大数据融合实战

邓 栋[1] 姜 禹[1] 王健楠[2]等
[1]清华大学
[2]美国加州大学伯克利分校

关键词：大数据 数据融合 实体解析

背景

随着社会信息化程度的不断提高，信息系统产生了越来越多的数据。面对大规模的数据，人们开始重新思考对数据的认识，开发新的数据分析工具，这些改变标志着大数据时代的来临。2012年，李国杰等[1]对大数据研究中的科学问题进行了归纳，介绍了大数据应用与研究中面临的问题与挑战。国内其它学者也从不同角度对大数据研究进行了探讨。孟小峰等[2]对近年来产业界和学术界在大数据方面的研究工作进行了综述，重点阐述了大数据集成、分析、隐私管理、能耗、易用性等方面的技术挑战。王飞跃[3]提出了面向大数据和开源信息的科技态势解析与决策服务的理念、概念，并构建了基本的系统框架和技术体系。王珊等[4]针对适合大数据分析的数据仓库架构进行了剖析。覃雄派等[5]则对关系型数据库技术及以MapReduce为代表的非关系数据管理技术进行了详细的比较分析。

大数据具有四个维度（简称4V）[6]：volume（数据体量：大）、velocity（数据增加和变化的速度：快）、variety（数据来源和类型：繁多）、veracity（数据的真实性：难以保证）。大数据往往是由来源不同的数据混合而成的，具有不同的数据模式以及数据表示方法，如何将这些数据融合起来是大数据研究的核心问题。大数据融合在实际中有很多重要的应用，比如获取互联网上各大电子商务网站的数据用于商品的价格比较。然而同一款商品在不同网站上可能具有不同的表示形式。例如：京东商城的“iPad 2”，淘宝商城的“iPad Two”，国美电器的“Apple iPad 2”代表完全相同的商品，但是产品名称却完全不同。如果采用数据的精确匹配，将丢失掉大量匹配的商品。所以在大数据融合的研究中，一般采用数据的近似匹配。如何快速从大规模数据中找到近似匹配的数据，是一个非常具有挑战性的课题。我们实验室从2009年开始从事这一问题的研究，相关研究成果多次发表在数据库顶级会议上。近期，我们参加了扩展数据库技术国际会议(International Conference on Extending Database Technology, EDBT）组织的大数据融合竞赛并夺得了冠军，证明了我们的技术在世界上处于领先地位。

大数据融合竞赛

数据融合算法一直是大数据研究领域的热门问题，来自世界各地的计算机、生物信息等不同领域的研究人员对其都有深入的研究。然而，由于之前在不同研究中的实验环境不同，人们无法准确地评判算法间的速度差距，而且实验环境中的数据规模往往较小，人们也无法掌握在大数据时代中，这些算法的扩展性是否能够满足新的需求。因此，EDBT会议主办方希望通过竞赛的方式来促进各领域之间的交流合作。2013年3月18~22日，EDBT2013在意大利热那亚召开，共有包括来自中国、美国、澳大利亚、英国等国的11支队伍参赛。

任务介绍

本次大数据融合的任务是从两个大规模的数据集合中，融合所有相似性较高的“数据对”（相似性通过相似性函数来定义，例如

编辑距离函数），参赛者可以使用任何语言来实现自己的算法，比赛按照程序运行时间进行排名。主办方给定两个真实世界中的数据集：一个是 DNA 序列数据集，其意义在于通过整合不同的生物信息数据，可以加速生命科学的研究和发展；另一个是地理位置信息数据集，通过融合不同的地理位置信息数据能够丰富地图信息。

问题挑战

假设给定的数据集含有 n 个数据，我们面对的是要从 $O(n^2)$ 的“数据对”集合中筛选出那些包含具有“非常相似”内容的“数据对”。在大数据时代，数据来源是多样的，同一个实体的表现形式可能并不完全相同。为了对这些数据进行融合和分析，我们必须能够判断出不同来源的数据之间的对应关系，这样才能进一步进行融合。如果采用蛮力的算法对这 $O(n^2)$ 的数据对进行一一验证，则在时间上无法承受。主办方提供的两个数据集的规模都在 10^7 量级，这也就意味着我们需要在 10^{14} 个“数据对”中找到所有内容相似的组合。假设计算机每秒可以处理一百万个“数据对”，那么这种算法将需要几年的时间。这在现实中显然是无法适用的，需要设计更精巧的解决方案。

