软件学报 ISSN 1000-9825, CODEN RUXUEW
Journal of Software,2013,24(Suppl.(2)):188−199
©中国科学院软件研究所版权所有.
E-mail: jos@iscas.ac.cn
http://www.jos.org.cn
Tel/Fax: +86-10-62562563

概率数据上基于 EMD 距离的并行 Top-*k* 相似性连接算法*

雷 斌, 许 嘉, 谷 峪, 于 戈

(东北大学 信息科学与工程学院 软件与理论研究所,沈阳 110819)

通讯作者: 于戈, E-mail: yuge@mail.neu.edu.cn

摘 要: 以无线传感器网络为代表的新型数据应用和以图像处理为基础的传统数据应用都产生了大规模的概率数据.在概率数据的管理中,Top-*k* 相似性连接操作返回最相似的 *k* 对概率数据,具有重要应用价值.直方图是最常用的概率数据模型之一,而 EMD(Earth Mover's Distance)距离因其较强的鲁棒性可更准确地量化直方图概率数据之间的相似性.然而 EMD 距离的计算却具有三次方的时间复杂度,给基于 EMD 距离的 Top-*k* 相似性连接带来巨大挑战.基于流行的 MapReduce 并行处理框架,利用 EMD 距离对偶线性规划问题的优良特性,提出了两种大规模概率数据上基于 EMD 距离的 Top-*k* 相似性连接算法.首先提出基于块嵌套循环连接思想的基本解决方法,命名为 Top-*k* BNLJ 算法.进而改进数据划分策略,提出基于数据局部性进行数据划分的 Top-*k* DLPJ 算法,有效降低了 MapReduce 作业执行过程中的数据传输量.使用大规模真实数据集对两种算法进行评估,证实了本文提出的 Top-*k* DLPJ 算法的高效性和处理大规模数据集时的良好扩展性.

关键词: 概率数据;EMD 距离;并行 Top-*k* 相似性连接;MapReduce 框架;对偶理论

中文引用格式: 雷斌,许嘉,谷峪,于戈.概率数据上基于 EMD 距离的并行 Top-*k* 相似性连接算法.软件学报,2013,24(Suppl.(2)):188−199. http://www.jos.org.cn/1000-9825/13036.htm

英文引用格式: Lei B, Xu J, Gu Y, Yu G. Parallel Top-*k* similarity join algorithm on probabilistic data based on earth mover's distance. Ruan Jian Xue Bao/Journal of Software, 2013,24(Suppl.(2)):188−199 (in Chinese). http://www.jos.org.cn/1000-9825/13036.htm

Parallel Top-*k* Similarity Join Algorithm on Probabilistic Data Based on Earth Mover's Distance

LEI Bin, XU Jia, GU Yu, YU Ge

(Institute of Computer Software and Theory, Information Science and Engineering College, Northeastern University, Shenyang 110819, China)

Corresponding author: YU Ge, E-mail: yuge@mail.neu.edu.cn

Abstract: Many new data applications, such as wireless sensor networks, and traditional data applications that process images produce massive probabilistic data. In probabilistic data management, the Top-*k* similarity join operation returns the most similar *k* pairs of probabilistic data and thus has important value. Histogram is one of the most frequently-used data models for representing probabilistic data. Earth Mover's Distance (EMD) is more robust in quantifying the similarities between histogram-represented probabilistic data. However, EMD computation has a cubic time complexity, which brings great challenges to the EMD-based Top-*k* similarity joins. Based on the MapReduce framework, this paper utilizes the good properties of EMD's dual problem and proposes two EMD-based Top-*k* similarity join algorithms on massive probabilistic dataset. A baseline solution named Top-*k* BNLJ algorithm is first developed on the idea of block-nested-loop join, and the novel Top-*k* DLPJ algorithm is then built on the improved idea of data locality preserving data partition strategy which significantly reduces the amount of data transferred during MapReduce jobs. Extensive experiments on large

* 基金项目: 国家重点基础研究发展计划(973)(2012CB316201); 国家自然科学基金(61272179, 61033007); 教育部博士点基金(20120042110028); 教育部-英特尔信息技术专项科研基金(MOE-INTEL-2012-06); 中央高校基本科研业务费专项资金课题(N100704001, N110404006)

