科学通报 2015 年 第 60 卷 第 5-6 期：485 ~ 490
专辑：大数据研究与应用 进 展 www.scichina.com csb.scichina.com
《中国科学》杂志社 SCIENCE CHINA PRESS

大数据哈希学习：现状与趋势

李武军[①②*]，周志华[①②*]

① 南京大学计算机软件新技术国家重点实验室，南京 210023;
② 软件新技术与产业化协同创新中心，南京 210023
* 联系人, E-mail: liwj@lamda.nju.edu.cn; zhouzh@lamda.nju.edu.cn

2014-09-01 收稿, 2014-11-03 接受, 2015-01-22 网络版发表
国家自然科学基金(61321491, 61472182)和国家高技术研究发展计划(2012AA011003)资助

摘要 随着信息技术的迅速发展，各行各业积累的数据都呈现出爆炸式增长趋势，我们已经进入大数据时代. 大数据在很多领域都具有广阔的应用前景，已经成为国家重要的战略资源，对大数据的存储、管理和分析也已经成为学术界和工业界高度关注的热点. 收集、存储、传输、处理大数据的目的是为了利用大数据，而要有效地利用大数据，机器学习技术必不可少. 因此，大数据机器学习(简称大数据学习)是大数据研究的关键内容之一. 哈希学习通过将数据表示成二进制码的形式，不仅能显著减少数据的存储和通信开销，还能降低数据维度，从而显著提高大数据学习系统的效率. 因此，哈希学习近年来成为大数据学习中的一个研究热点. 本文对这方面的工作进行介绍.

关键词
大数据
机器学习
哈希学习
大数据学习

随着近年来互联网、云计算、物联网、社交媒体以及其他信息技术的迅速发展，各行各业积累的数据都呈现出爆炸式增长趋势. 例如，欧洲粒子物理研究所(CERN)的大型强子对撞机每秒产生的数据高达40 TB(1 TB=1024 GB)，Facebook每天处理的数据超过500 TB，阿里巴巴拥有的数据量超过100 PB (1 PB=1024 TB)，新浪微博用户数超过5亿，每天产生的微博数超过1亿条. 中国移动研究院的一份简报中称，2011年人类创造的数据达到180亿吉字节(GB)，而且每年还在以高于60%的速度增长，预计到2020年，全球每年产生的数据将达到350万亿吉字节(GB). 大数据在科学、金融、气象、医疗、环保、教育、军事、交通等领域都具有非常广阔的应用前景[1,2]. 例如，在科学领域，包括天文、生物、物理、化学、信息等在内的各个领域的科学发现已经从实验型科学发现、理论型科学发现、计算型科学发现发展到第四范式，即基于大数据的数据密集型科学发现[3,4]. 因此可以说，大数据已经成为国家重要的战略资源，对大数据的存储、管理和分析也已经成为学术界和工业界高度关注的热点[1,2].

收集、存储、传输、管理大数据的目的是为了利用大数据，而要有效地利用大数据，机器学习技术[5]必不可少. 事实上，机器学习界一直在尝试对越来越大的数据进行学习[6]，但今日的大数据已不仅仅是数据量大，同时还伴随着数据的多源、动态、价值稀疏等特性，因此为机器学习界提出了一些新的挑战. 近年来，大数据机器学习(或简称为大数据学习)受到了广泛关注[7]，成为机器学习领域的研究热点之一，新成果不断涌现. 例如Kleiner等人[8]基于集成学习中Bagging的思想提出了新型数据采样方法BLB，用来解决Bootstrap在遇到大数据时的计算瓶颈问题；Shalev-Shwartz和Zhang[9]基于随机(在线)学习的思想提出了梯度上升(下降)的改进方法，用来实现大规模模型的快速学习；Gonzalez等人[10]提出了基于多机集群的分布式机器学习框架GraphLab，用以实现基于图的大规模机器学习；Gao等人[11]提出了“单遍学

引用格式: 李武军, 周志华. 大数据哈希学习：现状与趋势. 科学通报, 2015, 60: 485–490
Li W J, Zhou Z H. Learning to hash for big data: Current status and future trends (in Chinese). Chin Sci Bull, 2015, 60: 485–490, doi: 10.1360/N972014-00841

习”(one-pass learning)的思想，力图在学习中只扫描一遍数据、且使用常数级存储来保存中间计算结果，在AUC优化这样的复杂学习任务上已取得很好的效果. 此外还有很多新进展，本文不再赘述.

