求: 更高水平的大数据知识工程, 更好的 "恶意 (Malicious)" 过滤机制以及更合理的知识评价体系.

近几年, 人们对 "大数据" 一词似乎不再是那么陌生. 在数据挖掘和人工智能等科研领域内, 大数据的扩散速度随着相关研究的增多而加快. 研究者们逐渐认识到, 具有大数据特征的数据资源, 除去其固有的庞大的信息量, 似乎还可以挖掘出无法用我们现有的计算标准得出的隐含的 "大知识", 这些有用的知识我们无法快速、高效地处理和分析, 因此产生了一系列新的问题和挑战. 值得注意的是, 大数据的价值绝不仅仅是巨大的数据量而已, 虽然仅凭数据集的扩充, 确实能提升现有的统计和分析工作的精确度. 但是, 对于大知识的发现和表示, 仅仅通过提升对庞大数据的收集和存储能力是不足够的, 这些数据还包含对数据表示等方面的可伸缩性、数据分析算法本身的改进需求[6].

海量数据的收集和大数据知识发现技术可以应用到多个领域. 在科学研究方面, 目前国内外的天文学研究中海量数据的收集和应用已经非常普遍. 举例来说, 美国斯隆数字巡天项目 (Sloan digital sky survey, SDSS) 中所产生的海量的天文数据远远超出了预期, 至今其所收集的数据已多达 140 TB 之多[7]. 专业的科研领域内, 除了天文学的大量观测数据的应用, 移动终端等传感器产生的大数据也颇为重要: 大数据地理信息系统 (Geographic information system, GIS) 的构建、地震的勘探、雷达等非结构化信息的应用价值都不容小觑. 从政府推进力度来看, 美国将大数据作为事关国家战略和国家核心竞争力的问题, 并于 2012 年 3 月推出了 "大数据的研究与发展倡议", 这也让人看到了大数据应用广阔的前景. 除去科研工作, 文化领域也受到了大数据的影响. 微软纽约研究院的经济学家 David Rothschild 利用大数据技术, 成功预测了 2013 年 24 个奥斯卡奖项中的 19 个, 这一实例成为人们津津乐道的话题. 2014 年, David Rothschild 再次成功预测第 86 届奥斯卡 24 个奖项中的 21 个, 大数据知识的价值由此可见一斑. 除了各行业领域内的应用, 大数据精准的预测和分析手段、对用户的行为模式和偏好行为的挖掘、对商业和金融决策的意义, 以及在信息安全方面都能给现有的数据和信息处理模式带来变革.

然而, 利用现有的数据处理手段, 我们无法发挥出大数据真正的价值, 大数据的本质特征为我们在分析和应用上带来了一系列的问题. 大数据带来的挑战问题, 已经不仅仅是单纯意义上的数据规模的巨大, 还包含了对大数据分析技术的改进问题, 从而满足越来越多样化的对个性化服务和知识导航的需求. 接下来我们需要考虑的是如何从海量的数据中提取和分析出有价值的知识, 这也是对大数据进行研究的重要意义之一.

从数据量来说, 大数据庞大的数据量已经无法通过已有模型和计算平台简单处理, 面对大数据的数据规模, 我们无法单纯依靠并行计算和硬件方面的提升去突破计算平台上的瓶颈. 例如, 网络、电视、报纸等众多数据来源产生了不同结构的异构数据, 我们的首要挑战就是从这些看似杂乱无章的数据中提取出真正对我们后面的工作和预测有价值的数据信息, 选择合适的过滤机制[8]. 面对铺天盖地的数据资源, 我们需要的不再是通篇的文字、声音或者是图像信息, 数据的规模和数量在不断增长, 但无用数据的存在导致数据的价值并不会成比例增长. 针对这个问题, 现有的筛选机制对大数据的提取和分析显得尤为困难和低效. 由此, 在大数据环境下的数据的预处理和清洗也具有更高的要求. 数据的清洗过程既要过滤无用的数据, 也要保留对大知识提取有用的信息. 对大数据的知识处理来说, 通过一个稳定高效数据计算和清洗平台, 经过数据预处理过程, 得到高质量的数据集合进行下一步分析是关键的一步.

从大数据的产生和获取来源来说, 尽管网络规模的扩张为我们获取信息带来了便利, 但复杂网络结构和获取信息途径的多样化, 使得数据的异构问题日益凸显. 异构数据在数据的存储和表示上产生了困难, 单一的数据表示和存储已经无法满足需求. 数据的分析工作的价值远远高于简单的定位和识别, 数据间复杂的语义联系以及不同结构的数据, 需要我们寻找一种标准化的数据的表示方式. 标准化的数据表示形式的定义本身就存在相当大的挑战, 这

也会涉及到在对异构数据的集成过程中需要对大规模数据集进行数据的转换[9]. 以社交网络中的大数据分析为例, 通过对网络结构的刻画形式的改进, 我们集成多个网站上的异构自治信息源, 可能包括用户发送的微博、评论或者是上传的图片、音频等信息, 足以描绘出一个合理的网络结构描述数据间的语义关联.

从我们分析大数据的最终目的来说, 落实到实际应用上, 我们关心的是大数据能够提供的服务, 这些服务需要分析数据间的结构和关联, 面对简单的数据, 数据之间不存在动态的演化, 相应的知识挖掘和数据关联就易于发现和表示. 因此, 从以数据流形式到来的大数据中获取知识, 到近期的大数据知识工程模式, 都具有实时数据处理和更新数据的动态演变内容的需求, 其所得到的知识相较于单一数据也更具价值. 举例来说, 包括社区智能需求和提升个性化服务[10] 等以大数据知识为基础的导航服务, 在社会服务和个性化需求上具有更精准的导向.

通过大数据知识工程, 我们旨在获取大数据中的 "大知识": 大知识从异构、自治的大数据开始, 挖掘包括数据流和特征流的多源海量数据以发现数据对象之间复杂且演化的关联, 通过大数据知识工程, 以用户需求为导向, 提供具有个性化和实时使用价值的知识服务. 大知识源于大数据, 通过大数据知识工程的方法进行提取和处理. 数据流和特征流有别于传统的单个静态数据源, 以流的形式快速到来的大数据对实时性具有很高的要求, 数据之间的关联性和特征形成的特征流数据提出了新的数据挖掘和处理问题. 因此, 为了获取大知识, 我们需要了解大数据的本质特征和现有的大数据的一些挑战问题.

针对大数据的几大本质特征, 研究者们提出了几种目前被广泛接受的大数据模型, 包括 5V、5R、4P 和 HACE 定理. 这几个模型分别从不同的角度提出了在进行大数据分析和处理的过程中需重点关注的挑战, 其中 HACE 还对大数据挖掘提出了一种可行的多层框架. IBM 的 5V 模型着眼于大数据的核心特征, 注重以先进技术提高大数据的质量以得到有价值的知识, 每个 V 的维度都包含大数据工作中某一方面的严峻挑战[11]. 5R 模型从大数据的管理建模的角度, 注重大数据对于商业决策和商业回报的价值, 同时它也是本文介绍的大数据知识工程模型 BigKE 的支撑[12]. 4P 医学模型基于现有的 4P 医学模式, 包含预测性 (Predictive)、预防性 (Preventive)、个体化 (Personalized) 和参与性 (Participatory) 四个维度[13]. 4P 医学模型在强调专家知识的重要性的同时, 着眼于社会网络和个人信息的参与性. 然而, 专家知识和新加入的社会与个人因素同样产生了异构自治数据源和碎片化知识提取的问题, 这为大数据的数据集成以及碎片化知识的融合提出了新的技术要求[14]. 大数据的 HACE 定理考虑了大数据的本质特征, 包含了海量、异构、分布和分散式控制的自治源、数据间复杂和演化的关联等大数据的典型特征[15], 但是 HACE 定理也没有提出系统地解决碎片化知识的非线性融合问题的方法.

针对以上现有的大数据模型及其存在的问题, 本文从知识建模的角度介绍大数据知识工程模型 BigKE. 该模型针对海量异构数据中的碎片化知识的非线性融合问题, 提出了从数据流和特征流的在线学习为开端, 利用非线性知识融合手段形成有价值的知识图谱, 并以此为基础以满足需求为导向的知识服务的三层知识工程框架. BigKE 模型能够一定程度上应对大数据特征带来的知识工程的挑战, 从而在碎片化知识中提取出有价值的大知识, 最终满足大数据用户的个性化需求.

本文安排如下: 第 1 节介绍大数据的本质特征和知识工程的研究进展, 包括对现有的 5V 模型、5R 模型、4P 医学模型和 HACE 定理进行阐述, 这一节中对 HACE 定理的大数据多层处理框架做较为详细的介绍. 第 2 节, 介绍大数据知识工程的概念, 并对大数据背景下知识工程研究中的挑战问题做一些阐述. 第 3 节中, 我们从知识建模的角度, 详细介绍一种大数据知识工程模型 BigKE. 第 4 节中, 我们总结现有的大数据模型以及大数据知识工程模型 BigKE, 讨论 BigKE 模型后大知识的挑战问题和应用前景. 最后, 我们对从大数据到大知识的过程做出总结.

1 大数据特征与知识工程研究进展

1.1 大数据的本质特征

随着云计算、互联网、各种移动设备与物联网的发展和普及, 大数据已经成为一个耳熟能详的概念. 互联网的扩张, 使得人人都能感受到大数据的存在, 但各个领域对 "究竟什么是大数据" 或者 "具备怎样特征的数据可以称为大数据" 的问题, 都有各自不同的定义和理解. 早在 20 世纪 90 年代, 被称为 "数据仓库之父" 的 Bill Inmon 就开始关注大数据了, 只是当时的大数据还被称作海量数据. 维基百科和国际数据公司 (International Data Corporation, IDC) 对大数据分别做出了各自的阐述[16−17]. 简而言之, 大数据是无法在合理的时间内, 利用我们现有的数据处理手段, 对其进行诸如存储、管理、抓取等分析和处理的数据集合.

随着大数据科研项目的深入展开, 我们对大数据的定义, 以及对大数据蕴含的知识价值的认识, 从

最初单纯意义的 “大体量” 逐渐有了更深层次的阐述. 实际上, 大数据之 “大” 包含了数量与其蕴含的知识的价值两个方面, 大数据知识的目标和价值体现在对数据进行分析和处理之后, 加工后的数据在商业、科学、工程、教育、医疗和整个社会领域内的决策有着重要的导向意义[18].

为了从大数据中获取有价值的知识, 我们首先需要了解大数据的特征. 大数据的本质特征与大数据的来源密切相关. 首先值得关注的是大数据的大数据量. 随着互联网、云计算、物联网等技术的发展, 网络空间中数据的规模不断增加, 数据的计量从 GB、TB、PB 增长到 EB 和 ZB 的规模. IDC 研究报告显示, 全球大数据的数量规模在未来 50 年内会增加 50 倍, 管理数据仓库的服务器的数量将增加 10 倍以适应于大数据数量规模的 50 倍增长[19]. 在此之前, 由于数据的来源和数据的形式较为单一, 数据的获取、存储和挖掘的方法也相对比较单一, 从数据中获取知识的工作的复杂度也没有提升. 大数据的处理和知识发现与获取, 对算法的实时性具有较高的要求, 这也是由于大数据的海量特征. 实时处理的数据计算方法通常和流式计算相结合, 并且采用查询分类计算以提高响应的性能. 而传统的批处理计算和复杂数据挖掘计算则是非实时计算, 这就无法与大数据的海量特征相适应, 对大数据的处理和计算平台有了新的要求和挑战.

随着多种新型的数据获取渠道的出现, 不仅仅是音频、视频、广播、电视等多种媒体的混合, 包括复杂的网络在内的信息来源, 都显示出大数据的一个典型特征: 异构和多维度. 高维大数据的分布还产生了稀疏子空间聚类的问题. 大数据在高维通常分布在多个低维子空间的并上, 因此高维的数据在适当字典下的表示具有稀疏性[20]. 这需要我们寻找到合适的处理高维数据的聚类和分类的方法. 举个例子来说, 如果发生了一个热门的新闻事件, 那么在网络、电视、报纸等多个平台上就会引发热议. 大众对于事件的评价标准和意见各不相同, 信息和数据产生的形式可能是微博、视频、音频等. 不同的信息源产生的数据一般没有使用统一的数据收集、记录、存储和表达形式, 这使得异构的大数据在处理的过程中产生了诸多问题与挑战, 对数据的转换和集成提出了更高的要求.

