doi:10.3772/j.issn.2095-915x.2015.06.007

大数据知识计算的挑战

孙晓平

(中国科学院计算技术研究所智能信息处理重点实验室 北京 100190)

摘要: 知识计算是传统人工智能研究的一个重要领域。知识计算研究知识的表达、管理、获取和使用的计算问题。本文探讨了大数据时代,知识计算技术面临新的挑战 and 机遇,分析了知识的表达、获取和计算在大数据时代的新需求,以及知识计算模型和方法可能需要的新的突破以解决大数据的动态、开放、异构、海量等性质所带来的知识计算问题,提出了开放性、一致性、完备性和上下文等几个大数据知识计算所面临的关键问题。

关键词: 大数据, 知识计算

分类号: TP3-05

Challenges of Knowledge Computing in Big Data

Sun Xiaoping

(Key Lab of Intelligent Information Processing, Institute of Computing Technology, Chinese Academy of Sciences, Beijing 100190)

Abstract: Knowledge computing is an important research area in artificial intelligence. It includes knowledge representation, management, acquiring and utilization. In this paper, we investigate the new challenges and opportunities of knowledge computing research in big data era. We discuss the problems of the representation acquiring and utilization of knowledge in big data and propose that openness, consistence, completeness and context should be investigated in conducting knowledge computing in big data.

Key words: Big data, knowledge computing

作者简介: 孙晓平(1973年9月),男,副研究员,大规模网络和语义计算。北京市海淀区中关村科学院南路6号,中国科学院计算技术研究所,010-62600546, sunxiaoping@ict.ac.cn。

1 引言

大数据时代研究人员面临着新的挑战和机遇^[1]。与传统数据不同,3V或5V是大数据的主要性质¹。Volume: 体量,大数据时代的数据体量非常之大,如Google或Baidu等这样的搜索引擎公司每天几乎都需要产生几十PB(Peta Byte)的数据。Variety: 类别,多样性的数据结构类型包括如图片,视频,文本,半结构化数据等。Velocity: 速度,数据产生的速度相当快,如淘宝天猫节假日单日访问量就可以达到6000万次。Variability: 变化,所有这些特征都是在不断的变化。另外一个经常提及性质的是Veracity: 真实性,由于网络时代是个开放的时代,什么人都可以发布数据,造成数据的真实性难以得到保证,从海量大数据中去伪存真则成为一个巨大的挑战。大数据概念的提出并受到学术界和产业界的持续关注,成为技术热点,首先是由于大数据的商业价值^[2]。

数据意味着需求。Web时代自上世纪90年代起,网页数据指数级膨胀,研究人员已经开始关注海量异构Web数据的分析处理问题^[3]。然而,大数据作为一个新的概念出现,明显不同于之前海量Web数据分析处理,主要在于大数据时代的数据往往意味着用户的需求而非简单的Web页面展示。Web 2.0时代,电商,社交网络飞速发展,现在的大数据往往包含了大量的用户需求信息,如用户浏览产品,社交网络对产品服务的评价等等。信息不再孤立,大数据往往包含了多种用户信息,以往难以获取的用户特征在大数据时代成为最具价值的商业数据。如用户的兴趣,用户的生活习惯,日常活动,健康指标,情感好恶,流行趋势等等。这些信息都直接关系了市场扩展、

细分、发现、定位、产品推送改进等等。捕获这些信息对于企业的发展至关重要。这在以往单纯以信息发布为主要目的Web1.0页面时代是不可想象的。这些信息都直接反应的用户的需求,而用户是企业运营的基础与核心,掌握用户往往意味着掌握了市场。这是为什么企业开始关注大数据的关键原因。这些因素使企业看到了大数据的商业价值,从而对大数据格外的关注,不断推动大数据相关技术发展。Google, IBM, Microsoft, 阿里巴巴, 腾讯, 百度等纷纷加入大数据市场,推出了各自的大数据平台。Hadoop, MapReduce, Key-Value 数据库等开源数据管理和计算平台为大数据的推广使用提供了强有力的支持,使大数据的管理使用成为可能。