收稿时间: 2013-03-15; 定稿时间: 2013-07-11

© 中国科学院软件研究所 http://www.jos.org.cn

real-world datasets, with millions of probabilistic data, have verified the efficiency, effectiveness and scalability of Top-k DLPJ algorithm.
Key words: probabilistic data, earth mover's distance, parallel top-k similarity join, MapReduce, primal-dual theory

许多新型的数据应用,例如地理跟踪[1]、信息抽取[2]和无线传感器网络[3],由于测量误差、环境噪声以及网络延时等原因产生了大量的不精确数据.不精确数据通常用概率分布进行表示,因此通常被称为概率数据.同时,计算机视觉(computer vision)领域用图像视觉特征(例如灰度、颜色和纹理等)的概率分布来标识一幅图像,也产生了大量概率数据[4].在表示概率数据的多种数学模型中,直方图(histogram)因为提取过程简单,所表示的视觉特征分布不受图像旋转和平移变化的影响等原因被广泛应用于概率数据的建模,成为使用最为广泛的概率数据模型之一.例如概率数据管理系统 ProbView[5]中提出的 Probabilistic Tuple 模型,以及概率数据管理项目 Trio[6]中定义的 x-Tuple 模型都是基于直方图表示的概率数据模型.在概率数据的管理中,Top-k 相似性连接返回概率数据集中最相似的 k 个概率数据对,是非常重要的操作.由于 Top-k 相似性连接不需要用户事先提供搜索的相似性阈值,简化了用户操作,被广泛应用于概率数据的集成以及潜在副本检测等分析和挖掘工作中[7].在度量直方图概率数据之间的相似性时,传统的 L_p 范式距离($L_p(x,y)=\sqrt[p]{\sum_i (x_i-y_i)^p}$)仅量化了直方图对应数据桶之间的取值差异,被称为"对应桶距离函数".而 EMD 距离(earth mover's distance)不仅考虑了直方图对应数据桶之间的差异性,还考虑了相邻数据桶之间的差异性,被称为"交叉桶距离函数".由于兼顾考虑了相邻数据桶取值的相似性,EMD 距离对概率分布之间的微小偏移不敏感,比"对应桶距离函数"具有更强的鲁棒性,被广泛应用在基于内容的图像检索[8]以及数据库隐私保护[9]等重要领域.

基于 EMD 距离的 Top-k 相似性连接是一个颇具挑战的研究问题.一方面,Top-k 相似性搜索和数据库连接都是非常耗时的数据库操作.另一方面,EMD 距离的计算函数具有 $O(n^3 \log n)$的高时间复杂度(n 表示直方图所包含的数据桶数目).再加上现在越来越多的实际应用需要处理远超过单机内存容量的大规模概率数据,这使得基于 EMD 距离的 Top-k 相似性连接面临巨大困难.目前,已有研究者提出了多种基于 EMD 距离的相似性搜索算法[10−13].Ira 等人基于 EMD 距离的下界函数构建索引,避免了不必要的 EMD 求精计算[10],但因该索引是基于 EMD 距离的下界函数构建的,基于该索引过滤得到的结果候选集并不完备.Marc 等人提出了一系列有效的 EMD 距离下界函数,并基于这些下界函数利用"扫描-求精"的求解框架过滤无关数据记录[11].因缺乏索引机制而需要多次扫描数据集,该策略引入了较高的磁盘 I/O 代价.到目前为止,文献[12,13]为基于 EMD 距离的相似性搜索提供了最为有效的索引解决方案,在处理效率上显著优于 Ira 等人和 Marc 等人提出的方法.在文献[12]中,Brian 等人将概率数据进行投影处理,用正态分布拟合表示投影后的概率数据,然后将正态分布数据进行霍夫变换,并用四叉树索引变换空间.该索引的过滤性能极度依赖于投影向量的选择,且确定投影向量需要事先针对数据集进行耗时的主成分分析.在文献[13]中,我们基于线性规划的对偶理论[16],利用 EMD 距离对偶问题的可行解将直方图概率数据映射至一维实数空间.并使用经典的 B^+树结构索引直方图在一维空间内的映射点.通过将基于 EMD 距离的相似性查询转换为一维映射空间内的范围查询有效缩减了查询空间,提高了相似性搜索的效率.但上述研究工作都只关注单机环境下的处理,并不能有效处理大规模的概率数据.文献[7]中提出了一种利用 MapReduce 并行处理框架解决基于 L_P 范式距离的相似性连接算法,但 L_P 范式距离与 EMD 距离之间的差异性决定了无法复用文献[7]中的方案来解决本文讨论的问题.现有的研究工作都未能有效解决大规模概率数据上基于 EMD 距离的 Top-k 相似性连接问题.