多样化的数据来源产生了大数据的异构性问题, 当大数据投入到实际应用之中, 各个数据源在产生和收集数据的时候相互独立, 如同互联网中的自治系统, 能够自主地决定本网络中使用何种路由协议一样. 这样的数据特征显示出大数据的另一个本质特征: 分布式和分散式控制的自治数据源. 这些自治的数据源没有集中式控制, 能够自主地决定产生和收集的数据存储和表示的形式. 这在一定程度上使得数据之间的关联度有所下降, 也在一定程度上提升了数据和用户信息的安全性. 但这些自治源仍然带有分布式和分散式控制. 随着云计算和云终端的普及, 分布式控制方面的应用融入到生活的各个方面, 同样也保障了对于大数据惊人的规模增长同步的数据处理和分析能力的提升[21]. 在工业运用上, 以太网的计算机分散式控制也在电力系统上得到了应用[22]. 分散式控制过程中数据的安全提升了、数据处理的简便性增加了, 这使得在复杂的大数据环境和数据规模较大的控制环境下, 能够很好地适应数据分析和处理的需要.

同样, 由于大数据庞大的数据规模及其数据源的异构性和自治性, 数据间的关联显得更为复杂, 随着时间的推进, 数据之间的关联也会发生演化. 网络环境下的大数据信息则显得更加难以发现, 数据下隐藏的关键信息可能会有所重合, 并随着时间的推进发生演化. 大数据之间复杂和演化的关联的发现和早期集中式控制的信息系统有着明显的区分, 数据的内容无法再简单地由几个给定的特征值表示出来, 异构的数据无法统一其表示形式, 因而数据关联的发现和处理难度大大提升. 大数据的这一特征在社交网络中得到了充分的表现, 用户之间敌对或者友好的关系, 为我们对数据的聚合和分类提供了可能性[23]. 社交网络拥有庞大的用户群, 每日产生大量的图片和文字信息, 网络上充斥着各种形式不一的文本和音视频信息. 微博、推特、豆瓣等常见的社交平台上朋友圈之间和粉丝之间的联系隐藏了各种有用的信息, 包括事件的预测、真实性等. 用户在搜索引擎中搜索的信息, 也如实反映出了社交网络中数据的流动和演化倾向.

1.2 大数据特征: 5V 模型

2001 年, Gartner 公司的数据分析师 Laney 首次从大数据特征的角度明确定义了大数据, 强调了大数据的 3V 特征, 即海量(Volume)、快速 (Velocity) 与多样化 (Variety)[24]. 在 3V 的理论基础上, IBM 公司相继提出了大数据的 4V 和 5V 模型, 新加入了大数据的真实性 (Veracity) 与价值 (Value) 维度[2, 25]. IBM 的这种 5V 模型同样是着眼于大数据的本质特征, 反映出大数据规模巨大、数据的产生速度极快、数据的结构和框架不一致、数据的安全和隐私问题. 因此, 我们需要更优良的数据运算方法和平台, 以面对快速产生的数据流数据并给予更快的实时响应. 数据的有效性和真实性依赖于数据的质量, 高效地对数据和数据中的知识进行评估对此至关重要, 质量较好的数据对我们后期提取大知识和做出个性化服务具有重要意义, 高质量的数

据和知识也能够体现大数据的价值所在. 有效的数据管理和分析使得我们能够做出更好的商业决策, 甚至在医疗、隐私保护等多个领域都可以得到应用. 最经典的实例莫过于 "谷歌流感趋势 (Google flu trends, GFT)", Google 利用其用户的搜索数据, 准确预测了流感趋势的产生, 其预测的速度和准确度都远远高于美国疾病控制与预防中心 (Centers for Disease Control and Prevention, CDC) 检测报告的结果[26]. 谷歌的某些搜索关键词可以很好地表示流感疫情的现状, GFT 的工作原理就是利用经过汇总的谷歌搜索数据来估测流感疫情.

5V 模型较之于 3V 模型更着眼于使用先进的技术以提高数据的质量并且能够更加充分地探索大数据. "真实性 (Veracity)"[27] 和 "价值性 (Value)"[28] 结合了 3V 特征显然更加全面. IBM 公司对大数据特征的概括和应用更多的是在商业决策领域, 它更多地关注依据大数据知识做出的商业决策, 对于提高商业收益是否有现实的指导意义和价值. 但是, 即使是如同谷歌的流感预测这样典型的大数据应用实例, 也不会对决策产生完全的保障. 其主要原因不是由于大数据的价值被高估, 而是因为人们对大数据价值所在产生了误解: 大数据价值不在于其 "大小", 而是利用创新的数据分析方法来处理和分析数据[29]. 同样地, 大数据的价值不仅在于 "大" 也在于 "数据" 的价值. 而大数据的价值往往伴随着稀疏性的特点, 从 3V 模型到 5V 模型的扩充, 也反映出不当的大数据挖掘和处理所隐藏的陷阱. 接下来我们更多需要考虑的是在数据的分析和提取中, 利用更好的数据分析算法来提升数据的真实性和价值. 虽然 5V 模型对大数据的特征做了很好的阐释, 但是对于大数据本质特征所导致的问题和挑战并没有做出过多的描述和给出解决思路.

1.3 大数据管理与商用 — 5R 模型

从大数据中获取知识的过程, 如果采用数据管理的视角, 可以得到 5R 模型. 5R 模型由 Stidston 提出[12], 包括对大数据相关的 (Relevant)、实时的 (Real-time)、真实的 (Realistic)、可靠的 (Reliable) 以及投资回报 (Return on investment, ROI) 五大特征的阐述. 从 5R 模型的内容来看, 它和 5V 模型具有类似的地方. 它们都着眼于大数据的本质特征, 相比较而言, 5R 是基于商业用途而提出, 它对于大数据的五大特征的描述是基于数据管理在商业上的应用进行阐释. 从数据管理的角度来看待大数据, 其关键在于数据的组织形式. 大数据的海量多源异构特征已经得到了普遍的认可, 针对这些特征, 采取一种怎样的数据组织形式以提升数据收集、存储、处理和应用的效率, 获取对商业发展与决策具有价值的 "知识", 是 5R 模型中提出的需要解决的问题. 数据的组织和管理形式经历过人工管理、文件系统和数据库系统的发展历程, 对传统数据的组织已经满足用户的使用需求. 但是在大数据的背景下, 传统的关系型数据库技术对以数据流形式到来的巨型数据已经不再适应.

基于 5R 模型背景下的大数据管理系统的研究也成为一个热点并取得了一定的进展. 举例来说, Google 在网络规模的数据量下, 其采取的数据管理和分析方法 — 谷歌文件系统 (Google file system, GFS)[30] 具有较简单的思想. GFS 为客户端提供相似的操作系统水平上的字节抽象, 它对于非常大的文件的内容可以在众多的计算机之间跨平台共享, 且不需要创建共享集群, 这就使得硬件的消耗大大降低[31].

值得关注的是 5R 模型中的投资回报 (ROI). 许多的大数据项目最初关注的重点只是数据本身的利用, 而没有认识到对数据的利用怎么与整个商业计划相适应, 忽略了数据之下的知识的价值[32]. 尤其是对于投资回报 (ROI) 的关注显得很匮乏, 大数据项目中数据的来源和知识的获取应当提供最低的成本计划, 以对最终获取的知识进行价值评估. 对于一些数据层次本身就具有非常高的价值的项目, 项目本身就具有大数据的特征. 如果缺少了投资回报的评估, 我们就无法得知数据的价值与从某一个大数据项目中获取知识的项目的可行性, 无法评估在知识获取的过程中所花费在人力、软硬件等方面的投资是否具有意义.

5R 模型提出的大数据管理的实时性要求 (Real-time) 也是大数据分析的一个方向, 它和 5V 模型中的 Velocity 相契合. 在第 1.1 节中提及了大数据的本质特征含有分布式的特点. 在大数据的数据管理结构中, 目前普遍使用到的是分布式的文件系统和分布式数据库, 其中, Hadoop distributed file system (HDFS) 是比较具有代表性的分布式文件系统[33], 其较高的容错性适于部署在廉价的机器上, 和传统的分布式文件系统有着显著的区别, 它为用户提供高吞吐量的数据访问, 同时, HDFS 也面向流数据处理[34], 这些都利于我们在大数据规模下进行数据分析和处理工作, 高速处理海量数据成为了可能, 大数据管理的实时性要求得到了一定程度的满足.

1.4 4P 医学模型

知识工程概念的提出为专家系统 (Expert system, ES) 奠定了理论基础. 专家系统 (ES) 作为人工智能 (Artificial intelligence, AI) 的一个分支, 自 19 世纪 60 年代中期被提出以来, 已经被大量运用

到工程、科学、医学预测、商业等方面. 专家系统的基本思想是依赖于专业的知识, 对个性化应用做出预测等行为[35]. 然而, 随着大数据时代的到来, 仅依赖传统专家系统的领域知识提取大规模的异构数据集中的有价值信息, 这种方式的效率已经不能满足用户的需要. 基于大数据背景的知识工程, 为了提供更加智能的个性化服务, 在提取大知识的算法设计中, 需要考虑用户的社交和个人信息.

以大数据背景下的普适医疗应用为例. 普适医疗 (Pervasive healthcare)[36] 借助普适计算技术, 形成覆盖服务区域内各个医疗机构、家庭和个人的信息网络. 信息化的推进使得电子病历等一系列电子数据显现出大数据的特征, 同一种疾病的发病原因的多样化、同一种疾病采取多样化的治疗方法, 这些海量的异构医疗数据中同样隐藏着有价值的医疗知识. 针对这一问题, 4P 医学模型[37] 随之产生了. 在医学领域, 4P 医学模式的内容包含了预测性 (Predictive)、预防性 (Preventive)、个体化 (Personalized) 以及参与性 (Participatory) 四个维度. 这种新型的医学模式更强调病人个人, 以及周围亲属、朋友的参与和主动性, 强化个体生活行为对治疗和预防过程的干预. 由 4P 医学模型引申到大数据环境下, 我们发现对于个性化服务的设计和分析来说, 用户个人的行为因素、用户的参与度对用户数据的影响、数据的来源和专家知识的参与, 这三者是同样重要的. 可以说, 4P 医学模型的提出背景离不开大数据.

我们将 4P 医学模型与现有的大数据应用项目对比, 可以看出, 个体行为的重要性日益凸显, 病人的经历和治疗过程也成为知识的重要组成部分. 同 4P 医学模型提出的 “个体化” 与 “参与性” 相对应, 现代医学强调因人制宜, 包含了概念更新、理论框架的构建以及实践应用等一系列的创新举措, 这为从新的角度切入个体化诊疗的实现提供了可能[38]. 在注重用户个体性的同时, 我们也可以发现不同个体之间的相似性, 利用标签和聚类等数据处理手段, 将特定的用户和特定的行为表现相对应, 发现大数据下多个用户的相似的行为模式, 发现不同的个体与某一特定症状的相关性, 从而提高普适医疗信息管理和服务系统的准确性.

与现有的医疗系统相比较, 在大数据的背景下, 4P 医学模型对个性化医疗服务显然要更加适用, 它所提出的四个角度, 同大数据的本质特征也是相对应的. 专家系统对领域知识的依赖, 使得数据的来源过于单一, 会产生一系列的问题. 4P 医学模型中的 “预测性” 和 “预防性” 两个维度强调了先进医疗手段的重要性[39]. 然而对于普适医疗系统的应用来说, 个性化的服务更注重专家知识要和病人个体信息一致. 4P 医学模型将个性化的服务与预测相结合, 从而为病人提供基于大数据的个性化健康建议, 同时, 在诊断和治疗过程中的数据也被同时记录下来. 这种普适的个性化医疗服务已经渐渐渗透到生活中, 使得大数据和个人生活的关联显得不再遥不可及.

基于 4P 医学模型, 具备个性化诊疗功能的医疗系统的实现, 其核心技术在于融入了个性化的知识图谱. 专家系统相对个性化医疗系统而言, 数据和信息相对结构化, 虽然信息的处理和分析在一定程度上达到了较高的自动化水平, 但个性化知识的自动获取、分析和传播将会是更高的挑战. 目前, 网络空间里的许多信息系统正在越来越多地体现出 “人” 的智能. 这一趋势必然导致对大数据知识工程的更高要求.