哈希学习(learning to hash)[12~22]通过机器学习机制将数据映射成二进制串的形式，能显著减少数据的存储和通信开销，从而有效提高学习系统的效率. 哈希学习的目的是学到数据的二进制哈希码表示，使得哈希码尽可能地保持原空间中的近邻关系，即保相似性. 具体来说，每个数据点会被一个紧凑的二进制串编码，在原空间中相似的2个点应当被映射到哈希码空间中相似的2个点. 图1是哈希学习的示意图，以图像数据为例，原始图像表示是某种经过特征抽取后的高维实数向量，通过从数据中学习到的哈希函数h变换后，每幅图像被映射到一个8位(bit)的二进制哈希码，原空间中相似的两幅图像将被映射到相似(即海明距离较小)的2个哈希码，而原空间中不相似的两幅图像将被映射到不相似(即海明距离较大)的2个哈希码. 使用哈希码表示数据后，所需要的存储空间会被大幅减小. 举例来说，如果原空间中每个数据样本都被1个1024 B的向量表示，1个包含1亿个样本的数据集要占用100 GB的存储空间. 相反，如果把每个数据样本哈希到1个128位的哈希码，一亿个样本的存储空间只需要1.6 GB. 单台机器(包括配置很高的单台服务器)处理原始表示时，需要不断地进行外内存交换，开销非常大. 但如果用哈希码表示，所有计算都可以在内存中完成，单台普通的个人电脑(PC)也能很快地完成计算. 由于很多学习算法，比如k近邻(kNN)、支持向量机(SVM)等的本质是利用数据的相似性，哈希学习的保相似性将在显著提高学习速度的同时，尽可能地保证精度. 另一方面，因为通过哈希学习得到的哈希码位数(维度)一般会比原空间的维度要低，哈希学习也能降低数据维度，从而减轻维度灾难问题. 因此，哈希学习在大数据学习中占有重要地位.

需特别指出的是，数据库研究领域早已使用二进制哈希码来表示数据[23~25]，但他们使用的哈希函数是人工设计或者随机生成的；与之不同，哈希学习是希望从数据中自动地学习出哈希函数. 从哈希技术的角度来看，前者被称为数据独立方法，后者被称为数据依赖方法. 有研究表明[17,18]，与数据独立方法相比，数据依赖方法(即哈希学习方法)只需用较短的哈希编码位数就能取得理想的精度，从而进一步提高检索和学习效率，降低存储和通信开销.

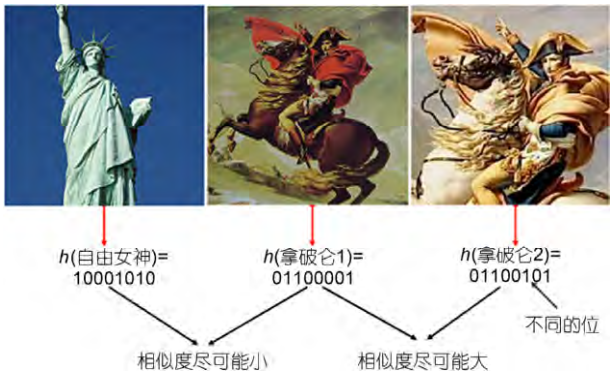

图 1 (网络版彩色)哈希学习示意图

Figure 1 (Color online)Illustration of learning to hash

1 研究进展

哈希学习由Salakhutdinov和Hinton[12,13]于2007年推介到机器学习领域，于近几年迅速发展成为机器学习领域和大数据学习领域的一个研究热点[14~22,26~37]，并广泛应用于信息检索[38,39]、数据挖掘[40,41]、模式识别[42,43]、多媒体信息处理[44,45]、计算机视觉[46,47]、推荐系统[48]、以及社交网络分析[49,50]等领域. 值得一提的是，国内学者在这方面也进行了有意义的探索[32~37,43,45~47,50,51].