设计与实现

我们采用基于过滤—验证两个步骤的框架来提升数据融合算法的效率：首先按照一定的方式，在较少的时间内对“数据对”进行过滤，产生“候选对”，然后对“候选对”进行一一验证。这样可以避免枚举所有 $O(n^2)$ 个数据对，大大减少算法所花费的时间。我们实验室之前已经进行了深入的探索工作，提出了键树连接 (trie-join)[7] 和传递连接 (pass-join)[9] 两种各具特色、世界领先的算法。trie-join 利用 trie 树来筛选结果，适合于短字符串集合；pass-join 是基于“特征”抽取的，如果两个字符串相似，那么必然会有重合的“特征”，因而适合于各种长度的字符串集合。为了选取合适的方法，我们首先利用主办方给定的两个规模仅为最终规模 5% 的抽样测试数据进行测试分析，最终选用了 pass-join。

接下来要考虑如何针对给定数据集进行优化。我们采用了 3 种优化策略。

“一对多”快速验证算法 目前基于过滤—验证的算法主要优化过滤阶段，缺少对验证阶段的优化。现有的验证算法只能逐一评判每对数据是否满足给定的相似性标准，导致计算过程较慢。那么有没有什么技术可以加速验证的过程？我们提出利用过滤的思想加快验证的过程，通过视角的转换，利用过滤算法的特殊性质来设计一对多的验证算法，快速过滤第一步遗留下来的“漏网之鱼”。

内容过滤机制 我们提出了“内容过滤”技术，该技术的侧重点与 pass-join 完全不同。pass-join 利用字符的相对位置进行过滤，不考虑具体的字符内容，而“内容过滤”则从字符内容角度进行过滤，可以很好地和 pass-join 进行互补。

硬件优化技术 为了使程序能够最大程度地发挥计算机硬件的优势，我们又对程序进行了挖掘。充分利用了现代计算机架构上的两个重要特性：多核多线程，单指令多数据 (single instruction multiple data, SIMD) 指令集。这两个特性获得了主流 CPU 的广泛支持，可以大大提高程序的运行速度。多核多线程可以直接应用在我们的算法上，使算法随着核数的提高具有线性的加速比；而 SIMD 指令集需要开发人员直接编写类似汇编的语句，我们通过该方法优化程序关键部分的性能，取得了良好的效果。

我们设计了基于“过滤—验证”的算法框架，并融入了快速验证机制、内容过滤技术、多核多线程技术、基于 SIMD 的汇编优化，最终设计出大数据融合算法。

比赛结果

图 1 是部分比赛结果，其中 1_A 是我们参赛队伍的程序。参数 k 用于控制输出结果的相似性，该参数越大，结果数量越多，计算量也就越大。从数据看，我们的程序在所有不同的参数设置下，均优于其他队伍的程序，在绝大多数参数设置下，我们更是在速度上以 10

[1] 一种基于划分的算法。

倍左右的优势胜过其他队伍。

研究展望

本次比赛我们采用了单机多核多线程算法来解决大数据融合问题。随着数据量的爆炸式增长，单台机器的力量已经不足以处理现实世界中的海量数据了。MapReduce 框架给了我们另外一条解决大数据融合问题的思路。MapReduce 是由谷歌公司提出并实现的一个分布式计算框架，它可以提供上千台计算机集群的高容错性并行计算。基于 MapReduce 框架的大数据融合算法的设计难点主要体现在数据倾斜以及算法可扩展性上。我们目前正在开展这方面的研究，通过 MapReduce 来提高大数据融合的可扩展性。为了避免 MapReduce 框架中传输大量数据，我们对传输的数据进行分组和聚集，从而大大减少了数据传输量，显著提高了算法的效率。目前我们正在研究如何使数据的传输量尽可能减少，从理论上保证算法最优。

线程数	程序	k=0	k=1	k=2	k=3	k=4
8	1_A	0.06	0.30	0.53	1.83	8.12
	4_A	10.40	10.35	11.15	17.06	46.00
	5_B	1.45	4.17	56.12	376.39	2513.64
24	1_A	0.08	0.27	0.38	0.94	3.14
	4_A	10.42	10.37	10.69	12.60	22.76
	5_B	5.76	4.07	65.28	760.71	2353.97
80	1_A	0.11	0.31	0.39	0.85	2.42
	4_A	10.47	10.46	10.48	11.37	16.76
	5_B	2.38	3.92	42.47	532.91	2051.15