为了向医疗服务提供者和医疗服务消费者提供有价值的和个性化的医疗服务, 需要挖掘海量医疗数据中的医疗知识, 这也是普适医疗信息管理与服务的关键技术与挑战问题. 4P 医学模型的启发性意义在于对病人的个人信息和异构的医疗信息源的处理, 以基于社会计算的普适医疗信息管理与服务体系 (Pervasive medical information management and service systems, PMIMSS) 为例, 现代的医疗服务模式涉及到医疗信息共享与集成、医疗知识发现与服务、医疗服务质量评价机制、个性化医疗服务推荐机制以及人与医疗信息系统交互的可信机制[36]. 这类系统的架构以及关键技术的出发点和设计理念, 与大数据的本质特征相匹配, 并且与知识工程的个性化服务推荐的目标相一致.

除了 PMIMSS, 还有其他个性化医疗服务的应用实例包含 4P 医学模型的思想. 比如, 医疗服务的移动客户端渐渐普及, 研究人员利用移动客户端的平台发布一系列的健康激励措施, 发送提醒大众关于疾病的预防等普适医疗信息[40]. 如果从用户的客户端中抽取有用的信息, 这些信息可能涉及运动频率、体重、社交活动等多方面的信息, 获取用户个人信息是碎片化的, 如何利用数据库中的专家知识对不同的用户信息进行有效的分析将会是知识集成的关键. 大数据在普适医疗的应用, 从技术层面来看, 其关键技术依赖于个人、社交信息以及专家知识等多源异构的大数据知识的融合[41]. 再比如, 患有某种特定疾病的病人会形成社交圈或者社区媒体, 病人们在社交网络中交换彼此的治疗进展或者患病信息, 这些信息作为整个社区的经验在社交网络中被分享. 大数据在病人和医生、病人和病人、医生与医生之间传播并产生一定的演化, 形成复杂的数据联系[42]. 同时, 这些涉及用户个人信息的数据, 需要结合已有的专业知识进行综合分析, 从而给出准确的预测和医疗建议. 专家知识可能来自于专家的建议、

医学著作和临床数据, 而用户个人信息的来源则更加多样化. 对这些大数据中所获得的大知识的提取与融合, 需要的大数据算法面对的是多源多样化的数据.

1.5 HACE 定理

大数据的 HACE 定理指出, 大数据始于异构 (Heterogeneous)、自治 (Autonomous) 的多源海量数据, 旨在寻求探索复杂的 (Complex) 和演化的 (Evolving) 数据关联的方法和途径. 5V 模型和 5R 模型介绍了大数据的本质特征, 4P 医学模型是大数据与普适医疗结合的实例. 接下来, 我们从大数据的本质特征介绍 HACE 定理提出的一种多层的大数据处理框架, 该多层框架分别从大数据的来源、大数据的复杂的数据结构以及数据之间的关系这三方面来描述[15]. 从大数据的来源来看, 异构和自治是大数据中多个数据源的最本质特征, 如盲人摸象中的每个盲人、物联网中的各个传感器和万维网上每位作者和读者, 他们可能用不同的语言 (中文、英文等)、不同的媒体形式 (文本、图像等) 和不同的表现形式 (如英国英文的 31/12/15 和美国英语的 12/31/15) 来描述和处理他们各自的信息. 大数据分析的最本质目标是探索异构、自治的多源海量数据中复杂且随时间和空间演化的数据关联.

依据 HACE 定理对大数据特征的阐述, 可以形成一个大数据的三层构架 (见图 1). HACE 定理的创新在于, 它把大数据的处理框架从单层扩展为多层. HACE 定理给出的多层处理架构关注对大数据的运算、大数据之间的语义联系和应用知识、大数据的挖掘算法设计[42]. HACE 定理给出的多层大数据处理框架本质上涵盖了分析大数据的科学方法, 下面我们给出每一层的细节介绍.

在构架的第一层中关注的是大数据计算平台. 对大数据的知识挖掘与分析, 首先是大数据计算的存储和计算问题. 在传统的处理方法中, 为了提升数据的运算能力, 我们可以从计算机硬件的方面加以改进: 利用密集型的计算单元, 或者是依赖高性能计算机提高抓取和计算大数据的能力. 在小规模和中型规模的数据量下, 我们可以仅通过硬件的提升来改进数据存储和计算的能力, 并达到数据的实时处理. 在大数据的海量多源异构的特点下, 传统的思路行不通了. 举个例子, 多个数据源中数据的采样和聚集就为我们的挖掘工作生成了一定的困难, 凭借少量计算机和传统的并行运算无法处理. 无论是采取流水线作业达成时间上的并行计算, 还是采用多个处理器达成空间上的并行, 虽然它们已经在诸如稀疏矩阵和迭代算法的运用中得到普及[43], 但对于大数据来说, 数据的稀疏性表现在一个较高的维度空间, 传统的并行算法并不是很有效, 尤其是对于以流数据形式到来的数据, 实时处理是非常困难的.

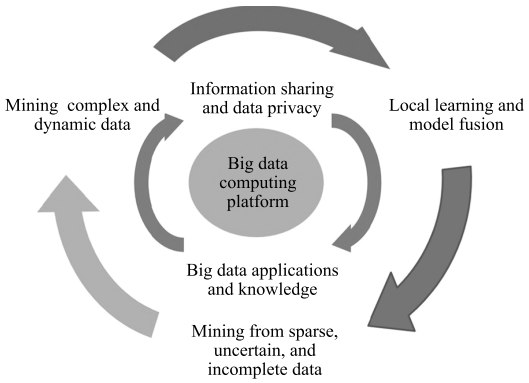


图 1　大数据处理框架的修改版[15]

Fig. 1　A big data processing framework updated form[15]

在 HACE 定理的第 1 层数据挖掘平台中, 提出使用带有高计算性能的集群计算机 (Cluster computers). 与中小规模数据集上的计算平台相比, 集群计算机上的每个计算节点都可以并行处理计算任务, 使得单个计算机的计算量有所降低, 从而减小对每个计算节点的硬件的依赖性. 利用这种结构的最典型的并行计算工具是 MapReduce. 谷歌的 MapReduce 模型是为了并行计算而提出的一种编程框架, 它将一个大规模的数据集上的计算任务拆分成多个小任务, 使得大规模数据集上的计算变得更加高效[44]. 传统的数据存储和处理工作, 使用最广泛的是关系型数据库结构. 但是大规模的数据下, 许多有用的信息隐藏在非结构化数据中, 诸如邮件、微博、视频等. 在这方面可以运用的技术包括 NoSQL 和谷歌提出的 “大表” (BigTable)[45]. BigTable 用分布式数据库存储系统管理大规模数据, 它将数据结构简化为键值之间的一种映射关系, 使得数据规模的大小和计算的延迟时间在 BigTable 中都得到了满足.

HACE 定理的第 2 层架构是大数据的语义和应用知识, 包含数据共享与隐私、领域和应用知识的问题. 第 1 层架构提出了集群式的大数据计算平台, 解决了对流数据存储的计算问题之后, 我们需要分析大数据中的隐含知识. 在对大数据下隐含知识的分析过程中需要数据的共享. 从数据的安全性来说, 由于大数据中包含大量的敏感信息, 或者是用户的一些不合法的数据操作, 都会影响到数据共享的效果, 并带来一些信息隐私的问题. 个人信息包含在大数据中, 也会引发关于数据可信度的度量和评估问题[46]. 大数据自治的分布式和分散式控制与数据的隐私有密切的联系, 为解决这一问题, 目前已经产生了一些适用于分布式的文件系统. 还是以 Google 的 GFS 文件系统为例, 该文件系统基于一

台主机和若干个备有 Linux 操作系统的 PC 机群构成了一个集群系统. GFS 系统对于用户从主机上得到的 Metadata, 从相应的位置产生通信过程从而获取文件数据[47]. 分布式文件系统的产生, 激励了诸如 Hadoop 和 Hive 这样的数据平台的产生, 数据仓库的数据处理在不断优化的程序中得到了更好的处理和分析.

在 HACE 定理的第 2 层架构中, 为了保护个人隐私信息, 同时提高所提取知识的可信度, HACE 主要提供了两种解决思路: 从数据存储角度, 对访问数据的权限进行限制可以一定程度上提高数据的可信度并减少对数据的误操作; 从信息共享的渠道来看, 对数据的一部分特征进行匿名化, 使得数据中包含敏感信息的部分不被公开或者进行一些模糊处理, 同样也可以起到保护隐私的目的[48]. 举例来说, 现有的关于数据匿名化的方法中, 使用最多的是 k 匿名方法[49], 用户通过对数据表的匿名工作指定一个 k 值, 限定发布的数据存在某些标识符与其他 $k-1$ 个具体个体没有方法区分开来, 从而保护了个体数据的隐私. 其次, 第 2 层架构需要考虑领域和应用知识[50], 它们能帮助我们辨别已收集到的大数据中哪些模式是用户希望去发现和使用的. 例如, 在医疗系统中对病人的数据信息进行分析时, 通过领域和应用知识可以识别我们需要的数据特征是诸如病人的血型、病史等信息, 从而刻画出有效的矩阵或者其他的数据特征表达方式, 同时为后期的数据挖掘工作清洗掉一部分无用的数据, 得到正确的数据语义联系.

HACE 的第 3 层从三个方面提出了大数据挖掘算法: 局部学习和多信息源的模型融合、稀疏不确定和不完整的数据挖掘、挖掘复杂的动态数据. 在网络数据的分析中, 出于保护数据隐私的考虑, 我们无法将从多个站点获取的局部数据简单地集成为一个集中式的站点. 因此, 大数据挖掘算法的设计存在许多挑战: 由局部数据特征到全局数据特征的转变, 稀疏的、不确定的和不完备的大数据需要有更高更快的实时性和准确性, 同时我们可能还要对缺失和不准确的数据进行填充[19]. 从数据建模的角度, 现有的文本模型, 包括向量空间模型 (Vector sapce model, VSP)[51]、潜在语义分析 (Latent semantic analysis, LSA)[52]、知识图谱 (Knowledge based graph)[53] 等, 都各有优劣, 比如, 在知识工程中知识图谱就能较好地表示实体之间的联系. 但这些基本模型无法满足动态环境中对整体大数据的特征刻画.

同时, 大数据之间的复杂的数据关联也随着动态数据而演化. 当数据流数据发生变化时, 我们需要考虑现有的数据结构是否适应于新的数据描述, 数据特征和数据变量在发生实时的变化. 对动态数据的挖掘, 对数据的变化如果只采取从头运行挖掘算法的方式, 就无法兼顾到实时处理的问题, 显然在动态数据中这不是一个有效的策略. 同时, 数据的动态改变导致了数据间关系的演化, 使得数据的规则和已获得的知识图谱无法匹配.

当然, 大数据的兴起不仅带来了挑战, 同时也促进了各领域的变革和发展. 例如, 研发针对社交网络之间的复杂联系以及演化关系的管理系统[54−55]. 诸如美国国防部高级研究计划局 (Defense Advanced Research Projects Agency, DARPA) 等机构也开始进行多种大数据研究项目, 其中包含多尺度的异常检测项目, 通过该项目研究大规模数据集的异常检测和特征化; Machine Reading 项目, 顾名思义, 该研究项目的着眼点在于实现人工智能的应用和发展学习系统, 对自然文本进行知识插入等[56].

2 大数据对知识工程的挑战

1977 年, 在第五届国际人工智能会议 (IJCAI 77) 上, 美国斯坦福大学计算机科学家费根堡姆 (Feigenbaum) 首次提出了知识工程 (Knowledge engineering) 的概念. 知识工程的概念提出之后, 人工智能的原理与方法在知识系统领域发挥了重大的作用. 知识工程包括五大活动: 对知识的获取、验证、表示、推论以及对知识的解释. 在知识的基础上, 知识工程通过这五大活动构建专家系统和各种智能系统[57]. 相对于知识管理技术, 知识工程关注的是知识产生和验证过程的动态变化, 它的创新性更强、对数据的操作更加复杂, 并且涉及多个相关领域的知识交叉. 在知识工程的五大活动中, 知识的获取具有更大的难度.

在大数据时代, 利用知识工程的思想和方法, 对大数据进行获取、验证、表示、推论和解释, 通过挖掘出的知识来形成解决问题的专家系统, 是本文所倡导的大知识, 也称为大数据知识工程[39]. 在大数据时代的背景下进行知识工程活动具有诸多挑战. 这主要是由于大数据的本质特征导致的, 涉及到异构、自治的海量多源数据, 隐藏在数据下的知识难以管理和发现. 下面分析一些大数据对知识工程的挑战问题.