为了有效处理大规模数据,谷歌公司提出了 MapReduce 并行处理框架[14].使用该框架可轻松完成处理逻辑在大规模无共享处理集群上的部署,实现大规模数据的分布式并行处理.Apache Hadoop[15]是 MapReduce 框架的成熟开源实现,现已受到工业和学术界的共同关注.一个 MapReduce 作业主要包括 3 个处理阶段,即 Map(映射),Shuffle(数据洗牌)和 Reduce(化简).Map 阶段和 Reduce 阶段分别由多个 Map 任务和 Reduce 任务组成.每个 Map 任务处理数据集的一个文件分片,并将文件分片中的所有数据记录转换成(*key*,*value*)对提交给 map 函数.map 函数处理数据后生成新的(*key*,*value*)对作为中间数据.Shuffle 阶段随即整合这些中间数据,并将它们通过网络传输给不同的 Reduce 任务.具有相同 *key* 值的(*key*,*value*)对将被传输给相同的 Reduce 任务.Reduce 任务

© 中国科学院软件研究所 http://www.jos.org.cn

将具有相同 *key* 值的(*key*,*value*)对合并成(*key*,*List*(*value*))形式并提交给 reduce 函数进行处理.reduce 函数随后将处理得到的新(*key*,*value*)对序列存储到 DFS 上.

本文讨论了如何利用 MapReduce 框架解决大规模概率数据上基于 EMD 距离的 Top-*k* 相似性连接问题.我们之前的工作[13]利用 EMD 距离对偶问题的可行解将直方图概率数据映射至一维实数空间,并使用经典的 B^+ 树结构索引一维空间内的映射点,进而有效提高了相似性搜索的效率.本文在文献[13]提出的基于 EMD 距离的相似性搜索技术的基础上,基于 MapReduce 框架提出了两种大规模概率数据上基于 EMD 距离的并行 Top-*k* 相似性连接算法.首先提出了基于块嵌套循环连接(block nested loop)思想的基本算法 Top-*k* BNLJ(block nested loop join)算法.Top-*k* BNLJ 算法将数据集等分成 *m* 块,每个 Reduce 任务对其中任意两块数据进行连接处理,通过构建 B^+树索引加速连接处理.这种算法需要在 Map 任务和 Reduce 任务之间传输平方级(m^2 个块)的数据量,在处理大数据时,算法的扩展性受到了限制.为了降低算法执行过程中的数据传输量,本文进一步改进了数据的划分策略,利用直方图在一维映射空间内具有良好数据局部性的特点,即相似直方图在一维空间内的映射点也相似[13],将直方图数据集划分成 *m* 个分块,并能够保证任何一个可能的 Top-*k* 相似性连接的候选直方图数据对都会处在其中一个分块中(而不是跨越两个分块).因此,每个 Reduce 任务仅需对一个数据块进行自连接处理.本文将这种基于数据局部性进行数据划分的 Top-*k* 相似性连接算法称为 Top-*k* DLPJ(data locality preserved join)算法.容易理解,Top-*k* DLPJ 算法在 Map 任务和 Reduce 任务之间仅需传输线性级(约 *m* 个块)的数据量,因而在处理大数据集时具有良好的扩展性.

本文的主要贡献归纳如下:

(1) 形式化定义了基于 EMD 距离的 Top-*k* 相似性连接,丰富了 Top-*k* 相似性连接的种类.

(2) 首次讨论了大规模概率数据上基于 EMD 距离的 Top-*k* 相似性连接问题,并基于流行的 MapReduce 并行处理框架提出了两种基于 EMD 距离的相似性连接算法,包括基本的 Top-*k* BNLJ 算法和高效的 Top-*k* DLPJ 算法.

(3) 使用大规模真实数据集对本文提出的算法进行实验验证.实验表明与 Top-*k* BNLJ 算法相比,本文提出的 Top-*k* DLPJ 算法有效提高了 Top-*k* 相似性连接的执行效率,并具有良好的扩展性.

本文第 1 节介绍基本概念,给出 EMD 距离和基于 EMD 距离的 Top-*k* 相似性连接的形式化定义.第 2 节描述基于块嵌套循环的并行 Top-*k* 相似性连接,给出 Top-*k* BNLJ 算法.第 3 节讨论基于数据局部性进行数据划分的并行 Top-*k* 相似性连接,给出 Top-*k* DLPJ 算法.第 4 节给出实验结果和分析.最后,第 5 节总结全文.