从科学研究角度看,大数据的价值早就为科学家所认识到。2007年,Jim Gray就提出了数据科学的问题,指出了第四次科学革命可能产生数据科学,即科学研究的第四范式^[4]。Jim Gray指出数据密集型的科学研究将与之前的模型驱动型的科学研究有本质不同。Jim Gray强调数据密集型的科学研究将给数据管理,科学计算乃至科学研究本身都带来新的挑战,更重要的是,Jim Gray强调更多来自于计算和数据管理领域的技术和方法将为传统科学研究方法提供新的思路和平台。他同时建议将所有的科研数据公开到网上,从原始数据到科学文献,形成一个系统的数据平台。今天,美国政府已经启动了大数据计划,开放了国家大数据平台²,开放了大量的科研数据和接口,供研究人员和开发人员使用,成为美国大数据行动计划的一部分。美国政府启动的大数据计划不仅包含了数据分析师的人才培养计划,也包含了卫生、健康、安全、军事等多个方面的大数据核心业务发展规划。

1 https://en.wikipedia.org/wiki/Big_data

2 <https://www.data.gov/>

2 大数据应用中的知识计算

大数据的应用中,无论是商业智能^[2],医药健康卫生^[5],图书情报^[6]或是科学研究^[4],数据的价值体现在数据中所蕴含的信息和知识。因此,如何从大数据中发掘信息和知识成为重要的目标和挑战。尽管数据管理、数据挖掘、信息检索、知识发现是长久以来的科学研究挑战^{[7], [8]},但这些领域的传统研究往往会限定目标,限定数据范围和规则,在处理大数据时往往需要进行调整^[9-15]。因此,开放领域信息抽取成为目前的一个研究热点^[16]。即不限定领域和问题的信息获取。在大数据时代,我们往往已经无法预知数据中到底有什么,没有什么,但因为数据量大,我们认为大数据中会蕴含更多有意义的信息。因此如何让机器去探索往往是一个迫切需要解决的问题。当然,无指导的开放信息检索模型还停留在初级阶段,但很多模型还是取得了不错的进展。尤其是在知识图谱^[17]、语义关系检索^[18]方面,一些新的模型对数据的处理取得了令人瞩目的进步。知识图谱、语义关系检索发现来自语义网(Semantic Web)的发展。语义网是万维网时代对HTML等半结构化数据表示模型的一次语义扩充尝试^[19]。T. Bernes Lee等人希望借助语义网标准来扩充并统一万维网的语义信息表达,实现数据在语义层面的互联和共享。语义网提供了不同语义层次的标准语义描述语言,RDF(Resource Description Framework)定义了最基本的语义元素和关系的标准关键词,标准语义和一般性约束^[20]。OWL则可以扩展到描述逻辑,甚至是一阶逻辑语义层面,用于表示类,实例,推理等语义知识信息^[21]。RDF通常以三元组的形式出现,表示主体、关系、客体对象的这三者之间信息。RDF的作为知

识表示的基础,可以扩展描述各种语义关系。如DBPedia^[22]和Yago^[23]等三元组本体集合是对RDF的基本扩展,包含了大量的现实世界对类,对象,关系的描述信息。一方面,这些语义信息作为本体可以支持大量的语义计算。另一方面,如何从现实数据信息中获取三元组关系信息也成为研究重点。在三元组之上搭建大数据分析平台成为可能,如OpenKN平台旨在提供系统的知识检索服务^[24],甚至是搭建更为宏观的社会计算系统^[25]。因此,知识图谱的抽取构建成为一个重要的研究问题^[17]。从大数据的角度来看,发现、抽取知识是大数据研究和应用的关键^[1, 25-27]。知识计算成为大数据研究的一个顶层应用问题^[28, 29]。一方面,业界普遍期望更智能的信息服务。另一方面,大数据本身提供了这样的可能解决方案。

3 大数据应用中知识计算的挑战

知识计算是与知识表达、存储、管理相关的计算模型方法的统称。人工智能的一个终极目标是图灵测试,想要逼近这样一个目标,知识的分析处理是必不可少的一步。然而知识是什么存在着大量的争议和不确定性。知识可以看作是对某个或某些事物的认知和理解³,通过经验或接受教育(如感知,发现或学习)来获得的如事实,信息,描述,技能等等。不论知识到底是如何存在于人脑中,至少从人类智能活动的表象来看,知识在智能的实现和运作中起到了至关重要的作用,这不能不让我们考虑解决知识计算问题可能是实现智能至关重要的一部。另一方面知识计算的一个基础问题是知识的存在性问题,即数据中有什么。