首先, 大数据知识工程需要对获取的数据进行合理的存储和表示, 清晰的数据存储形式更有利于发现数据的有用特征, 剔除一些无用的数据属性. 从数据本身来看, 大数据知识工程涉及大量的非结构化数据, 其数据结构多以数据流的形式到来. 数据流数据是一种由实时、连续、有序的数据组成的序列, 它是一种动态变化的数据. 与传统的静态结构化数据相比, 数据流数据具有连续、快速、难以预测数据

趋势等特点[58]. 考虑到大数据特征, 数据的存储要求具有三个变化: 1) 数据量升至 PB 级; 2) 数据分析需求从常规分析转向深度分析 (Deep analytics); 3) 硬件平台从高端转向中低端[59]. 从数据的表示来看, 已有的数据模型包括聚类分析、决策树、分类方法、频繁模式挖掘等. 常见的聚类分析方法是通过寻找数据点的 k 个中心点来获取数据间的距离总和的最小值[60]. 对数据流数据的易变特点产生的概念漂移问题, 已有使用 k 棵随机决策树组成的基分类器的双层窗口的分类算法[61]. 对数据流的频繁模式挖掘, 往往存在实时性较差且查询粒度粗的问题. 而采用快速启发式的方法可以兼顾到对数据流数据的实时处理和更细的查询粒度[62]. 这些模型在提取和刻画数据特征方面各有优劣, 但它们都针对的是静态的数据, 对大数据的表示和数据建模难以适应.

同数据流相对应的是特征流的问题. 含有特征流的应用中, 无法预知整个特征空间的相关知识. 特征流是在时间上连续到来的特征序列, 随着特征数量的不断增加, 训练集的个数可能是固定的[63]、也可能在变化之中[64]. 在线特征的选择具有三大挑战问题: 1) 特征的规模和数量随着时间不断增长; 2) 巨大的特征空间具有未知和规模无限大的可能性; 3) 整个空间的特征过于庞大, 为了学习整个空间的特征, 学习算法无法从最初处理整个特征集. 这三大挑战问题, 同大数据的海量有着密切的关系. 传统的特征选择面对有规律增长的特征数量, 可以不必对特征流加以考虑. 但大数据为特征的选择增加了新的难度, 从而引发了新的研究热点. 针对特征流的问题, 在现有的特征选择算法的基础上, 对特征之间的相关性和特征冗余加以考虑, 能够提高特征选择的效率, 基于特征更为精确和清晰的表示方式[65].

除了大数据的存储和表示方面的挑战, 我们需要考虑的是大数据中知识的获取. 考虑到大数据的多源异构的特征, 数据源通常还含有自治性质, 数据的获取通常是从局部的数据源中获取碎片化的知识[15]. 对观测到的数据, 现有的标准在线学习算法大都使用线性拟合的方式, 多源的数据使得获得的知识往往成碎片化, 碎片化知识的融合无法通过线性拟合完成. 大数据对知识工程的又一挑战是碎片化知识的刻画和融合. 从碎片化知识的获取来说, 现有的拟合方式无法对碎片化数据特征的分布形成合适的拟合, 甚至会产生过度拟合的问题[66]. 其次, 现有的在线学习方法, 尤其是基于 Kernel 算法的在线学习, 随着数据量的上升, 模型的参数设置会变得很复杂. 比如, 使用表示定理 (The representation theorem)[67] 可知 Kernel 函数的数量随着观测值的上升呈现出线性增长, 这样数据分析和处理的复杂度就会提升. 然而, 相应的一个使用机器学习分析大数据的好处是, 许多的数据样本是可获得的, 相应的减小了过度拟合的可能[68].

除此以外, 对数据的训练时间或者使用批处理来处理观测值的时候, 我们对在线学习的响应时间是有要求的, 如果响应时间过长, 那么由于数据隐藏的信息可能会随着时间演化, 则我们得到的信息也许就会对我们在生产、生活、商业决策方面的应用产生误导. 基于处理大规模高维数据的目的, 目前已经提出了多种有效的算法. 大数据环境下的知识发现所需要的算法, 需要避免输入数据时在数值或者特征上的冗余, 否则数据的维度会过高. 同时在学习的过程中不断更新以降低计算的复杂度, 对于高维的数据, 我们还可以使用在线增量学习方法, 实现模型和函数的足够的精确度和近似过程具有足够的泛化[69]. 从碎片化知识的融合来看, 碎片化知识的融合是为了从单个数据源的局部数据中获取整个大数据集合的全局数据特征. 碎片化知识的融合使用现有的线性融合方法会产生一些问题, 例如, 如果我们采用基于形式化逻辑的知识融合[70], 知识融合的过程中会被局部知识的表示形式限制, 对于结构化数据这样的融合方式没有问题, 但在非结构化的数据中, 提取出的碎片化知识不具有统一的数据结构和形式. 大数据环境下, 为了获取数据中的知识, 我们可以采用在线学习的方式. 在线学习面对数据流数据, 对流中可能出现的概念漂移问题能有效地解决[71]. 它不仅仅是把碎片化知识 “拼凑” 在一起, 而是从碎片化知识之间的关联得到新的全局知识, 这和对单数据源的批处理有所区别.

在大数据的知识工程中, 还存在着一些数据可用性的挑战. 我们这里所探讨的大数据的可用性, 包含数据的一致性、完整性、精确性、时效性和实体统一性五个方面[72]. 举例来说, 提高数据的可用性可以增强银行卡的安全性. 如果数据库中存在同一用户的数据主体的不统一, 例如说同一张银行卡的某一段较小的时间间隔内, 发生了两笔空间位置距离较远的消费记录, 则可能存在欺诈消费或者是银行卡被盗刷的可能. 同样, 数据隐私的问题也会影响到数据的可用性. 为了保护个人数据的隐私, 大规模数据集中可能对部分敏感字段采取匿名的方法, 但是这样也使得数据的使用风险增加和巨大的信息损失. 为了在数据的隐私和数据的可用性之间寻找平衡, 研究人员提出了多种方法, 例如轨迹匿名算法[73−74], 通过对用户的轨迹数据的匿名化, 同时融入对时间、位置、速度和方向等外在的轨迹特征信息, 以及对轨迹中邻近位置的改变, 来刻画出轨迹数据之间的相似度.

大数据的知识工程旨在形成对个性化服务有价值和指导作用的专家系统. 从融合的碎片知识, 我们

可以用知识图谱表示大数据中隐藏的大知识. 知识图谱的节点表示碎片化的知识, 连接节点的边我们可以看作是碎片化之间的关联. 我们需要应对的问题是如何量化这些边和节点的关系, 尤其是在动态变化的大数据关系中, 已得到的知识图谱结构也会产生变化. 现有的算法需要从头推算整个数据的结构并更新知识图谱, 这种做法相当耗费时间. 并且, 在海量数据中形成的知识图谱, 由于我们无法对每个观测数据都做到保留, 经过数据处理和清洗的大数据集形成的知识图谱, 必然存在诸如数据值的丢弃、噪声[75]、不平衡数据[76] 等问题. 因此, 大数据的知识工程需要对获取的知识的真实性提出评估机制和演化关系的更新标准.

大数据知识工程还应考虑知识自动化带来的问题. 互联网、大数据、云计算等技术的发展, 虽然带来了更好的数据处理和分析手段, 但许多数据和信息管理应用中仍然存在数据过载的问题. 大数据知识工程最终希望提供以需求为导向的知识服务, 但过载数据的存在降低了服务的可用性和精确性. 知识的自动化指的不是知识本身自动产生, 但可以诱发知识的传播、获取、分析、影响、产生等方面的重要变革[77]. 知识的自动化是信息自动化的自然延伸和提高, 对于具有较大不确定性、冗余性、不一致性的数据和社会信息, 仅依靠人类的智力很难对海量大数据进行更有效分析[78]. 采用以数据作为驱动的方法, 将物理空间产生的数据和虚拟空间产生的数据结合起来进行分析, 将会更有利于解决数据的过载.

3 大数据知识工程模型 — BigKE

基于第 2 节中大数据对知识工程中的各种挑战问题, 本节介绍一种由吴信东等在 2015 年提出的大数据知识工程模型 BigKE[39] (见图 2). 该模型用以解决碎片化知识建模与多数据源的在线学习、碎片知识的非线性融合、需求驱动下的自动化知识导航问题. BigKE 模型采用一种三层次的知识建模方法, 最终获取个性化的知识导航服务. 下面分别从该模型的三个层次来进行介绍.

3.1 多源异构数据中的碎片化知识建模

与传统的知识工程比较, 大数据知识工程着重于提取碎片化知识, 同领域专家知识相结合, 不同于传统的知识工程只基于领域专家的专家知识. 这是因为大数据来源于多源的异构数据, 数据中存在不确定、不完整和异构的问题. 同领域知识相比较, 碎片化知识的精确度有所降低, 但由于它对于有个人偏好的专家知识的依赖度降低了, 换个角度说, 碎片化知识的无偏性和效率也就提高了. 碎片化知识隐藏在多源异构的自治源下, 从这样的数据源中发现知识是一项富有挑战和趣味性的工作. 以社交媒体为例, 2012 年 10 月, 美国总统奥巴马和罗姆尼州长之间的总统辩论在 2 小时内就引发了超过 1 000 万条的推特信息[79]. 如此庞大的信息数量中, 隐藏着复杂的语义关系, 每个用户的评论行为和情感倾向相互独立但又相互影响, 这和大规模数据集的自治性相符合. 随着数据量的不断增大, 数据的来源、数据的结构、数据之间的关联难以使用现有的知识工程技术进行整合. 如何分析用户的行为变化和用户行为之间的相互影响, 成为了一个大数据知识工程问题.

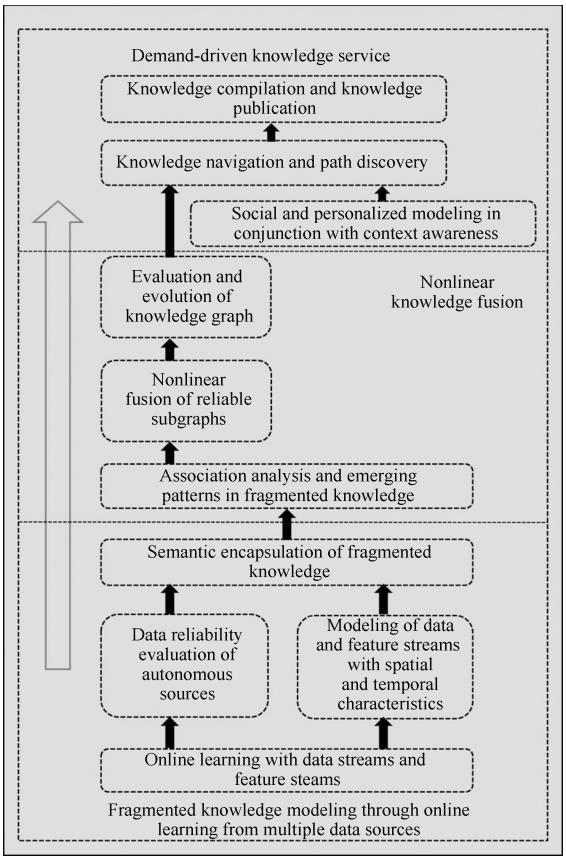


图 2 大数据知识工程模型 — BigKE[39]

Fig. 2 Big data knowledge engineering — BigKE[39]

BigKE 模型的第一步采用合适的模型对多数据源中的碎片化数据进行建模. 从多数据源中获取的碎片化知识对评估数据的可靠性和数据质量有重要的作用. 为了更好地表示数据的特征, BigKE 利用在线学习方法, 同时考虑 “数据流” 和 “特征流”, 因为大数据知识工程首先需要关注的是数据的获取和存储[80]. 在第 2 节中我们讨论了大数据对于知识工程提出了数据存储方面的挑战, 在 BigKE 模型中, 目前可以利用的诸如并行数据库和 MapReduce 技术的混合架构[59]. 对于快速到来的数据, 其中含有

大量的时间和空间信息, 这些时空信息可能隐藏数据的有用特征, 对流数据的处理需要选择动态的模型来刻画数据的特征. 所以同传统的知识工程相比, 动态大数据的知识提取是一个重要方面. 对数据流数据的知识工程已经取得一部分进展, 例如, 针对数据流数据的算法研究和数据模型的改进工作[81], 以及数据流数据的聚类算法研究[82].

