从大数据到大知识:HACE+BigKE

吴信东

本报告主要包括:1)大数据研究背景与动机;2)5V,5R,4P与HACE定理;3)大数据知识工程;4)大知识的挑战与前景展望4个方面的内容。

1 大数据的研究背景与动机

20世纪90年代,"数据仓库之父"Bill Inmon开始关注大数据(海量数据)。简而言之,大数据是无法在合理的时间内利用现有的数据处理手段进行诸如存储、管理、抓取等分析和处理的数据集合。数据收集、存储和互联网技术的快速发展使得数据复杂度得到提升,主要表现在:1)海量数据(Large-volume)——信息的爆炸式增长、用户由信息接收者转变为创造者;2)异构和自治的数据源(Heterogeneous & Autonomous)——网络规模的扩张和复杂的网络结构;3)多源数据(Multi-sources)——音频、视频、移动数据源等。上述背景促使我们挖掘和分析大数据间复杂和演化的关联,即如何寻找一种标准化的数据表示方式并集成异构数据,以一种合理的网络结构描述数据间的语义关联。在实际应用中,主要是从大数据中提取和分析有价值的大知识,而人们关心的是大数据能提供的服务。

大数据与大知识的关系如图1所示。

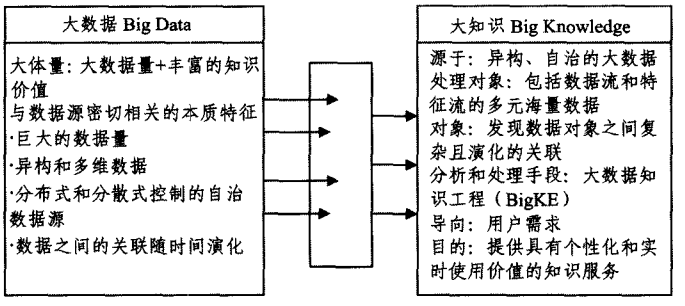

图1 大数据与大知识的关系

2 5V,5R,4P与HACE定理

现有的大数据模型具有5V特征:海量(Volume)、快速(Velocity)、多样化(Variety)、真实性(Veracity)、价值(Value)。Gartner 2014-2015新兴技术发展周期简评表明大数据已从概念转入实践应用阶段。因此在数据的分析和提取中,应该利用更好的创新数据分析算法来提升数据的真实性和价值。

大数据的管理与商用——5R模型。5R模型的主要内容包括Relevant(相关性)、Real-Time(实时性)、Realistic(真实性)、Reliable(可靠性)、ROI(Return on Investment,投资回报)。5R模型基于商业应用的背景,提出对数据的来源和知识的获取需要提供成本预算,用于评估获取知识的价值和可行性。同时,关注数据的组织形式,采取可信的数据组织形式,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"。

大数据与医学应用——4P医学模型。专家系统(Expert System,ES)作为人工智能的分支,已被大量运用到工程、科学、医学预测等方面。大数据背景下的普适医疗(Pervasive Healthcare)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。"4P"具体是指Predictive(预测性)、Preventive(预防性)、Personalized(个性化)、Participatory(参与性)。将个性化的服务和预测相结合能提供基于大数据的个性化医疗建议。而"个体化"强调个体行为的重要性,应因人制宜,由于同种疾病发病原因的多样化,因此同一种疾病采取多样化的治疗方法;参与性是一个科学问题,医学概念的更新、理论框架的构建应该与实践、个人、社交信息和专家知识的异构数据融合。基于4P医学模型,形成了以医学大数据知识工程为核心,包括预测医学、参与医学、预防医学、个性化医学在内的4P医学应用。

HACE定理是指大数据始于异构(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。

一种多层的大数据处理框架如图 2 所示。

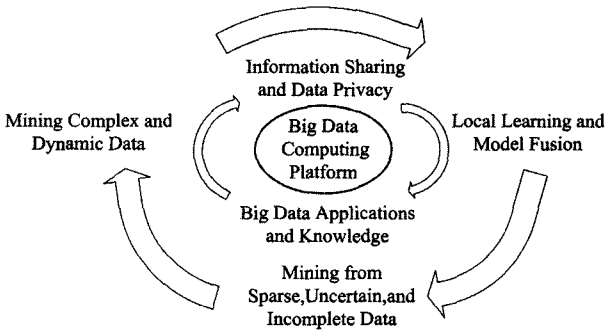

图 2　一种多层的大数据处理框架

第一层:数据计算平台。使用带有较高计算性能的集群计算机(cluster computers),每个计算节点都可以并行处理计算任务,以降低单个计算机的计算量,从而减小对每个计算节点的硬件的依赖性。