由于从原空间中的特征表示直接学习得到二进制的哈希编码是一个NP难问题[14]. 现在很多的哈希学习方法[14,17~20]都采用两步学习策略：第一步，先对原空间的样本采用度量学习(metric learning)[52]进行降维，得到1个低维空间的实数向量表示；第二步，对得到的实数向量进行量化(即离散化)得到二进制哈希码. 现有的方法对第二步的处理大多很简单，即通过某个阈值函数将实数转换成二进制位. 通常使用的量化方法为1个阈值为0的符号函数，即如果向量中某个元素大于0，则该元素被量化为1，否则如果小于或等于0，则该元素被量化为0. 例如，假设样本在原空间中的特征表示为1个5维实数向量(1.1, 2.3, 1.5, 4, 3.2)，经过某种度量学习(通常把降维看成度量学习的一种)处理后得到1个三维的实数向量(1.8, –2.3, 0.6)，然后经过符号函数量化后，得到的二进制哈希码为(1, 0, 1). 一般来说，度量学习阶段首先得构建学习模型，然后对模型的参数进行优化和学习. 下面我们将从学习模型、参数优化和量化策略3方面

来介绍哈希学习的最新进展.

根据学习模型(一般指度量学习阶段的模型)是否利用样本的监督信息(例如类别标记等), 现有的哈希学习模型可以分为非监督模型[18~20]、半监督模型[17,36,47]和监督模型[26,31,42]. 非监督模型又可以进一步细分为基于图的模型[19]和不基于图的模型[18,20], 监督模型又可以进一步细分为监督信息为类别标记的模型[26,42]和监督信息为三元组或者排序信息的模型[31]. 实际上, 这每一个细分的类对应于机器学习中一个比较大的子方向, 例如基于图的模型. 由此可以看出, 现有的哈希学习模型虽然总数比较多, 但是在各个子方向上还仅仅只是进行了初步的尝试. 此外, 度量学习是机器学习领域的研究热点之一, 而度量学习方面的工作刚好可以用来实现哈希学习的第一步, 因此目前很多哈希学习模型(包括非监督、半监督和监督)只是直接利用或者简单改进已有度量学习模型, 然后采用上述的符号函数进行量化, 得到哈希编码. 经过一些摸索[20,22,29], 我们发现度量学习得到的结果通常是在模型目标函数的限制下使得信息损失最小, 因此得到的总是最优的结果; 而在将度量学习应用到哈希学习中时, 除了第一步的度量学习可能造成信息损失外, 第二步量化过程的信息损失对性能的影响也非常大, 有时候甚至超过第一步造成的信息损失, 因此, 第一步度量学习得到的最优结果并不能保证最终量化后的二进制编码为最优. 目前, 很多哈希学习方法没有将量化过程中的信息损失考虑到模型构建中去.

现有的参数优化方法大概可以分为两类. 第一类是采用与传统度量学习的优化方法类似的策略, 对所有位对应的(实数)参数一次性全部优化[14,19]. 这种策略带来的一个不利后果是没办法弥补量化过程带来的信息损失, 有可能导致的结果是随着哈希码长度的增大, 精确度反而下降. 第二类是避免一次性全部优化所有位对应的(实数)参数, 而采用按位(bitwise)优化策略[17,22,36], 让优化过程能够自动地弥补量化过程中损失的信息. 实验结果表明, 即使学习模型的目标函数相同, 采用按位优化策略能取得比一次性全部优化所有参数的策略更好的性能. 但按位优化策略对模型目标函数有一定的要求和限制, 比如目标函数可以写成残差的形式[22]. 目前, 大部分哈希学习方法还是采取一次性全部优化所有参数的策略.