对在线获取的碎片知识, 还需要评估数据的可靠性. 这是因为在处理大规模数据的过程中无法对所有的数据进行建模, 采用的数据抽样方式对数据的可信度产生了影响, 同时, 传统的数据学习方法和建模方式无法处理在大数据环境下的概念漂移问题. 数据的精确度和可靠性评估可以通过对数据的来源来进行排序和评价, 在对碎片化数据进行筛选和清洗时, 选择具有较高质量的数据. 关注特征流的在线学习方法和传统的在线学习方法相比, 不再仅仅是关注所处理的数据的处理顺序, 而是对不断增长的大数据的数据量、巨大的数据的特征空间等都有所关注[39], 这样提取出来的碎片化知识具有更高的精确度和可信度. 其次, 碎片化知识建模时, 概念漂移的问题对数据的影响也需要注意. 概念漂移发生时, 现有的数据对象的统计性质可能会随着时间的推进产生变化, 那么我们运用的模型如果是固定不变的, 所得到的碎片化知识的真实性会产生偏差. 我们需要算法和模型具有自适应性, 以得到我们需要的碎片化知识[83]. 概念漂移的情况下, 可能对数据的存储和记忆需要设置时间值, 用以保障对数据特征的存储和描述是最新的.

除了考虑大数据的来源, BigKE 还着眼于数据挖掘和融合的方法来评估数据的质量. 通过改变传统的学习思路, BigKE 在进行大数据的碎片化知识建模时, 采用协同学习 (Co-learning), 这样可以利用具有相似数据特征的数据之间的联系, 从相似的数据中互相评价和调用信息, 以达到提高数据质量的目的, 同时对于碎片化建模的模型质量也会有所提升. 碎片化知识建模的重要性是不言而喻的. 举个例子来说, 如果一个健身的手机 APP 想要为用户提供合适的健身计划, 需要结合这位用户在饮食、运动能力、作息时间甚至是疾病历史等多方面的信息, 涉及到的时间轴和空间轴的刻画是很复杂的, 况且涉及到用户个人隐私的信息, 诸如个人收入, 有时候是难以获得真实完整的数据. 在这样的情况下, 对碎片知识的建模挑战不仅来源于数据模型的挑战, 还涉及到数据的可靠性和完整性等问题.

3.2 从局部知识到全局知识 —— 碎片化知识融合

通过对碎片化知识的建模和语义封装, 我们得到了 BigKE 第一阶段的产物, 即用合适的模型表示的碎片化知识. 为了进一步得到整个大数据集的全局知识, BigKE 需要对碎片化知识进行非线性融合. 多源异构的数据环境下, BigKE 采用知识图谱对碎片化数据进行表示. 将大数据知识工程同传统知识工程相比较, 后者先对收集提取出的知识进行聚合得到全局的知识, 进而在全局知识上进行一系列的知识推断工作, 前者与它的区别在于通过推断工作, 得到现有的局部的碎片化知识中可能没有表现出的有用信息. BigKE 对碎片化知识的融合具有两个创新点: 1) 考虑到碎片化知识的融合无法采用简单的线性处理方式; 2) 将碎片化知识之间的关联表示, 转化成知识图谱的子图来处理.

BigKE 采用知识图谱来表示和融合碎片化知识具有许多优点. 首先, 由于碎片化知识之间的动态的和演化的语义关联, 传统的线性融合方法和模型无法反映出局部知识之间的联系. 碎片化知识之间的关系是复杂的, 其复杂性来源于数据源的异构性, 异构导致了不同的碎片化知识具有不同的记录、存储和表示的形式. 而知识图谱给出了局部知识到全局知识的统一的表示形式, 这使得碎片化知识的融合过程更加简便. 其次, 知识图谱的点与点之间的路径可以看做不同的碎片化知识之间可能的关联, 这为个性化服务的实现提供了实现的可能性. 举个例子来说, 目前的搜索引擎和购物网站可以通过用户的搜索和浏览记录, 推荐给用户相关的新闻网页或者是相关的物品. 诸如亚马逊的相关商品推荐和微博上可能认识的用户的推送信息[84]. 每个用户的记录是局部的, 关于用户的需求的发现由此转变为寻找知识图谱中用户的碎片化知识的相邻节点, 或者是路径导航.

采用知识图谱来进行碎片化知识的非线性融合时, 我们除了要应对复杂的异构数据, 还需要处理好碎片化知识之间固有的语义联系. 例如, 对同一事件的讨论, 从微博、微信、推特等不同的社交网站上获得的碎片化知识可能包含的是同一种意见倾向, 或者是存在敌对的意见, 那么进行知识融合时我们需要在知识图谱中有所体现. 通过知识图谱表示的知识的节点和所连接的尚在演化的关系中需要作出相应的调整. 因此, 我们需要关注的是, 碎片化知识融合时这些联系和节点的表示[85]. BigKE 模型中对碎片知识的融合, 需要对现有的子图进行一定的筛选, 碎片化知识反映出的局部信息是多数据源的自治性的一种表现, 这些局部的信息对获取全局知识的重要性, 也需要通过子图的可信度来刻画.

与现有的推荐网站和个性化服务有所不同, 大数据知识工程模型 BigKE 的知识图谱结构需要动态更新, 这是碎片化知识之间复杂的动态联系所导致的. BigKE 对碎片化知识的融合过程同样引入了

评估机制. BigKE 模型的第 1 层中, 需要评估的是所获得数据的质量, 在知识融合的过程中, 评估的是知识图谱的可靠性. 这是由于碎片化知识的复杂关系同样受到漂移的影响, 由此带来了关系的演化. 评估这些碎片化知识之间的关联可以提升所得到的知识图谱的精确度, 对后期知识导航奠定基础, 评估的标准可以参考碎片化知识联系的关系强度等来表述. 举例来说, 在不同的关系中, 关系强度可以被描述为显式的强度或隐式的强度关系[86]. 碎片化数据之间的关联被表示为知识图谱的边, 通过对数据间关联的强度刻画, 能够动态更新知识图谱的边, 从而刻画出大数据中动态的数据关联.

3.3 个性化知识导航

大数据知识工程的最终目标是提供以用户需求为导向的知识服务. BigKE 模型通过对碎片化知识的非线性融合得到了大数据的全局知识, 为了将从大数据中获取的知识应用到知识服务中, 需要考虑用户的社交信息等个性化的信息, 并需要合适的方法对用户的个性化查询提供精确的推荐和导航服务. 前两小节中提到的知识图谱, 其节点和边对应的是知识的单元和知识之间的语义关联, 提供知识服务可以看作寻找某两点之间的最佳路径. 我们利用用户需求作为导向, 使用知识图谱中的连接关系, 寻找用户节点包含内容之间的关系.

个性化服务的一个案例是病人之间的同病不同源, 因而在 4P 模型里需要不同的治疗方案。在数据挖掘技术快速发展的时代, 我们应该站到数据科学发展的最前沿, 积极探索将全新的数据分析技术和个性化服务相融合的方法. 类似地, 已经成功应用到网络学习等系统中的知识导航服务, 是根据对文本语境和浏览记录等数据的分析, 寻找到最感兴趣的另一个知识节点[87]. 以社交网络为例, 我们可以从客户端的浏览器中获取细节信息, 就是用户选择浏览的链接列表及其运行的时间[88]. 其他可获取的一些信息的类型, 还包括用户从某一链接到另一网站的路径消耗时间. 这些信息可以用来形成用户的个性化查询, 从而我们可以找到最短路径的链接来预估用户需求从而做出合适的推荐和导航服务. 发现路径后, 对从知识图谱中提取出的路径结构还要加以适当的整理工作.

对用户的个性化需求和查询要求, BigKE 模型在知识图谱上直接进行推断工作, 从而进行用户未来可能行为的预测. 这也是大数据知识工程和传统知识工程的一个区别. 前者注重的是预测未来, 后者注重的是管理和使用已获取的数据和知识. 但由于庞大的数据量和知识图谱规模的巨大化, 在数据维度过高时会产生问题, 使得新知识图谱的构建和导航服务的质量下降, 因而, 发现算法采取近似手段是必要的. 同时, 对所提供知识导航服务的用户, 其潜在需求往往需要结合到上下文感知、协同过滤等技术[89]. 开发和应用知识导航算法时, 上下文感知技术向我们提供调整知识系统运行的可能, 尤其是涉及到移动设备等的知识导航服务, 上下文感知技术能够大大提高所获取的知识的可用性, 提升知识导航服务的价值. 协同过滤技术同样是为了降低知识导航的模糊性, 提升个性化服务的准确度, 基于用户的系统通过对推荐和导航服务预测工作的评价, 可以获得更优良的精度评价指标.

对用户的需求和个性化查询, BigKE 基于知识图谱给出的结果, 还需要用一种直观、简便的形式展现给用户, 以提高知识服务的可用性和可操作性. 例如, 用户总是希望手机推荐的热点新闻是以简洁的标题和某一张新闻图片结合的方式呈现的, 如果推荐系统只是将推荐的内容以长文本的形式推送给用户, 那么就会降低用户的阅读兴趣, 从而使获取的大知识被用户忽略.

4 大知识的挑战和前景

与 5V 模型、5R 模型、4P 医学模型和 HACE 相比较, BigKE 具有它的优越性. 面向海量多源的动态数据, BigKE 考虑到大数据的异构和自治特征, 提供基于互联网的知识服务. 5V 模型、5R 模型和 4P 医学模型提炼出的大数据特征, 在大数据知识工程中为大数据中的 “大” 知识的存储和分析工作提供了导向, 但它们没有强调大数据中数据流和特征流的处理方式. 对数据流数据的碎片化知识提取和非线性融合可以依靠 BigKE 的第 1 层和第 2 层得到. 4P 医学模型强调用户个人信息的参与, 这需要对基于互联网的大数据流之间的语义关系建立合适的模型. BigKE 对碎片化知识的语义封装能够提供更可靠的个人信息及它们之间的演化关系的表示, 体现出大数据动态的特点. HACE 定理给出了处理大数据的多层框架, BigKE 在它的基础上对大数据挖掘形成的知识图谱提出了个性化服务的导航, 更有利于和具体的应用实例结合. 尽管 BigKE 同已有的大数据模型相比具有自身的优势, 但涉及到大知识的发现和挖掘, 仍具有进一步的挑战.

大数据知识工程模型 BigKE 旨在解决大数据对知识工程提出的挑战, 本节我们讨论 BigKE 中几个挑战问题和可能的应用场景.

挑战 1. 碎片化知识的非线性融合. 首先, 在 BigKE 的第 2 层, 碎片化知识生成于异构自治的多源数据. 这些数据没有统一的数据表示形式, 这些碎片化知识也缺乏统一的逻辑结构, 所以知识融合起来十分困难. 传统的知识工程处理的信息通常含有

一定的逻辑和统一的格式, 而 BigKE 面对多种形式的数据, 诸如微博、短信息、传感器数据、音视频和邮件等, 这项挑战工作也正在形成一个研究热点. 现有的数据融合方法大多采用的是有偏估计, 例如, 利用多传感器的有偏估计, 可以将数据的融合近似的收敛于无偏的估计, 从一定程度上提高数据融合的精确度[90]. 将异构的碎片化知识进行融合时, 为了形成统一的知识图谱形式, 我们无法兼顾到所有的信息, 因此必然存在对数据和信息的取舍问题, 如果单纯采用加权和阈值的形式决定融合过程中对信息的丢弃, 则融合后的全局知识的精度会下降. 因此, 我们需要一个合适的机制来选择在碎片化知识融合的过程中, 对数据信息的取舍做出判断, 期望在尽可能保留原有信息以提高知识图谱的准确度, 同时也能够以一种简便的形式表现出用户需要的知识.