第二层:大数据的语义和应用知识。包括信息的共享和隐私,以及大数据应用与知识。

第三层:大数据挖掘算法。出于保护数据隐私的考虑,无法将多个节点的局部数据简单地线性集成为一个全局数据。其中涉及的问题包括:1)局部学习和多信息源的模型融合;2)稀疏、不确定和不完整的数据挖掘;3)挖掘复杂的动态数据。

3　大数据知识工程

知识工程是人工智能的原理和方法,对那些需要专家知识才能解决的应用难题提供求解的手段。恰当运用专家知识的获取、表达和推理过程的构成与解释,是设计基于知识的系统的重要技术问题。中国科学院数学与系统科学研究院陆汝钤院士于 2013 中国大数据技术大会-大数据研究与发展专题论坛中总结并提出了知识工程的发展进程,如图 3 所示。

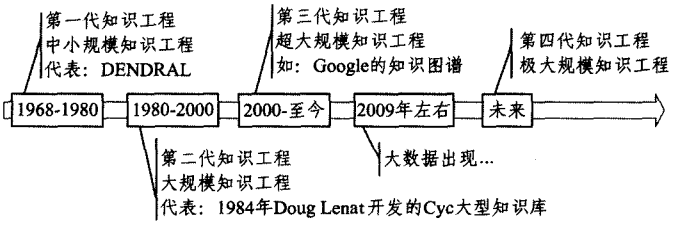

图 3　知识工程发展进程

User Generated Content 形式的数据源作为人类知识的重要载体,有助于突破基于专家知识的传统知识工程中的"知识获取瓶颈"问题,加剧了知识碎片化:与特定主题相关的碎片化知识高度分布在自治的数据源中。

知识碎片化有 3 个表现:即时动态性,低质化,无序化。其所对应的 3 个挑战分别为:1)多源分布的碎片化知识具有多模态、多粒度、时变特性,如何表示既能刻画其演化规律,又能支持知识获取与融合? 2)碎片化知识质量上良莠不齐,在真实性、自洽性、完备性等方面存在不足,如何实现量质转换与知识增殖(知识推理和发现)? 3)如何对碎片化知识进行动态有序化,以导航方式适配用户需求与场景的多变性? 针对大数据环境下多源异构自治数据源产生的碎片化知识,刻画其表示与演化规律,通过融合碎片化知识实现量质转换与知识增殖,同时对融合知识进行动态有序化组织,以满足不同用户需求与场景的多变性。

奥巴马政府 2012 年发布的"大数据行动计划"中指出:大数据背景下,利用已有的数据收集、存储、管理、分析和共享技术来转换教育和学习方式恰逢其时。欧盟在 2020 年欧洲信息化发展愿景中指出,利用大数据实现通过个性化服务削减经济成本及为公共领域增效等目标。

大数据知识工程旨在融合碎片化知识,以知识导航的形式提供个性化服务,以满足不同用户需求与场景的多变性。这些特性为公共领域(如医疗、在线教育等)增效提供了有利的保障。