哈希学习跟传统度量学习的一个很本质的区别是需要量化成二进制码. 现有的哈希学习方法大多采用很简单的量化策略, 即通过某个阈值函数将实数转换成二进制位. 最近出现一些专门研究量化策略的工作[29,39,53], 并且发现量化策略也会影响哈希学习方法的性能, 至少跟第一步的度量学习阶段同等重要. 我们在文献[29,39]中, 采用对度量学习阶段得到的每一个实数维进行多位编码的量化策略, 取得了比传统的单位编码策略更好的效果. 一般来说, 度量学习的结果中, 各维的方差(或信息量)通常各不相等[18,20]. 而现有的很多方法采用“度量学习+相同位数编码”的策略[14,19], 导致的结果是随着哈希码长度的增大, 精确度反而下降. 一种更合理的量化策略是, 采用更多的位数编码信息量更大的维. 目前, 有部分工作在这方面进行了尝试, 取得了不错的结果[53].

综上所述, 目前哈希学习的研究现状是: 已被广泛关注并在某些应用领域取得了初步成效, 但研究才刚刚开始, 有的学习场景和应用领域甚至还没有研究者进行哈希学习的尝试; 问题本质和模型构建有待于进一步深入思考; 模型参数的优化方法有待于进一步探索; 量化阶段的重要性已经引起注意, 但量化策略期待进一步突破.

2 发展趋势

目前大部分哈希学习研究的思路为: 针对某个机器学习场景(比如排序学习场景[31])或者应用场景, 只要以前没有人尝试过用哈希学习的思想来加速学习过程, 就可以考虑把哈希学习用进去, 然后在一个传统模型(这个传统模型不用哈希)解决不了的数据或者应用规模上进行实验验证. 从解决实际问题的角度来讲, 这些工作虽然初步, 但还是很有研究价值的, 毕竟为大数据中传统模型不能解决的问题提供了一种可行的解决思路. 但从哈希学习本身的研究来讲, 目前大部分工作还没有从哈希学习问题的本质上进行考虑. 我们认为以后的工作可以从理论分析、模型构建、参数优化、量化策略等几个方面进行进一步探索.

目前哈希学习理论分析方面的工作还很少. 用哈希码表示数据后, 数据相当于进行了有损压缩. 在后续的处理中, 比如检索或者挖掘过程中, 基于哈希码表示的数据得到的模型能在多大程度上逼近从原

始数据得到的模型, 即精确度如何, 目前还没有相关的理论分析. 另外, 针对一个具体问题或应用, 到底需要多少位编码才能保证结果达到一定的精确度, 目前都是根据在验证集上的实验结果来进行选择, 是否存在一些理论上的指导也非常值得研究.

针对哈希学习的量化过程会存在信息损失这一本质特征, 更好的策略是在度量学习的模型构建过程中将量化过程中可能的信息损失考虑进去. 但如果把量化过程中可能的信息损失考虑到模型的构建过程中, 量化结果的离散性将使得模型构建变得异常复杂. 因此, 如何构建考虑到量化过程信息损失的有效哈希学习模型是哈希学习研究需要解决的又一重要问题.

在参数优化过程中, 虽然按位优化策略能自动地弥补量化过程中损失的信息, 但目前大部分模型的目标函数并不适合于这种优化方式. 为其他模型设计能弥补量化过程信息损失的优化策略, 还需要进行进一步的研究. 另外, 目前的监督模型中[26,42], 对监督信息的利用大多通过构建样本对之间的关系来实现. 例如, 如果样本i和j属于同一类, 则$\boldsymbol{Y}(i, j)=1$, 否则$\boldsymbol{Y}(i, j)=0$. 然后再基于$\boldsymbol{Y}$矩阵来建模. 这种监督信息利用方式的一个后果是: 存储和计算开销都至少是训练样本数的平方级. 哈希学习研究近几年之所以这么热, 正是因为它能够被用来处理大数据. 当存在海量的训练数据, 尤其是海量有监督信息的数据时, 模型的参数训练和优化过程非常慢或者甚至不可行. 如何保证参数优化算法能快速地完成, 也是有待解决的关键问题之一.

虽然最近出现的量化策略[29,39,53]已经取得了比传统量化策略更好的性能, 但还没有很好地跟保相似性或者监督信息结合起来. 因此, 研究更优的量化策略, 以更好地保持原始空间的相似性或者跟监督信息尽可能保持一致, 也是值得进一步探索的方向.