挑战 2. 大知识图谱的动态更新. 大数据知识工程与传统知识工程的一大区别在于大数据知识工程具有预测未来趋势的要求. 大数据不断地到来, 现有的知识图谱无法一劳永逸地表现出每时每刻的数据特征. 大数据的数量可能呈现惊人的增长速度, 现有数据之间的关联随着时间的推移也会产生变化. 碎片化知识的关联随着原始数据关联的变化而变化, 表现在知识图谱中可能是某个节点的消失和新节点的产生, 以及一些新产生的边的构建. 知识图谱的动态更新主要涉及到两大问题: 1) 如何设置合理的时间点更新现有知识图谱, 2) 如何确定对某一数据关联的取舍问题. 第 1 个问题可采用事先设置好的时间阈值, 以当前时间点为起始, 到达规定的阈值范围时, 则重新扫描数据集构建新的知识结构. 这样的方式虽然可以提高所得到的知识的质量, 但是大规模数据集的重新扫描过于耗费时间, 不满足对大数据知识工程的时间要求. 因此, 相比较于采用事先设置的固定时间阈值, BigKE 的后续工作可以考虑对时间阈值的动态设置. 设置扫描时间阈值的动态指标可以参考新的数据到来的速度, 根据新数据产生的多少来调整更新算法运行的时间间隔. 针对第 2 个问题, BigKE 的挑战在于要建立一个数据关联度的评估评价机制, 因为现有的数据关联, 无论是数据节点还是联系, 都会随着新数据的到来和时间的推移发生变化. 在进行知识图谱更新时, 为了确定一条现有的边的保留或者丢弃, 现有的数据关联强度算法很少考虑到大数据的动态性, 接下来的工作需要考虑对数据关联强度的评价机制中加入动态的因素.

挑战 3. 基于集成和拆解的知识重组. BigKE 的核心思想是集成碎片化数据, 产生新的知识面向个性化服务. 然而, 碎片有大有小, 有些大碎片必须首先分割成小的碎片以后才能有效集成. 这就是粒度问题. 人们一般不认为一本完整的书是知识碎片. 然而, 在浩如烟海的书库前面, 一本书就可以看成是一个知识碎片. 一篇文章可能会被看成是知识碎片. 然而, 如果分开考察它所包含的许多定理, 以及这些定理所组成的知识体系, 那么文章本身又不是碎片了. 因此. 是碎片还是知识, 是相对而非绝对的. 如何分拆, 如何重组? 如何根据重组的目标来分拆? 既是技术问题, 也是科学问题.

挑战 4. 海量碎片化知识的约化表示. 海量并不能完全刻画大数据, 但是大数据一定是海量的, 而且大数据存在着不确定、不完整、含噪音的数据质量问题。我们不能在要用到大数据时每次都临时到网上去找, 所以必须考虑大数据和从大数据中生成的碎片化知识的海量存储和管理问题. 在许多的相关技术中, 大数据及其碎片化知识的存储、访问和利用可以采取约化表示. 约化的含义是把同一知识的复杂表示 A 转换为简单表示 B, 使得 B 的容量大大小于 A, 但是 B 已经包含了 A 的绝大部分有用信息, 已经可以在绝大部分场合代替 A “出场”. 一个实例是机器学习中的流形学习, 它的主要作用是降维, 把高维数据降为低维数据而不影响, 或很少影响其特征性质. 该方法在各种模式识别中有重要应用.

挑战 5. BigKE 的分布式实现. 高效的大数据知识工程一定要走分布式处理的道路, 不仅是为了存储和管理, 更重要的是为了计算效率. 我们在前文中提到了一种可能的选择是采用 Map-Reduce 方法. 该方法的核心在于把大数据分拆成许多小块数据, 分配到许多节点上, 通过分布式方式计算后再集成其结果. 但这个方法也不是万能的. 对解决某些问题来说, 例如统计问题, 其结果可能会不理想. 除了前文已经提到的把大量分散模块的数据合并计算可能会模糊了某些统计阈值以外, 还可能出现统计值不正确的问题, 徐宗本院士指出, Hadoop 类型的大数据回归算法, 只有在满足所谓 “一致相合” 条件下才能提供合理结果[91]. 这样的挑战是我们在把大数据集成为知识时必须应对的.

挑战 6. 个性化用户行为的建模. 大数据中的大知识为我们提供了个性化的大知识服务, 个性化大知识服务的关键在于对个人和社交信息的建模. 由于 BigKE 提出大数据的知识工程需要直接在知识图谱上进行知识的推断, 那么接下来的工作重点应该着眼于过滤和选择算法的实时性. 在知识图谱上的直接推断可能会产生几个相类似的结果, 除了知识图谱的结构在随着时间变化, 用户的需求也会产生变化, 所以, BigKE 模型面对的另一大挑战问题是对用户行为的建模. 通过聚集个人和社交的信息, 知识图谱可望涵盖用户的行为和情感倾向, 由此 BigKE 可以对用户未来的行为做出推断, 从而动态地改善现有的知识服务质量. 从协同过滤或上下文

感知的过滤和选择机制开始, 加入用户行为的推断, 这样给出的结果带有实时性, 但同时对 BigKE 的挑战又进一步提升了, 因为多一个考虑的维度, 带有需求驱动的大数据算法的编译效率可能就会下降很多, 训练集和测试集的划分也会对算法的效率有所影响, 因此 BigKE 的后续工作还涉及到大数据算法效率的提升.

大知识面向国民经济的主战场, 在各个科技领域都会有着广泛的应用. 下面我们分析几个大知识的应用场景.

应用场景 1. 动态网络大词典. 本文在第 1 节中已经对大知识给出定义. 大知识所具有的海量、异构和多源的特性源于大数据的来源. 将大知识应用到动态词典的建立和更新中具有广阔的前景. 动态词典是相对于传统的静态数据而言, 词典的建立和更新是动态的, 其动态性体现在随着社会和网络语言知识的变化, 在较短的时间间隔内动态词典能够更新词汇的内容和语言的规范. 从文本语言中抓取即时的语料库, 实现动态词典的动态特征. 事实上, 无论是文本数据挖掘还是动态词典的建立, 都需要对语料库加以动态的扩充和更新来不断适应伴随数据流和特征流到来的新数据. 除了对语料库的动态更新, 网络动态词典所应用的大知识还能体现词汇的关联和兼容. 这是由于大知识来源于异构的大数据, 从多种媒体抓取的词语信息, 需要经过加工和融合形成新的词语信息对语料库进行更新. 异构的多源信息是否能产生新的大知识, 取决于对新知识的评估体系, 评估内容应当包含新知识与当前已有词汇信息的重合度比较和关联性分析, 以降低动态词典内知识的重合和冗余. 大知识应用在动态词典的建立和更新中, 除了有上述的两个关键问题, 考虑动态词典的内容, 还应当包含有方言的相关知识. 大知识的多源特征决定了它应当涵盖尽可能多和广的信息, 应用在动态词典中, 表现为词汇的覆盖范围需要考虑到时间和空间两个因素. 时间维度上表现为词典的动态更新, 空间维度上表现为词典的内容考虑到地域的不同, 则应当涵盖尽可能多的方言知识.

应用场景 2. 网络新闻的动态跟踪和总结. 大知识应用到多源新闻分析领域, 具有新的应用前景, 可以做新闻的动态跟踪和总结. 在互联网 2.0 时代, 可供获取的新闻信息增长过快, 然而新闻的数量快速增长的同时, 并没有使得新闻的质量同步提升, 重复阅读的信息耗费了用户大量的时间. 新闻事件中的大知识, 应当伴随时间轴清晰地梳理和表示出新闻事件的多个主题, 包括对频繁发生的新闻事件的当前关注焦点和后期演变形式的跟踪, 以便用户更加全面和具有针对性地获取新闻中重要的本质. 新闻的动态跟踪和总结基于大量的新闻网页和文本, 利用词共现图的构建提取出用户感兴趣的新闻中的多个主题, 对与新闻事件相关的多个主题建立各自的摘要集合, 从而生成各主题的动态跟踪和总结. 在整个新闻主题的抓取和动态跟踪过程中, 产生了大量的知识. 在这个应用背景下, 大知识表现为与用户感兴趣的新闻最具相关性的新闻主题和摘要总结. 动态的新闻跟踪在考虑新闻查询和新闻相关性的基础上, 考虑新闻文档中的多个主题, 针对同一个新闻事件, 建立了更加清晰的主题演化过程的展示和更加全面的新闻内容的总结.

应用场景 3. 普适医疗信息的管理与服务. 在医疗应用方面, 大知识的应用具有广阔的前景. 大知识与普适医疗的结合, 可以建立和动态更新医疗推荐系统. 通过分析用户的个人信息, 包括地理位置、个人病史和社交偏好等, 实时更新用户附近的医院、药房等医疗保障系统的信息. 在某一时刻, 用户根据需要查询当前针对某一病症可获得的最佳诊断和治疗方案. 查询信息表现为现有的病症表现和疼痛程度等, 个性化推荐信息可以包括距离最近和治疗效果最佳的药房和医院等信息. 这一过程需要大知识作为普适医疗系统的支撑. 医疗数据中在地理、多种类医疗器械和软件上的分布, 由此导致的异构性造成了信息集成的困难. 同时, 利用收集到的医疗数据挖掘出有价值的医疗知识成为了能否提供准确的推荐信息的关键. 病人的病史分析和现有医疗知识图谱的比对是否精准, 也需要通过用户的评价系统不断加以改进.

应用场景 4. 万维网就业培训. 与普适医疗类似, 个性化的推荐服务中大知识还可以渗透到网上创业培训当中. 基于万维网的就业信息, 可以构建大型的知识图谱, 其子图的划分可以参考就业的种类选择、求职人的文化水平以及地域划分等. 就业技能的数据包含多个职业分类, 数据的来源也各不相同, 含有地域性的差异, 由此导致了数据的集成和融合问题. 比如, 在农业发达地区, 对种植指导专家岗位的需求远远大于渔业和工商业发达地区. 那么网上就业培训系统需要依据用户的地理信息进行数据的筛选和过滤, 结合用户的个人就业倾向和现有的岗位的地理位置, 进行就业培训内容的推荐. 事实上, 个人通过网上就业培训系统学习就业技能时, 系统根据用户所提出的限定条件, 反馈出的信息是从已有的大知识图谱中寻找针对某一问题的映射, 为用户提供市场分析和技能培训.

例如, 某个本科即将毕业的计算机专业的学生希望策划一份上海的软件开发工作, 该学生已具备的知识可能有高等数学和数据结构等基础知识, 但某一符合他就业期望的岗位还需要具备高级编程语言的技能. 通过将大知识图谱中的某一针对性映射

同用户个人的知识图谱进行比对, 可以发现相似的节点以及缺失的节点, 从而寻找到用户就业需要学习的技能, 提高知识学习导航的准确性. 网上就业培训的关键在于个人图谱和大知识图谱的比对以及大知识图谱的构建, 这些关键问题随着大知识应用范围的扩大会成为进一步的挑战.

应用场景 5. 自动编辑和出版. 上面提到的挑战 3, 如果能够很好地解决, 则自动知识编辑的前景就可以实现. 例如: 要求计算机根据库中的一万本计算机科学电子书, 自动编辑下列新书: 计算机科学百科全书、计算机软件教程、大数据发展史要、计算机专业大学生用操作系统习题集等. 从长远来看, 只要我们有一个数量巨大、组织合理、不断更新的 "知识碎片库", 那么编辑和出版新书以满足各种社会需求就不再是一个大量耗费人力和财力的事业.

应用场景 6. 智慧城市的动态认知与决策. 面向智慧城市及城市重大事件管理的实际需求, 大数据知识工程可以针对城市大数据在自然属性、地理属性、时间属性、社会属性以及交互行为等方面的异构、自治、多介、高维、低质等特点, 发现伴随时空维度推进下蕴含的内在关联语义一致性, 实现复杂关系的动态认知和演化计算, 探索多源感知信息的多层次关联、语义提取与融合分析的机制和方法, 实现多源异构城市数据的紧耦合. 智慧城市的动态认知可以进行跨时空城市感知数据的关联推理和深度挖掘, 研究多维 (时间、空间、属性、语义) 数据分析的城市重大事件管理方法, 包括同类、异类城市事件的相关性分析、以及预测未来一段时间内同地区发生类似事件的可能性, 对城市群体行为或个别重大事件数据进行理解与分析, 建立城市行为动力学理论体系. 智慧城市的动态决策可以通过城市重大事件的交互式临场分析, 实现协同感知下城市大数据的推理模型, 研究城市行为事件间相互作用、渗透和扩散的物理模型, 以揭示城市行为涌现、传播和演化机制, 对面向公共安全的敏感事件进行语义理解、检测跟踪和预测预警. 智慧城市的动态认知与决策基于数据和知识的联合驱动以及多模态数据的关联增强技术, 采用多源信息的视觉转换机制和自适应交互可视化方法, 旨在推进人机智能的深度耦合, 实现 "数据 – 信息 – 知识" 阶进式服务, 从而实现大数据时代的大知识精细化城市模拟及管理.