大数据一个特点是大,也即意味着大数据很

3 <https://en.wikipedia.org/wiki/Knowledge>

可能涵盖了更大范围内的信息,也就是涵盖了更多的知识信息,可以被机器所发现获取。或者说,如果数据中根本没有想要的信息和知识,那么机器也不可能发现这样的信息。而大数据则可以更大程度上避免这样的问题。但这里面我们还是需要对知识和知识计算有一个更清晰的认识。

3.1 什么是知识

什么是知识?首先我们的目的不是探讨知识的本源、本质、内在属性等哲学问题,或者设计一个万全的模型来解释人类知识的来龙去脉。我们需要讨论的是如何借助计算机来表示某种知识,知识的某个方面,某个层次,某个维度,某个度量性质。比如,我们可以认为 RDF 三元组是一种面向计算机的知识表达和查询检索模型。再比如,用 lambda 代数来表达知识^[30]。这样,我们所能表示的知识局限于我们所设定的语言及其表达能力和约束。当需要对大数据中的知识进行提取分析时,这种预先定义的知识模型则成为重要的先验知识施加于计算模型中,以实现我们的目标。比如,当我们需要提取“x 位于 y”,“x 反对 y”之类语义信息时,我们需要一个形式化的知识定义,即 $\langle x, R, y \rangle$ 这样的三元组。当然,我们可以对 x, y 和 R 做进一步限制,比如, x, y 需要是名词, R 是某一个指定的关键词或一类关键词。命名实体关系识别模型可以进一步扩展,如加入否定约束,数量约束,时间约束,空间约束等语义约束。当我们在定义这样一个语义模型时,实际上我们就是在定义一个知识模型。再比如在 lambda 演算模型 (Lambda calculus) 中, $\lambda x.name(x, people)$ 可以用来定义一个人的姓名这样的信息。用 lambda 演算已经可以表达非常多的语义信息,因为作为一种语言 lambda 演算等同于图灵机模型。但如果这样还不够明显的话,我们还可以用一阶逻辑来描述复杂的约束信息,如,“年龄小于 15 岁的

学生”可以用 $\exists x \text{ age}(x) < 15 \text{ and student}(x)$ 这样的方式来表示。推理也是知识表示的一部分。比如,“奥巴马是美国总统” “美国总统必须是美国人”,所以“奥巴马是美国人”。得出这样的结论当然需要一定的先验推理规则和先验知识,比如这里面用到了“美国总统必须是美国人”这样的先验知识,以及“奥巴马是美国总统”这样的知识,同时,还需要运用一个三段论式的推理规则才能得出后面的结论。因此,针对大数据中的知识计算问题,我们需要首先明确我们要找到的知识的形式是什么,如何解释利用这样的知识。三元组,一阶逻辑,图,树,都可以是作为知识表示的具体形式,只要他们能够表达我们需要的特定信息,如“奥巴马是美国总统”这样的知识条目。

从上述来看,目前知识的定义仍然沿用之前已有的知识定义模型和方法,如语义网,三元组,一阶逻辑等等,这些定义本身并不区分知识是源于基于大数据或其他数据源。对于大数据来说,知识定义模型是否需要新的研究工作来解决新的问题或挑战呢?我们可能需要进一步分析探究从需求到数据本身的特点。我们假定的大数据或知识主要是文本或自然语言所描述的信息,比如社交网络上的信息,网页信息等等,暂时不考虑图片,视频以及各种结构化数据源。那么在如果考虑大数据中的知识的定义和计算时,我们至少需要考虑一下几个问题。

3.2 知识计算的开放性

知识模型的严谨性体现在一致性,完备性等特点上。如语义网,一阶逻辑等,都需要处理知识的一致性,正确性,完备性。如,不能有矛盾的结论,所有事物要么存在,要么不存在,要么正确,要么不正确,推出的一定是正确的结论,正确的结论一定会被推出等。但在开放世界中,不存在的并不一定是否定的,而可能是不知道的,