3 总结与展望

本文对大数据哈希学习的研究进展和发展趋势进行了介绍. 可以看出, 哈希学习虽然已被广泛关注并在某些应用领域取得了初步成效, 但研究才刚刚开始, 大部分学习场景和应用领域到目前为止还只出现很少的哈希学习方法, 有的场景和应用甚至还没有研究者进行哈希学习的尝试. 例如, 推荐系统是个很大的应用方向, 但到目前为止这方面采用哈希学习的工作还不多[48]. 因此, 怎样将哈希学习的思想和方法拓展到新的学习场景和应用领域, 用来解决传统方法在遇到大数据时不能解决的问题, 将是非常有意义的工作. 特别值得一提的是, 大数据学习中的另一重要研究方向是基于多机集群的分布式机器学习[10], 而很多分布式机器学习的瓶颈在于节点间的通信开销. 因此, 将哈希学习引入到分布式机器学习算法, 并验证哈希学习在减小通信开销方面的有效性, 也是非常有意义的研究方向.

参考文献

1 Mayer-Schönberger V, Cukier K. Big Data: A Revolution That Will Transform How We Live, Work, and Think. Boston: Eamon Dolan/Houghton Mifflin Harcourt, 2013

2 Tu Z P. The Big Data Revolution (in Chinese). Guilin: Guangxi Normal University Press, 2013 [涂子沛. 大数据. 桂林: 广西师范大学出版社, 2013]

3 Hey T, Tansley S, Tolle K. The Fourth Paradigm: Data-Intensive Scientific Discovery. Redmond: Microsoft Research, 2009

4 Bryant R E. Data-intensive scalable computing for scientific applications. Comput Sci Engin, 2011, 13: 25–33

5 Zhou Z H. Machine learning and data mining (in Chinese). Commun Chin Comput Fed, 2007, 3: 35–44 [周志华. 机器学习与数据挖掘. 中国计算机学会通讯, 2007, 3: 35–44]

6 Zhou Z H, Chawla N V, Jin Y, et al. Big data opportunities and challenges: Discussions from data analytics perspectives. IEEE Comput Intell Mag, 2014, 9: 62–74

7 Jordan M. Message from the president: The era of big data. ISBA Bull, 2011, 18: 1–3

8 Kleiner A, Talwalkar A, Sarkar P, et al. The big data bootstrap. In: Proceedings of the 29th International Conference on Machine Learning (ICML), Edinburgh, 2012, 1759–1766

9 Shalev-Shwartz S, Zhang T. Accelerated proximal stochastic dual coordinate ascent for regularized loss minimization. In: Proceedings of the 31st International Conference on Machine Learning (ICML), Beijing, 2014, 64–72