大数据知识工程模型称为 BigKE,BigKE 采用一种 3 层次的知识建模方法,最终获取个性化的知识导航服务。

它包括 3 次层次:1)多源异构数据中的碎片化知识建模;2)从局部知识到全局知识——碎片化知识融合;3)个性化知识导航。详细信息如图 4 所示。

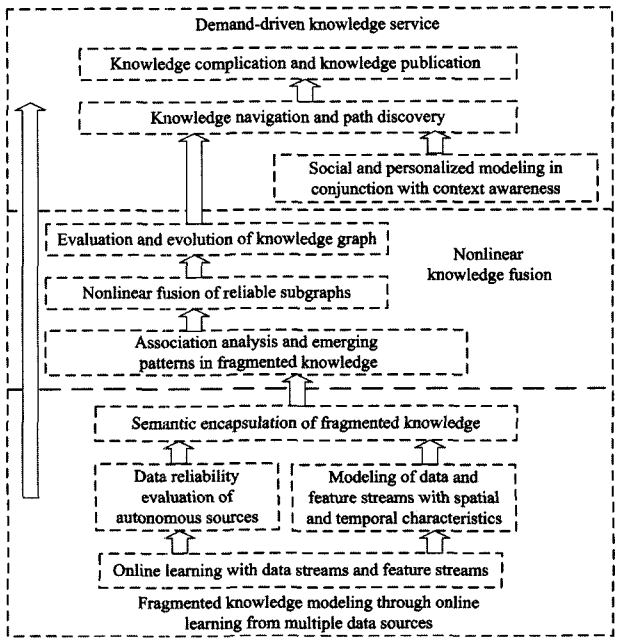

图 4　BigKE 模型

4　大知识的挑战与前景展望

对大知识的挑战与应用前景总结如下几点。

挑战:

(1)碎片化知识的非线性融合

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。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。

(2)大知识图谱的动态更新

基于对预测未来趋势的需求,大数据知识工程需要根据不断到来的数据,对大知识图谱的结构和内容进行动态更新。大知识图谱的动态更新主要涉及两大问题:1)如何设置合理的时间点来更新现有知识图谱;2)如何确定对某一数据关联的取舍问题。

(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。然而对碎片打下的粒度划分,涉及到对知识或碎片的划分问题,对知识的分拆和重组也提出了新的挑战问题。

(4)海量碎片化知识的约化表示

约化的含义是把统一知识的复杂表示 A 转换为简单表示 B,使得 B 的容量远远小于 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”。海量碎片化知识的约化表示涉及到降维等问题。

(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。如何提高数据的处理和计算效率,同时解决一些统计值不正确的问题,在满足“一致相合”的条件下提供合理的结果,是我们在大数据知识集成时需要应对的。

(6)个性化用户行为的建模

个性化的大知识服务的关键在于对个人和社交信息的建模。知识图谱的结构随着时间变化,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。

应用前景:

(1)构建动态网络大辞典

动态网络大辞典的构建涉及到动态数据特征的抓取、表示以及通过加工合成形成新的词语信息对语料库进行更新。异构的大数据信息源还需要对内容建立合理的评估体系。大数据的多源特征决定了动态网络大辞典应当在时间和空间维度涵盖尽可能多和广的信息。

(2)网络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,然而在新闻的数量快速增长的同时,新闻的质量并没有同步提升,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树立和表示出新闻事件的多个主题,全面和针对性地获取新闻中重要的本质。

(3)普适医疗信息的管理与服务

大知识与普适医疗相结合,可以建立和动态更新医疗推荐系统。对地理位置、用户病史和社交偏好等信息的融合与分析,以及对医疗数据和知识的推荐进行评价系统的改进,能够提供个性化的医疗管理与服务。

(4)万维网就业培训

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,依据用户的信息进行数据筛选和过滤,结合用户个人的就业倾向等进行比对,反馈从已有万维网的大知识图谱中寻找某一针对性问题的映射,为用户提供市场分析和技能培训。

(5)自动编辑和出版

对"基于集成和拆解的知识重组"这一挑战问题如果能够得到很好的解决,自动知识编辑的目标就可以实现。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社会需求就不再是一个大量耗费人力和财力的事业。

(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。

参 考 文 献

[1] Wu Xindong, Zhu Xing-Quan, Wu Gong-Qing, et al. Data Mining with Big Data[J]. IEEE Transactions on Knowledge and Data Engineering, 2014, 26(1): 97-107

[2] Wu Xin-dong, Chen Huan-huan, Wu Gong-qing, et al. Knowledge Engineering with Big Data[J]. Intelligent Systems IEEE, 2015, 30(5): 46-55

本文根据 2016 年第六届中国数据挖掘会议(CCDM2016)大会报告"从大数据到大知识:HACE+BigKE"整理。

(整理:李亚辉,何进)

吴信东 美国佛蒙特大学计算机科学系教授,合肥工业大学长江学者,主要研究方向为数据挖掘、基于知识的系统、万维网信息探索。