5 总结

从大数据中获取有价值的大知识具有许多问题和挑战, 这与大数据的本质特征密不可分. 由大数据的 HACE 定理, 我们了解到大数据异构和自治的本质特征, 其分布式和分散式控制的特点, 以及大数据之间复杂和演化的关联. 大数据的本质特征使得知识工程存在诸多挑战, 利用传统的数据分析和处理手段无法解决这些问题. 现有的几种大数据模型, 包括 5V, 5R, 4P 和 HACE 定理, 在解决大数据知识工程的问题和挑战中具有各自的优劣. 本文从知识建模的角度介绍一种大数据知识工程模型 BigKE, BigKE 模型从大数据源中提取碎片化知识, 进而对这些碎片化知识进行非线性的知识融合, 最终根据用户的需求作为导向, 提供个性化的大知识服务. 将 BigKE 模型和一些现有的大数据模型相比较, BigKE 通过其三层架构给出了现有大数据模型提出的一些挑战问题的解决思路. 为了从大数据中获取更高质量的大知识, BigKE 模型还有许多有价值的后续工作, 主要针对 BigKE 中无法解决的挑战给出后续工作的方向. 大知识的进一步挑战与其广泛的应用前景密切相关, 在动态网络大词典的构建、新闻的动态跟踪和总结、普适医疗、网上就业培训、自动编辑和出版、以及智慧城市的动态认知和决策等应用场景中, 大知识还大有可为.

致谢

HACE 定理和 BigKE 模型是文献 [15, 39] 的合作者们共同研究的成果, 本文的讨论和展望也得益于同这些作者和其他大数据知识工程方向合作者的广泛交流, 这里对国内合肥工业大学、西安交通大学、中国科学院数学所、中国科学技术大学、华东师范大学、广西师范大学、百度和国外众多单位的同行和合作研究者们一并表示感谢.

References

1 Beyer M A, Laney D. The importance of "Big Data": a definition [Online], available: https://www.gartner.com/doc/2057415, February 17, 2016

2 Marr B. Big data: the 5 Vs everyone must know [Online], http://www.linkedin.com/pulse/20140306073407-64875646-big-data-the-5-vs-everyone-must-know, January 21, 2016

3 Mervis J. Agencies rally to tackle big data. *Science*, 2013, **336**(6077): 22−22

4 Wang Fei-Yue. Software-deined systems and knowledge automation: a parallel paradigm shift from Newton to Merton. *Acta Automatica Sinica*, 2015, **42**(1): 1−8
(王飞跃. 软件定义的系统与知识自动化: 从牛顿到默顿的平行升华. 自动化学报, 2015, **42**(1): 1−8)

5 Fish A N. *Knowledge Automation: How to Implement Decision Management in Business Processes*. USA: Wiley, 2012.

6 Fernández A, Del Río S, López V, Bawakid A, Del Jesus M J, Benítez J M, Herrera F. Big data with cloud computing: an insight on the computing environment, MapReduce, and programming frameworks. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 2014, **4**(5): 380−409

7 Kent S M. Sloan digital sky survey. *Science with Astronomical Near-Infrared Sky Surveys*. France: Springer, 1994. 27−30

8 Labrinidis A, Jagadish H V. Challenges and opportunities with big data. *Proceedings of the VLDB Endowment*, 2012, **5**(12): 2032−2033

9 Knoll A, Meinkoehn J. Data fusion using large multi-agent networks: an analysis of network structure and performance. In: Proceedings of the 1994 IEEE International Conference on MFI′94, Multisensor Fusion and Integration for Intelligent Systems (MFI). Las Vegas, NV: IEEE, 1994. 113−120

10 Nature Editorial. Community cleverness required. *Nature*, 2008, **455**(7209): 1−1

11 Che D R, Safran M, Peng Z Y. From big data to big data mining: challenges, issues, and opportunities. In: Proceedings of the 18th International Conference on Database Systems for Advanced Applications. Wuhan, China: Springer, 2013. 1−15

12 Stidston M. Business leaders need R′s not V′s: the 5 R′s of big data [Online], available: https://www.mapr.com/blog/business-leaders-need-r%E2%80%99s-not-v%E2%80%99s-5-r%E2%80%99s-big-data#.U2qmcq1dWIU, December 21, 2015

13 Wang Ji, Wang Qi. Chinese constitution research and the practice of 4P medical model. *Chinese Journal of Integrated Traditional and Western Medicine*, 2012, **32**(5): 693−695
(王济, 王琦. 中医体质研究与 4P 医学的实施. 中国中西医结合杂志, 2012, **32**(5): 693−695)

14 Auffray C, Charron D, Hood L. Predictive, preventive, personalized and participatory medicine: back to the future. *Genome Medicine*, 2010, **2**(8): 57−57

15 Wu X D, Zhu X Q, Wu G Q, Ding W. Data mining with big data. *IEEE Transactions on Knowledge and Data Engineering*, 2014, **26**(1): 97−107

16 Wikipedia. Big data [Online], available: https://en.wikipedia.org/wiki/Big_data#Definition, December 12, 2015

17 IDC 权威定义大数据概念: 满足 4V 标准 [Online], available: http://www.d1net.com/bigdata/news/237143.html, December 12, 2015

18 Tien J M. Big data: unleashing information. *Journal of Systems Science and Systems Engineering*, 2013, **22**(2): 127−151

19 Wang Yuan-Zhuo, Jin Xiao-Long, Cheng Xue-Qi. Network big data: present and future. *Chinese Journal of Computers*, 2013, **36**(6): 1125−1138
(王元卓, 靳小龙, 程学旗. 网络大数据: 现状与展望. 计算机学报, 2013, **36**(6): 1125−1138)

20 Wang Wei-Wei, Li Xiao-Ping, Feng Xiang-Chu, Wang Si-Qi. A survey on sparse subspace clustering. *Acta Automatica Sinica*, 2015, **41**(8): 1373−1384
(王卫卫, 李小平, 冯象初, 王斯琪. 稀疏子空间聚类综述. 自动化学报, 2015, **41**(8): 1373−1384)

21 Armbrust M, Fox A, Griffith R, Joseph A D, Katz R H, Konwinski A, Lee G, Patterson D A, Rabkin A, Stoica I, Zaharia M. Above the Clouds: A Berkeley View of Cloud Computing, Technical Report UCB/EECS-2009-28, EECS Department, University of California, Berkeley, 2009

22 Blaabjerg F, Teodorescu R, Liserre M, Timbus A V. Overview of control and grid synchronization for distributed power generation systems. *IEEE Transactions on Industrial Electronics*, 2006, **53**(5): 1398−1409

23 Leskovec J, Huttenlocher D, Kleinberg J. Signed networks in social media. In: Proceedings of the 2010 SIGCHI Conference on Human Factors in Computing Systems. New York: ACM, 2010. 1361−1370

24 Zikopoulos P, Eaton C. *Understanding Big Data: Analytics for Enterprise Class Hadoop and Streaming Data*. USA: McGraw-Hill Osborne Media, 2011.

25 The four V′s of big data [Online], available: http://www.ibmbigdatahub.com/sites/default/files/infographic_file/4-Vs-of-big-data.jpg, January 21, 2016

26 Lazer D, Kennedy R, King G, Vespignan A. The parable of google flu: traps in big data analysis. *Science*, 2014, **343**(6176): 1203−1205

27 IBM. What is big data? [Online], available: http://www-01.ibm.com/software/data/bigdata/what-is-big-data.html, December 2, 2015

28 Barwick H. The "four Vs" of big data. Implementing information infrastructure symposium [Online], available: http://www.computerworld.com.au/article/396198/iiis_four_vs_big_data, December 2, 2015

29 数据并非越大越好: 谷歌流感趋势错在哪儿了? [Online], available: http://www.guokr.com/article/438117/, December 2, 2015

30 Ghemawat S, Gobioff H, Leung S T. The Google file system. In: Proceedings of the 19th ACM Symposium on Operating Systems Principles. New York: ACM, 2003. 29−43

31 Dean J, Ghemawat S. MapReduce: simplified data processing on large clusters. In: Proceedings of the 6th Symposium on Operating Systems Design and Implementation. Berkeley, CA, USA: USENIX Association, 2004. 137−149

32 Big data solution offering [Online], available: http://mike2.openmethodology. org /wike /Big_ Data_ Solution_ Offering, November 28, 2015

33 White T. *Hadoop: The Definitive Guide* (2nd Edition). USA: Yahoo Press, 2010. 1−4

34 Gupta P, Kumar P, Gopal G. Sentiment analysis on Hadoop with Hadoop streaming. *International Journal of Computer Applications*, 2015, **121**(11): 4−8

35 Liao S H. Expert system methodologies and applications — a decade review from 1995 to 2004. *Expert Systems with Applications*, 2005, **28**(1): 93−103

36 Wu Xin-Dong, Ye Ming-Quan, Hu Dong-Hui, Wu Gong-Qing, Hu Xue-Gang, Wang Hao. Pervasive medical information management and services: key techniques and challenges. *Chinese Journal of Computers*, 2012, **35**(5): 827−845
(吴信东, 叶明全, 胡东辉, 吴共庆, 胡学钢, 王浩. 普适医疗信息管理与服务的关键技术与挑战. 计算机学报, 2012, **35**(5): 827−845)

37 Auffray C, Chen Z, Hood L. Systems medicine: the future of medical genomics and healthcare. *Genome Medicine*, 2009, **1**(1): 2−2

38 Luo Xu, Chen Bo, Luo Li-Ya, Zhang Hong-Yan, Wu Hao, Li Jing-Bo. Discussion on reconstructing hospital healthcare management under 4P medical conception. *Chinese Hospitals*, 2014, **18**(7): 61−63
(罗旭, 陈博, 罗莉娅, 张宏雁, 吴昊, 李景波. 4P 医学理念下医院健康管理体系重构思考. 中国医院, 2014, **18**(7): 61−63)

39 Wu X D, Chen H H, Wu G Q, Liu J, Zheng Q H, He X F, Zhou A Y, Zhao Z Q, Wei B F, Li Y, Zhang Q P, Zhang S C, Lu R Q, Zheng N N. Knowledge engineering with big data. *IEEE Intelligent Systems*, 2015, **30**(5): 46−55

40 Klasnja P, Pratt W. Healthcare in the pocket: mapping the space of mobile-phone health interventions. *Journal of Biomedical Informatics*, 2012, **45**(1): 184−198

41 Vassis D, Belsis P, Skourlas C, Pantziou G. Providing advanced remote medical treatment services through pervasive environments. *Personal and Ubiquitous Computing*, 2010, **14**(6): 563−573

42 合肥工业大学吴信东: 大数据 Processing Framework 多层架构 [Online], available: http://www.csdn.net/article/2012-07-27/2825305, December 7, 2015

43 Petersen W P, Arbenz P. *Introduction to Parallel Computing*. Oxford: Oxford University Press, 2004.

44 Corbett J C, Dean J, Epstein M, Fikes A, Frost C, Furman J J, Ghemawat S, Gubarev A, Heiser C, Hochschild P, Hsieh W, Kanthak S, Kogan E, Li H Y, Lloyd A, Melnik S, Mwaura D, Nagle D, Quinlan S, Rao R, Rolig L, Saito Y, Szymaniak M, Taylor C, Wang R, Woodford D. Spanner: Google′s globally-distributed database. *ACM Transactions on Computer Systems*, 2012, **31**(3): Article No. 8

45 Chang F, Dean J, Ghemawat S, Hsieh W C, Wallach D A, Burrows M, Chandra T, Fikes A, Gruber R E. BigTable: a distributed storage system for structured data. *ACM Transactions on Computer Systems*, 2008, **26**(2): Article No. 4

46 Peel M, Rowley J. Information sharing practice in multi-agency working. *ASLIB Proceedings*, 2010, **62**(1): 11−28

47 Wang M D, Li B, Zhao Y X, Pu G G. Formalizing Google file system. In: Proceedings of the 20th IEEE Pacific Rim International Symposium on Dependable Computing (PRDC). Singapore: IEEE, 2014. 190−191