10 Gonzalez J E, Low Y, Gu H, et al. PowerGraph: Distributed graph-parallel computation on natural graphs. In: Proceedings of the 10th USENIX Symposium on Operating Systems Design and Implementation (OSDI), Hollywood, 2012, 17–30
11 Gao W, Jin R, Zhu S, et al. One-pass AUC optimization. In: Proceedings of the 30th International Conference on Machine Learning (ICML), Atlanta, 2013, 906–914
12 Salakhutdinov R, Hinton G E. Semantic hashing. In: Proceedings of SIGIR Workshop on Information Retrieval and Applications of Graphical Models, Amsterdam, 2007
13 Salakhutdinov R, Hinton G E. Semantic hashing. Int J Approx Reasoning, 2009, 50: 969–978
14 Weiss Y, Torralba A, Fergus R. Spectral hashing. In: Proceedings of the 22nd Annual Conference on Neural Information Processing Systems (NIPS), Vancouver, 2008, 1753–1760
15 Kulis B, Jain P, Grauman K. Fast similarity search for learned metrics. IEEE Trans Pattern Anal Mach Intell, 2009, 31: 2143–2157
16 Weinberger K Q, Dasgupta A, Langford J, et al. Feature hashing for large scale multitask learning. In: Proceedings of the 26th International Conference on Machine Learning (ICML), Montreal, 2009, 1113–1120
17 Wang J, Kumar S, Chang S F. Semi-supervised hashing for large-scale search. IEEE Trans Pattern Anal Mach Intell, 2012, 34: 2393–2406
18 Gong Y, Lazebnik S, Gordo A, et al. Iterative quantization: A procrustean approach to learning binary codes for large-scale image retrieval. IEEE Trans Pattern Anal Mach Intell, 2013, 35: 2916–2929
19 Liu W, Wang J, Kumar S, et al. Hashing with graphs. In: Proceedings of the 28th International Conference on Machine Learning (ICML), Washington, 2011, 1–8
20 Kong W, Li W J. Isotropic hashing. In: Proceedings of the 26th Annual Conference on Neural Information Processing Systems (NIPS), Nevada, 2012, 1655–1663
21 Rastegari M, Choi J, Fakhraei S, et al. Predictable dual-view hashing. In: Proceedings of the 30th International Conference on Machine Learning (ICML), Atlanta, 2013, 1328–1336
22 Zhang D, Li W J. Large-scale supervised multimodal hashing with semantic correlation maximization. In: Proceedings of the 28th AAAI Conference on Artificial Intelligence (AAAI), Quebec, 2014, 2177–2183
23 Gionis A, Indyk P, Motwani R. Similarity search in high dimensions via hashing. In: Proceedings of the 25th International Conference on Very Large Data Bases (VLDB), Edinburgh, 1999, 518–529
24 Datar M, Immorlica N, Indyk P, et al. Locality-sensitive hashing scheme based on p-stable distributions. In: Proceedings of the 20th ACM Symposium on Computational Geometry (SOCG), New York, 2004, 253–262
25 Andoni A, Indyk P. Near-optimal hashing algorithms for approximate nearest neighbor in high dimensions. Commun ACM, 2008, 51: 117–122
26 Norouzi M, Fleet D J. Minimal loss hashing for compact binary codes. In: Proceedings of the 28th International Conference on Machine Learning (ICML), Washington, 2011, 353–360
27 Norouzi M, Fleet D J, Salakhutdinov R. Hamming distance metric learning. In: Proceedings of the 26th Annual Conference on Neural Information Processing Systems (NIPS), Nevada, 2012, 1070–1078
28 Zhen Y, Yeung D Y. Co-regularized hashing for multimodal data. In: Proceedings of the 26th Annual Conference on Neural Information Processing Systems (NIPS), Nevada, 2012, 1385–1393
29 Kong W, Li W J. Double-bit quantization for hashing. In: Proceedings of the 26th AAAI Conference on Artificial Intelligence (AAAI), Toronto, 2012, 634–640
30 Strecha C, Bronstein A M, Bronstein M M, et al. LDAhash: Improved matching with smaller descriptors. IEEE Trans Pattern Anal Mach Intell, 2012, 34: 66–78
31 Li X, Lin G, Shen C, et al. Learning hash functions using column generation. In: Proceedings of the 30th International Conference on Machine Learning (ICML), Atlanta, 2013, 142–150
32 Xu B, Bu J, Lin Y, et al. Harmonious hashing. In Proceedings of the 23rd International Joint Conference on Artificial Intelligence (IJCAI), Beijing, 2013, 1820–1826
33 Huang L K, Yang Q, Zheng W S. Online hashing. In: Proceedings of the 23rd International Joint Conference on Artificial Intelligence (IJCAI), Beijing, 2013, 1422–1428
34 Zhai D, Chang H, Zhen Y, et al. Parametric local multimodal hashing for cross-view similarity search. In: Proceedings of the 23rd International Joint Conference on Artificial Intelligence (IJCAI), Beijing, 2013, 2754–2760
35 Zhang Y M, Huang K, Geng G, et al. Fast kNN graph construction with locality sensitive hashing. In: Proceedings of the European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases (ECML/PKDD), Prague, 2013, 660–674
36 Wu C, Zhu J, Cai D, et al. Semi-supervised nonlinear hashing using bootstrap sequential projection learning. IEEE Trans Knowl Data Eng, 2013, 25: 1380–1393