这是开放世界假设^[31]和封闭世界假设^[32]的区别。为了应对这一问题,语义网采用了开放世界假设,即没有提到的知识不是不存在的或否定的信息。即,没有被描述的信息有可能也是真的信息。例如,我们的系统中只有这样一条表示为三元组的信息,论断或知识: <克林顿,是,美国人>,即此推断为真,那么给另一个推断<奥巴马,是,美国人>,判断其是否为真。在开放世界假设中这一推断的结果为“未知”,而在封闭世界假定前提下,此论断则为假,因为没有任何一条系统已知的知识可以推出<奥巴马,是,美国人>。

可以看出,开放性假设对推断结果起着至关重要的影响。换言之,我们在从大数据中提取三元组等逻辑推断时,界定没有发现,不存在,或暂时不存在的知识在系统中如何处理是需要关注的首要问题。如果系统仅仅是提取三元组而不考虑三元组之间的二次推理或融合的结果,问题似乎还不算大,一旦涉及推理时,如何处理不存在、未知的知识是一个重要的工作。通常来说,开放世界假设是必然的,即,不知道的知识当作是未知而非否定。

3.3 知识计算的一致性

一致性:我们希望所发现提取的知识中不要存在互相矛盾的知识。如果说开放性还不具备一定破坏性,那么存在矛盾的知识则是更为严重的问题。比如,“奥巴马是加拿大人”,“奥巴马是美国总统”这样的本质上是矛盾的信息。相较未知信息而言,不一致的知识往往意味着一个错误的知识。一致性问题的难点在于检测不一致:即,如何让一个知识发现或提取模型能够自动的检测知识的不一致。但即便在封闭的逻辑系统中这也是一个难题,如结构化数据中的一致性检测问题,甚至是不可判定的问题,如一阶逻辑中的矛盾命题的判定。精确判断不一致的知识实际上几乎是

不可能的。我们可以给每一个知识点赋予一个向量,然后判断向量之间距离来决定知识点之间的相似程度,但不相似并不意味着相反,或矛盾。如何区分不相似、相反和矛盾是一个极具挑战的任务。借助先验的本体知识进行判断是一个可行的思路。即,我们可以借助预先建立的本体知识来辅助判断所提取的知识是否存在矛盾,或借助本体的严格一致性来推导出提取知识的一致性。因此,在上一例子中,如果我们有这样的本体知识,即美国总统必须是美国人,我们还是有可能推断出“奥巴马是加拿大人”和“奥巴马是美国总统”这样的信息是矛盾的,二者必有其一是错误的。但问题是,这样的一致性检测复杂度非常之高。

3.4 知识计算的完备性

所谓完备性是指正确的论断一定能够推出。这一性质在严格的形式化逻辑系统中可以得到保障。但在开放大数据环境下,这一性质几乎不可能得到满足。从应用的角度看,我们的确不需要保证所有正确的论断都是可以推出的,尤其是大量平凡的结论,比如“奥巴马是奥巴马”,“奥巴马的夫人的先生是奥巴马”这样的知识。但我们往往希望知识计算模型能够发现提取所谓重要的知识,也就是说对于重要的知识的完备性。我们希望能够得到一定程度的保证。但问题来了,我们如何判断一个知识是重要的还是不重要的,甚至平凡的呢?实际上,我们自身有时候也很难判断一条知识是否是重要的或不重要的,比如“奥巴马是美国总统”这样的知识在不同场合下,不同背景知识下,不同应用场景中的重要性显然不同,对于政治类话题,对于中学生,对于一般性国际政治新闻,这样的信息可以看作是重要的知识点。但对于重要的科学文献,技术报告,政府工作报告,财政报告,政府间合同,法律条文等专业知识丰富的文本资源来说,“奥巴马是美国总统”

的知识点则显得过于平凡了,对于这些文本的读者来说,大量的专业知识可能是更为重要的知识点,比如“奥巴马政府的预算分配政策”等等。当然,不重要的知识可能是重复非常少的知识,在此类专业文献中,“奥巴马是美国总统”这样的信息可能出现次数很少,可以忽略。但问题是,重要的知识可能未必是重复出现次数多的知识。比如,奥巴马的简历可能只有一个地方是关于“奥巴马是美国总统”的信息,但这样的知识点对于一份简历来说还是非常重要的知识信息。因此,我们需要考虑的不仅仅是重要性问题的完备问题,还需要考虑的是上下文的问题。