48 Cormode G, Srivastava D. Anonymized data: generation, models, usage. In: Proceedings of the 2009 ACM SIGMOD International Conference on Management of Data. Providence, RI: ACM, 2009. 1015−1018

49 Sweeney L. *k*-anonymity: a model for protecting privacy. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, 2002, **10**(5): 557−570

50 Kopanas I, Avouris N M, Daskalaki S. The role of domain knowledge in a large scale data mining project. *Methods and Applications of Artificial Intelligence*. Thessaloniki, Greece: Springer, 2002. 288−299

51 Salton G M, Wong A, Yang C S. A vector space model for automatic indexing. *Communications of the ACM*, 1975, **18**(11): 613−620

52 Deerwester S C, Dumais S T, Furnas G W, Landauer T K, Harshman R. Indexing by latent semantic analysis. *Journal of the American Society for Information Science*, 1990, **41**(6): 391−407

53 Freedman E G, Shah P. Toward a model of knowledge-based graph comprehension. *Diagrammatic Representation and Inference*. Callaway Gardens, GA, USA: Springer, 2002. 18−30

54 Aral S, Walker D. Identifying influential and susceptible members of social networks. *Science*, 2012, **337**(6092): 337−341

55 Centola D. The spread of behavior in an online social network experiment. *Science*, 2010, **329**(5996): 1194−1197

56 Strassel S, Adams D, Goldberg H, Herr J, Keesing R, Oblinger D, Simpson H, Schrag R, Wright J. The DARPA machine reading program — encouraging linguistic and reasoning research with a series of reading tasks. In: Proceedings of the 7th International Conference on Language Resources and Evaluation. Valletta, Malta: European Language Resources Association, 2010. 986−993

57 Studer R, Benjamins V R, Fensel D. Knowledge engineering: principles and methods. *Data and Knowledge Engineering*, 1998, **25**(1−2): 161−197

58 Pan Yun-He, Wang Jin-Long, Xu Cong-Fu. State-of-the-art on frequent pattern mining in data streams. *Acta Automatica Sinica*, 2006, **32**(4): 594−602
(潘云鹤, 王金龙, 徐从富. 数据流频繁模式挖掘研究进展. 自动化学报, 2006, **32**(4): 594−602)

59 Wang Shan, Wang Hui-Ju, Qin Xiong-Pai, Zhou Xuan. Architecting big data: challenges, studies and forecasts. *Chinese Journal of Computers*, 2011, **34**(10): 1741−1752
(王珊, 王会举, 覃雄派, 周xx亘. 架构大数据: 挑战、现状与展望. 计算机学报, 2011, **34**(10): 1741−1752)

60 Guha S, Mishra N, Motwani R, O′Callaghan L. Clustering data streams. In: Proceedings of the 41st Annual Symposium on Foundations of Computer Science. Redono Beach, USA: IEEE, 2000. 359−366

61 Zhu Qun, Zhang Yu-Hong, Hu Xue-Gang, Li Pei-Pei. A double-window-based classification algorithm for concept drifting data streams. *Acta Automatica Sinica*, 2011, **37**(9): 1077−1084
(朱群, 张玉红, 胡学钢, 李培培. 一种基于双层窗口的概念漂移数据流分类算法. 自动化学报, 2011, **37**(9): 1077−1084)

62 Zhang Xin, Li Xiao-Guang, Wang Da-Ling, Yu Ge. A high-speed heuristic algorithm for mining frequent patterns in data stream. *Journal of Software*, 2005, **16**(12): 2099−2105
(张昕, 李晓光, 王大玲, 于戈. 数据流中一种快速启发式频繁模式挖掘方法. 软件学报, 2005, **16**(12): 2099−2105)

63 Wu X D, Yu K, Ding W, Wang H, Zhu X Q. Online feature selection with streaming features. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2013, **35**(5): 1178−1192

64 Zhang Q, Zhang P, Long G D, Ding W, Zhang C Q, Wu X D. Towards mining trapezoidal data streams. In: Proceedings of the 2015 IEEE International Conference on Data Mining (ICDM′15). Atlantic City, NJ, USA: IEEE, 2015. 1111−1116

65 Wu X D, Yu K, Wang H, Ding W. Online streaming feature selection. In: Proceedings of the 27th International Conference on Machine Learning. Haifa, Israel, 2010. 1159−1166

66 Kivinen J, Smola A J, Williamson R C. Online learning with kernels. *IEEE Transactions on Signal Processing*, 2004, **52**(8): 2165−2176

67 Kimeldorf G, Wahba G. Some results on Tchebycheffian spline functions. *Journal of Mathematical Analysis and Applications*, 1971, **33**(1): 82−95

68 Zhou Z H, Chawla N V, Jin Y C, Williams G J. Big data opportunities and challenges: discussions from data analytics perspectives [Discussion forum]. *IEEE Computational Intelligence Magazine*, 2014, **9**(4): 62−74

69 Vijayakumar S, D′Souza A, Schaal S. Incremental online learning in high dimensions. *Neural Computation*, 2005, **17**(12): 2602−2634

70 Hunter A, Summerton R. Fusion rules for context-dependent aggregation of structured news reports. *Journal of Applied Non-Classical Logics*, 2004, **14**(3): 329−366

71 Žliobaitė I. Learning under concept drift: an overview. *Computer Science — Artificial Intelligence* [Online], available: http://arxiv.org/abs/1010.4784, May 31, 2015

72 Li Jian-Zhong, Liu Xian-Min. An important aspect of big data: data usability. *Journal of Computer Research and Development*, 2013, **50**(6): 1147−1162
(李建中, 刘显敏. 大数据的一个重要方面: 数据可用性. 计算机研究与发展, 2013, **50**(6): 1147−1162)

73 Samarati P, Sweeney L. Protecting privacy when disclosing information: *k*-anonymity and its enforcement through generalization and suppression. In: Proceedings of the 1998 IEEE Symposium on Research in Security and Privacy. Palo Alto, CA: IEEE, 1998. 1−19

74 Wang Chao, Yang Jing, Zhang Jian-Pei. Research on trajectory privacy preserving method based on trajectory characteristics and dynamic proximity. *Acta Automatica Sinica*, 2015, **41**(2): 330−341
(王超, 杨静, 张健沛. 基于轨迹特征及动态邻近性的轨迹匿名方法研究. 自动化学报, 2015, **41**(2): 330−341)

75 Wu X D, Zhu X Q. Mining with noise knowledge: error-aware data mining. *IEEE Transactions on Systems, Man, and Cybernetics — Part A: Systems and Humans*, 2008, **38**(4): 917−932

76 He H B, Garcia E A. Learning from imbalanced data. *IEEE Transactions on Knowledge and Data Engineering*, 2009, **21**(9): 1263−1284

77 王飞跃. 迈向知识自动化 [Online], available: http://www.cas.cn/xw/zjsd/201401/t20140103_4009925.shtml, June 1, 2016

78 Deng Jian-Ling, Wang Fei-Yue, Chen Yao-Bin, Zhao Xiang-Yang. From industries 4.0 to energy 5.0: concept and framework of intelligent energy systems. *Acta Automatica Sinica*, 2015, **41**(12): 2003−2016
(邓建玲, 王飞跃, 陈耀斌, 赵向阳. 从工业 4.0 到能源 5.0: 智能能源系统的概念、内涵及体系框架. 自动化学报, 2015, **41**(12): 2003−2016)

79 Twitter Blog. Dispatch from the Denver debate [Online], available: http://blog.twitter.com/2012/100dispatch-reom-denver-debate.html, October 1, 2012

80 Chun D X, Jun C J, Zhong C Y, Chao T M, Cong P. Data engineering in information system construction. In: Proceedings of the 2012 IEEE Symposium on Robotics and Applications (ISRA). Kuala Lumpur: IEEE, 2012. 135−137

81 Aggarwal C C. *Data Streams: Models and Algorithms (Advances in Database Systems)*. US: Springer, 2007.

82 Silva J A, Faria E R, Barros R C, Hruschka E R, de Carvalho A C P L F, Gama J. Data stream clustering: a survey. *ACM Computing Surveys*, 2013, **46**(1): Article No. 13

83 Patil P D, Kulkarni P. Adaptive supervised learning model for training set selection under concept drift data streams. In: Proceedings of the 2013 International Conference on Cloud and Ubiquitous Computing and Emerging Technologies. Pune: IEEE, 2013. 36−41

84 Hakkani-Tür D, Heck L, Tur G. Using a knowledge graph and query click logs for unsupervised learning of relation detection. In: Proceedings of the 2013 IEEE International Conference on Acoustics, Speech, and Signal Processing. Vancouver, BC: IEEE, 2013. 8327−8331

85 Dantas J R V, Farias P P M. Conceptual navigation in knowledge management environments using NavCon. *Information Processing and Management*, 2010, **46**(4): 413−425

86 Xu C J, Li A P, Liu X M. Knowledge fusion and evaluation system with fusion-knowledge measure. In: Proceedings of the 2nd International Symposium on Computational Intelligence and Design. Changsha, China: IEEE, 2009. 127−131

87 Shahabi C, Zarkesh A M, Adibi J, Shah V. Knowledge discovery from users web-page navigation. In: Proceedings of the 7th International Workshop on Research Issues in Data Engineering. Birmingham: IEEE, 1997. 20−29

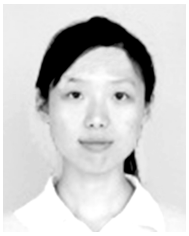
88 Baldauf M, Dustdar S, Rosenberg F. A survey on context-aware systems. *International Journal of Ad Hoc and Ubiquitous Computing*, 2007, **2**(4): 263−277

89 Herlocker J L, Konstan J A, Terveen L G, Riedl J T. Evaluating collaborative filtering recommender systems. *ACM Transactions on Information Systems*, 2004, **22**(1): 5−53

90 Yue Yuan-Long, Zuo Xin, Luo Xiong-Lin. Improving measurement reliability with biased estimation for multi-sensor data fusion. *Acta Automatica Sinica*, 2014, **40**(9): 1843−1852
(岳元龙, 左信, 罗雄麟. 提高测量可靠性的多传感器数据融合有偏估计方法. 自动化学报, 2014, **40**(9): 1843−1852)

91 Xu C, Zhang Y Q, Li R Z. On the feasibility of distributed kernel regression for big data. *Statistics* [Online], available: http://arxiv.org/abs/1505.00869, May 31, 2016

吴信东 长江学者, “千人” 计划特聘教授, IEEE Fellow, AAAS Fellow. 合肥工业大学计算机与信息学院教授. 美国佛蒙特大学计算机与科学系教授. 1993 年获得英国爱丁堡大学人工智能博士学位. 主要研究方向为数据挖掘, 知识库系统, 万维网信息探索. 本文通信作者.
E-mail: xwu@hfut.edu.cn
(**WU Xin-Dong** Professor at the College of Computer Science and Information Engineering, Hefei University of Technology; professor in the Department of Computer Science, the University of Vermont. He received his Ph. D. degree from the University of Edinburgh in 1993. His research interest covers data mining, knowledge based systems, and Web information exploration. Corresponding author of this paper.)



何 进 合肥工业大学计算机与信息学院硕士研究生. 2015 年获得安徽财经大学计算机科学与技术系学士学位. 主要研究方向为数据挖掘和大数据分析.
E-mail: flyingfish93319@126.com
(**HE Jin** Master student at the College of Computer Science and Information Engineering, Hefei University of Technology. She received her bachelor degree from Anhui Finance and Economics University in 2015. Her research interest covers data mining and big data analytics.)

陆汝钤 中国科学院院士. 1959 年获得德国耶拿大学数学系学士学位. 主要研究方向为知识工程, 基于知识的软件工程, 人工智能. E-mail: rqlu@math.ac.cn
(**LU Ru-Qian** Member of the Chinese Academy of Sciences. He received his bachelor degree from the University of Jena (Germany) in 1959. His research interest covers knowledge engineering, knowledge based software engineering, and artificial intelligence.)

郑南宁 中国工程院院士, IEEE Fellow, 西安交通大学教授. 1985 年获得日本庆应大学工学博士学位. 主要研究方向为模式识别, 机器视觉与图像处理.
E-mail: nnzheng@mail.xjtu.edu.cn
(**ZHENG Nan-Ning** Member of the Chinese Academy of Engineering, IEEE Fellow, and professor at Xi′an Jiaotong University. He received his Ph. D. degree from Keio University (Japan) in 1985. His research interest covers pattern recognition, machine vision, and image processing.)