37 Zhang P, Zhang W, Li W J, et al. Supervised hashing with latent factor models. In: Proceedings of the 37th ACM Conference on Research and Development in Information Retrieval (SIGIR), Queensland, 2014, 173–182
38 Zhang D, Wang F, Si L. Composite hashing with multiple information sources. In: Proceedings of the 34th ACM Conference on Research and Development in Information Retrieval (SIGIR), Beijing, 2011, 225–234
39 Kong W, Li W J, Guo M. Manhattan hashing for large-scale image retrieval. In: Proceedings of the 35th ACM Conference on Research and Development in Information Retrieval (SIGIR), Portland, 2012, 45–54
40 He J, Liu W, Chang S F. Scalable similarity search with optimized kernel hashing. In: Proceedings of the 16th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD), Washington, 2010, 1129–1138
41 Zhen Y, Yeung D Y. A probabilistic model for multimodal hash function learning. In: Proceedings of the 18th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD), Beijing, 2012, 940–948
42 Liu W, Wang J, Ji R, et al. Supervised hashing with kernels. In: Proceedings of the 25th IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Providence, 2012, 2074–2081
43 Shen F, Shen C, Shi Q, et al. Inductive hashing on manifolds. In: Proceedings of the 26th IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Portland, 2013, 1562–1569
44 Zhu X, Huang Z, Shen H T, et al. Linear cross-modal hashing for efficient multimedia search. In: Proceedings of the 21st ACM Multimedia (MM), Barcelona, 2013, 143–152
45 Wu F, Yu Z, Yang Y, et al. Sparse multi-modal hashing. IEEE Trans Multimedia, 2014, 16: 427–439
46 Xu H, Wang J, Li Z, et al. Complementary hashing for approximate nearest neighbor search. In: Proceedings of the 13rd IEEE International Conference on Computer Vision (ICCV), Barcelona, 2011, 1631–1638
47 Kan M, Xu D, Shan S, et al. Semi-supervised hashing via kernel hyperplane learning for scalable image search. IEEE Trans Circuits Syst Video Technol, 2014, 24: 704–713
48 Zhou K, Zha H. Learning binary codes for collaborative filtering. In: Proceedings of the 18th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD), Beijing, 2012, 498–506
49 Sarkar P, Chakrabarti D, Jordan M. Nonparametric link prediction in dynamic networks. In: Proceedings of the 29th International Conference on Machine Learning (ICML), Edinburgh, 2012
50 Ou M, Cui P, Wang F, et al. Comparing apples to oranges: A scalable solution with heterogeneous hashing. In: Proceedings of the 19th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD), Chicago, 2013, 230–238
51 Zhang Q, Wu Y, Ding Z, et al. Learning hash codes for efficient content reuse detection. In: Proceedings of the 35th ACM Conference on Research and Development in Information Retrieval (SIGIR), Portland, 2012, 405–414
52 Bellet A, Habrard A, Sebban M. A survey on metric learning for feature vectors and structured data. arXiv:1306.6709, 2013. http://arxiv.org/abs/1306.6709
53 Moran S, Lavrenko V, Osborne M. Variable bit quantization for LSH. In: Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (ACL), Sofia, 2013, 753–758

Learning to hash for big data: Current status and future trends

LI WuJun[1,2] & ZHOU ZhiHua[1,2]

[1] *National Key Laboratory for Novel Software Technology, Nanjing University, Nanjing 210023, China;*
[2] *Collaborative Innovation Center of Novel Software Technology and Industrialization, Nanjing 210023, China*

With the rapid development of information technology, explosion of data has occurred in most areas, which means that we have entered the era of big data. Big data has become one of the most important national strategic resources owing to its wide application in a large variety of areas. As a result, research in both academia and industry has focused greatly on big data processing, including storage, management, and analysis. Because the ultimate goal of big data processing is to mine value from big data, in which machine learning plays a key role, big data machine learning (BDML) has become one of the core directions for big data research. By representing the data as binary code, learning to hash (LH) can dramatically reduce the storage and communication cost, thereby improving the efficiency and scalability of BDML systems. Furthermore, LH can also alleviate the curse of dimensionality in BDML systems. Hence, LH has become a hot research topic in machine learning and BDML. This paper gives a brief introduction to LH.

big data, machine learning, learning to hash, big data machine learning

doi: 10.1360/N972014-00841