3.5 知识计算的上下文

所谓上下文,用不准确的语言描述就是确定一个知识点所处的当下知识语境是什么。或者说所讨论的知识的范畴是什么。比如,我们阅读一份简历与阅读一份政府报告,其中所获得知识的上下文是截然不同的。当然,这需要我们确定一个问题,即,什么是知识的上下文。

我们可以这样定义^[33]:上下文通常由一组知识点来组成,这组知识点是其中每一个知识点的

上下文。对于同一个知识点,可以处于不同的上下文环境中,甚至处于不同层次结构的上下文之中。也就是说,上下文环境是相互包含的。如果我们将上下文看做是知识的一种,这就再次涉及到知识的定义,或者说如何将上下文引入到知识的表示和处理中^[34-36]。然而,这方面的工作开展甚少,应该说是知识计算中的一个重要但尚待进一步发掘的问题。

4 总结

文中我们回顾了大数据的发展和挑战,并着重探讨了知识计算在大数据中的位置,以及所面临的挑战,并指出了几个可能需要进一步发掘探讨的问题,包括知识的表示,知识的一致性,完备性,上下文等重要性质的处理和应对。近年来,深度神经网络模型在统计机器学习领域取得了突飞猛进的进展^[37,38]。如已有工作借助递归神经网络来处理逻辑推理、存在、否定、嵌套等多种复杂语义关系的发现和提取^[39]。这些模型和方法或许将有助于我们进一步应对大数据知识计算中遇到的多种问题。

参考文献

[1] 李国杰. 大数据研究的科学价值 [J], 中国计算机学会通讯, 2012, 8 (9): 8-15.

[2] CHEN H, CHIANG R H, Storey V C. Business Intelligence and Analytics: From Big Data to Big Impact [J]., MIS quarterly, 2012, 36 (4):1165-1188.

[3] ABITEBOUL S, BUNEMAN P, SUCIU D. Data on the Web: from relations to semistructured data and XML [M]. Morgan Kaufmann, 2000.

[4] HEY A J, TANSLEY S, TOLLE K M. The fourth paradigm: data-intensive scientific discovery [D]. Microsoft Research Redmond, WA, 2009.

[5] MARX V. Biology: The big challenges of big data [J]. Nature, 2013, 498 (7453):255-260.

[6] 韩翠峰. 大数据带给图书馆的影响与挑战 [J]. 图书与情报, 2012, 147 (05):37-40.

[7] BERSON A, SMITH S J. Data warehousing, data mining, and OLAP [M]. McGraw-Hill, Inc., 1997.

[8] BERRY M J, LINOFF G S. Data mining techniques: for marketing, sales, and customer relationship management[M]. John Wiley & Sons, 2004.