图1　在地理位置信息数据集中前三名程序的性能比较（单位：秒）

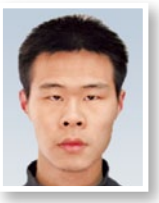
大数据融合研究中的另外一个问题是融合方法的准确率。在现实中，单纯依靠机器算法很难达到令人满意的融合结果，这主要是由于判断一组“数据对”是否可以被融合往往需要对自然语言的理解，而目前的机器算法还很难精准地剖析出文字背后所蕴含的语义。例如：

第一组：“iPad 2nd generation”和“iPad 3rd generation”

第二组：“iPad 2nd generation”和“Apple iPad Two”

从字面上看，第一组“数据对”看起来非常相似，但却代表苹果公司两款不同的产品；第二组数据虽然字面上看起来很不相似，但是通过对自然语言的理解，可以断定它们代表同一款产品。如果只是依靠文本相似性来决定是否进行数据融合，会得到完全错误的结果。

为了解决这一问题，我们目前开始研究基于众包的大数据融合算法，即利用群体智慧的力量来提高大数据融合算法结果的准确率。与机器算法相比，众包技术虽然可以获得更高质量的融合结果，但是开销变大了，等待时间也变长了。所以如何设计出有效的解决办法来更好地均衡结果质量、人工开销以及完成时间，是基于众包的大数据融合算法需要研究的主要问题。目前我们已在这一方向上取得了一些阶段性的研究成果，提出了一种混合机器和众包的数据融合框架[11,12]。实验表明，这项技术可以通过花费很少的费用来大幅度提升数据融合的准确率。今后，我们计划更深层地融合机器算法和众包技术，即：利用众包技术帮助机器学习算法训练更好的模型，从而进一步节省费用、提高质量。■



邓　栋
清华大学计硕士研究生。主要研究方向为数据质量以及数据融合。dd11@mails.tsinghua.edu.cn

姜　禹
清华大学硕士研究生。主要研究方向为数据融合。y-jiang12@mails.tsinghua.edu.cn

王健楠
美国加州大学伯克利分校博士后。主要研究方向为大数据融合与众包数据库。wjn08@mails.tsinghua.edu.cn

李国良　冯建华

参考文献

[1] 李国杰,程学旗.大数据研究.中国科学院院刊，2012,27(6):647~657

[2] 孟小峰，慈祥．大数据管理:概

念、技术与挑战. 计算机研究与发展, 2013, 50(1): 146~169

[3] 王飞跃. 知识产生方式和科技决策支撑的重大变革—面向大数据和开源信息的科技态势解析与决策服务. 中国科学院院刊, 2012, 27(5): 527~537

[4] 王珊, 王会举, 覃雄派, 周烜. 架构大数据:挑战、现状与展望. 计算机学报 34(10): 1741~1752

[5] 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. 软件学报, 2012, 23(1): 32~45

[6] IBM: What is big data? [2013-04-12]. http://www.ibm.com/software/data/bigdata/

[7] Jiannan Wang, Jianhua Feng, and Guoliang Li. Trie-Join: efficient trie-based string similarity joins with edit-distance constraint. VLDB 3(1):1219~1230, 2010

[8] JiannanWang, Guoliang Li, and Jianhua Feng. Fast-Join: an efficient method for fuzzy token matching based string similarity join. ICDE 2011:458~469

[9] Guoliang Li, Dong Deng, JiannanWang, and Jianhua Feng. Pass-Join: a partition-based method for similarity joins. VLDB 5(3):253~264 (2011)

[10]JiannanWang, Guoliang Li, and Jianhua Feng. Can we beat the prefix filtering? An adaptive framework for similarity join and search. SIGMOD 2012: 458~469

[11]Jiannan Wang, Tim Kraska, Michael J. Franklin, and Jianhua Feng. CrowdER: crowdsourcing entity resolution. VLDB 5(11):1483~1494, 2012

[12]Jiannan Wang, Guoliang Li, Tim Kraska, Michael J. Franklin, and Jianhua Feng. Leveraging transitive relations for crowdsourced joins. SIGMOD 2013

[13]Dong Deng, Yu Jiang, Guoliang Li, Jian Li, Cong Yu. Scalable column concept determination for Web tables using large knowledge bases. VLDB 2013