1 基本概念

1.1 问题定义

定义 1. EMD 距离.

已知包含 *n* 个数据桶的直方图概率数据 $P=\{p[1],\ldots,p[n]\}$和 $Q=\{q[1],\ldots,q[n]\}$,以及地面距离矩阵 $D=[d_{ij}]$,则 *P* 和 *Q* 之间的 EMD 距离,表示为 *EMD*(*P*,*Q*),是以下线性规划问题的最优解.

$$\begin{aligned} Minimize: \quad & \sum_{i=1}^{n}\sum_{j=1}^{n} f_{ij}d_{ij} \\ \text{s.t.} \quad & \forall i: \sum_j f_{ij} = p[i] \\ & \forall j: \sum_i f_{ij} = q[j] \\ & \forall i,j: f_{ij} \geqslant 0 \end{aligned} \tag{1}$$

地面距离 $d_{ij}\in D$ 表示直方图第 *i* 个数据桶到第 *j* 个数据桶的搬运距离.以上线性规划问题包含 n^2 个变量,记为 $F=\{f_{ij}\}$,$f_{ij}\in F$ 表示从直方图 *P* 的第 *i* 个数据桶搬运至直方图 *Q* 的第 *j* 个数据桶的概率值.不难理解 EMD 距离表示将直方图 *P* 转换成直方图 *Q* 的最小搬运代价.EMD 距离的求解具有高达 $o(n^3\log n)$的时间复杂度.本文假设地面移动距离是一种距离测度(metric distance),例如常见的 Manhattan 距离和 Euclidean 距离,因此 EMD 距离也满足距离测度的特性[4],即距离的非负性、对称性和三角不等性.

© 中国科学院软件研究所 http://www.jos.org.cn

定义 2. 基于 EMD 距离的 Top-k 相似性连接.

给定包含 m 个直方图概率数据的概率数据集 $D=\{P_1,\ldots,P_m\}$,其中每个直方图概率数据都包含 n 个数据桶,即 $P_i=\{p_i[1],\ldots,p_i[n]\}$,且 $p_i[j]\geqslant 0$ 和 $\sum_j p_i[j]=1$,则 D 上的 Top-k 相似性连接返回 D 上最相似 k 个直方图概率数据对,表示为 $S_{\mathrm{Top}k}=\{(P_{i1},P_{j1}),\ldots,(P_{ik},P_{jk})\}$,且满足:

- 对于直方图概率数据 $P_{il},P_{jl}\in D$,若概率数据对$(P_{il},P_{jl})\in S_{\mathrm{Top}k}$,则为了避免出现重复概率数据对,限定 $il<jl$;
- $EMD(P_{il},P_{jl})\leqslant\ldots\leqslant EMD(P_{ik},P_{jk})$;
- 对于直方图概率数据 $P_{il},P_{jl}\in D$ 且 $il<jl$,若概率数据对$(P_{il},P_{jl})\notin S_{\mathrm{Top}k}$,则一定满足 $EMD(P_{ik},P_{jk})\leqslant EMD(P_{il},P_{jl})$.

1.2 构建面向EMD距离的高效索引

根据线性规划中的对偶理论,每一个以目标函数最小化为优化目标的线性规划问题(设为原问题)都有且仅有一个以目标函数最大化为优化目标的对偶线性规划问题.求出对偶问题的解时,也给出了原问题的解.不难得出,求解 EMD 距离的线性规划问题(公式 1)的对偶问题为:

$$\begin{aligned}\text{Maximize}: \quad & \sum_{i=1}^{n}\phi_i\cdot p[i]+\sum_{j=1}^{n}\pi_j\cdot q[j]\\ \text{s.t.} \quad & \forall i,j:\phi_i+\pi_j\leq d_{ij}\\ & \forall i:\phi_i\in\mathbb{R}\\ & \forall j:\pi_j\in\mathbb{R}\end{aligned} \tag{2}$$

EMD 距离的对偶问题拥有 $2n$ 个变量 $\{\phi_1,\ldots,\phi_n\}$ 和 $\{\pi_1,\ldots,\pi_n\}$.依据对偶理论,公式 2 中目标函数的最大值即为 $EMD(P,Q)$的值.下面介绍索引构建过程,构建索引依赖于 EMD 距离对偶问题的可行解.