[9] WITTEN I H, FRANK E. Data Mining: Practical

- machine learning tools and techniques [M]. Morgan Kaufmann, 2005.
- [10]MADDEN S. From databases to big data [J]. IEEE Internet Computing, 2012, 2012 (3):4-6.
- [11]FAN W, BIFET A. Mining big data: current status, and forecast to the future [J]. ACM SIGKDD Explorations Newsletter, 2013, 14 (2): 1-5.
- [12]LIN J, AND RYABOY D. Scaling big data mining infrastructure: the twitter experience [J]. ACM SIGKDD Explorations Newsletter, 2013, 14(2): 6-19.
- [13]WU X, ZHU X, WU G Q, DING W. Data mining with big data [J]. Knowledge and Data Engineering, IEEE Transactions on, 2014, 26 (1):97-107.
- [14]孟小峰, 慈祥. 大数据管理: 概念, 技术与挑战 [J]. 计算机研究与发展, 2015, 50(1):146-169.
- [15]覃雄派, 王会举, 杜小勇, 王珊. 大数据分析 -RDBMS 与 MapReduce 的竞争与共生 [J]. 软件学报, 2012, 23(1):32-45.
- [16]BANKO M, CAFARELLA M J, SODERLAND S, et al. Open information extraction for the web[C]//IJCAI. 2007, 7: 2670-2676.
- [17]WANG Z, ZHANG J, FENG J, et al. Knowledge Graph Embedding by Translating on Hyperplanes[C]//AAAI. 2014: 1112-1119.
- [18]LIN Y, LIU Z, SUN M, et al. Learning Entity and Relation Embeddings for Knowledge Graph Completion[C]//AAAI. 2015: 2181-2187.
- [19]BERNERS-LEE T, HENDLER J, LASSILA O. The semantic web [J], Scientific american, 2001, 284 (5):28-37.
- [20]KLYNE G, CARROLL J J. Resource description framework (RDF): Concepts and abstract syntax. W3C Recommendation, 2004[J]. World Wide Web Consortium, <http://w3c.org/TR/rdf-concepts>, 2004.
- [21]MCGUINNESS D L, VAN HARMELEN F. OWL web ontology language overview[J]. W3C recommendation, 2004, 10(10): 2004.
- [22]AUER S, BIZER C, KOBILAROV G, et al. Dbpedia: A nucleus for a web of open data[M]. Springer Berlin Heidelberg, 2007.
- [23]FABIAN M S, GJERGJI K, GERHARD W. Yago: A core of semantic knowledge unifying wordnet and wikipedia[C]//16th International World Wide Web Conference, WWW. 2007: 697-706.
- [24]程学旗, 靳小龙, 王元卓, 等. 大数据系统和分析技术综述 [J]. 软件学报, 2014, 25(9): 1889-1908.
- [25]孟小峰, 李勇, 祝建华. 社会计算: 大数据时代的机遇与挑战 [J]. 计算机研究与发展, 2013, 50(12): 2483-2491.
- [26]李国杰, 程学旗. 大数据研究: 未来科技及经济社会发展的重大战略领域—大数据的研究现状与科学思考 [J]. 中国科学院院刊, 2012, (6): 647-657.
- [27]王元卓, 靳小龙, 程学旗. 网络大数据: 现状与展望 [J]. 计算机学报, 2013, 36(6): 1125-1138.
- [28]HILBERT M. Big data for development: From information-to knowledge societies[J]. Available at SSRN 2205145, 2013.
- [29]ZHUGE H. Mapping Big Data into Knowledge Space with Cognitive Cyber-Infrastructure[J]. arXiv preprint arXiv:1507.06500, 2015.
- [30]BARENDREGT H P. The lambda calculus[M]. Amsterdam: North-Holland, 1984.
- [31]RUSSELL S, NORVIG P. Artificial Intelligence: A modern approach[J]. Artificial Intelligence. Prentice-Hall, Egnlewood Cliffs, 1995, 25: 27.
- [32]MINKER J. On indefinite databases and the closed world assumption[C]//6th Conference on Automated Deduction. Springer Berlin Heidelberg, 1982: 292-308.
- [33]CAO M, TIAN J, CHENG D, et al. What Makes it Difficult to Understand a Scientific Literature?[J]. arXiv preprint arXiv:1512.01409, 2015.
- [34]BOUQUET P, GIUNCHIGLIA F, VAN HARMELEN F, et al. C-owl: Contextualizing ontologies[M]//The Semantic Web-ISWC 2003. Springer Berlin Heidelberg, 2003: 164-179.
- [35]WANG X H, ZHANG D Q, GU T, et al.

Ontology based context modeling and reasoning using OWL[C]//Pervasive Computing and Communications Workshops, 2004. Proceedings of the Second IEEE Annual Conference on. Ieee, 2004: 18-22.

[36]BOUQUET P, SERAFINI L, STOERMER H. Introducing Context into RDF Knowledge Bases[C]//SWAP. 2005, 5: 14-16.

[37]BENGIO Y, COURVILLE A, Vincent P. Representation learning: A review and new perspectives[J]. Pattern Analysis and Machine

Intelligence, IEEE Transactions on, 2013, 35(8): 1798-1828.

[38]HERMANN K M, KOCISKY T, GREFFENSTETTE E, et al. Teaching machines to read and comprehend[C]//Advances in Neural Information Processing Systems. 2015: 1684-1692.

[39]BOWMAN S R, POTTS C, MANNING C D. Recursive neural networks can learn logical semantics[J]. arXiv preprint arXiv:1406.1827, 2014.