线性规划问题的可行解需要满足所有约束条件,但不一定使目标函数的取值达到最优.一个线性规划问题可以拥有任意多组可行解.现设 $\Phi=\{\phi_i,\pi_j\}$ 为 EMD 距离对偶线性规划问题的一组可行解,因此对于任意一对概率数据 P 和 Q 满足:

$$\sum_{i=1}^{n}\phi_i p[i]+\sum_{j=1}^{n}\pi_j q[j]\leqslant EMD(P,Q) \tag{3}$$

公式(3)表明,根据 EMD 距离对偶问题的一组可行解 Φ,可以计算任意一对概率数据之间 EMD 距离的下界值.同时由 EMD 距离对偶问题的限制条件(公式(2))可知,Φ 的求取仅依赖于地面距离矩阵$[d_{ij}]$,并不依赖任何概率数据.因而可以基于可行解 Φ 为直方图概率数据集 D 构建与数据无关的索引.

已知 EMD 距离对偶线性规划问题的一组可行解 $\Phi=\{\phi_i,\pi_j\}$,现定义任一概率数据 P 的键值为 $bkey(P,\Phi)=\sum_i\phi_i.p[i]$.通过计算直方图概率数据集 D 中的每一个概率数据的键值(每次键值计算仅耗费 $o(n)$ 时间),即可实现从 D 到一维实数空间的映射.从而可以使用经典的 B^+树结构索引 D 在一维实数空间的各个映射点.构建索引的目的在于过滤无关查询对象.基于公式(3)的结论,文献[13]已证明以概率数据 Q 为查询对象 τ 为相似性阈值的范围查询的查询候选概率数据 P(即满足 $EMD(P,Q)\leqslant\tau$),其键值 $bkey(P,\Phi)$ 必然落在以下键值区间内:

$$[\min_i(\phi_i+\pi_i)+bkey(Q,\Phi)-\tau,\tau-bckey(Q,\Phi)] \tag{4}$$

利用公式(4)可以将基于 EMD 距离的相似性范围查询成功转换为对一维键值空间的范围查询.此处 $bkey(Q,\Phi)$定义为概率数据 Q 的对应键值,其计算公式为 $bkey(Q,\Phi)=\sum_j\pi_j q[j]$.基于 EMD 距离对偶线性规划问题的一组可行解可以构建一棵 B^+树.容易理解,基于不同可行解构建出的 B^+树用自己的方式索引概率数据集中的每个概率数据.因此可以基于 L 组可行解构建 L 棵 B^+树,利用不同 B^+树返回的查询结果候选集之间的差异性来协同过滤更多无关数据.文献[13]中提出了两种有效的启发式策略用于选择优良的 EMD 对偶问题的可行解,一是基于聚类的策略,二是基于随机采样的策略.本文选用基于随机采样的策略生成 EMD 距离对偶问题的可行解,因为实验证实其更简单有效,只需要随机选取两个直方图,就可在计算它们之间 EMD 距离的过程中获得

© 中国科学院软件研究所 http://www.jos.org.cn

EMD 距离对偶问题的一组可行解.

2 基于块嵌套循环的并行 Top-*k* 相似性连接

在 MapReduce 框架下实现基于 EMD 距离的 Top-*k* 相似性连接的一种基本方法就是使用传统的块嵌套循环连接的思想(block-nested-loop-join),在此命名为 Top-*k* BNLJ 算法.Top-*k* BNLJ 算法在 Map 阶段通过线性扫描将概率数据集 D 等分成 m 个数据块.经过 Map 阶段,任意两个数据块组成一组(包括块与其自身组成的分组),并交由一个 Reduce 任务对这两个数据块进行块嵌套循环 Top-*k* 相似性连接.每个 Reduce 任务只能找出局部的 S_{Topk},所以还需要启动一次 MapReduce 作业从 $m(m+1)/2$ 个局部 $S_{\text{Top}k}$ 中找出全局 $S_{\text{Top}k}$.同时为了加速 Top-*k* 相似性连接,在执行连接算法之前,需要启动一次 MapReduce 作业基于样本数据集计算出 Top-*k* 相似性阈值(即 $S_{\text{Top}k}$ 中第 k 近的数据对之间的 EMD 距离)的上界值 τ.τ 将作为第 2 次 MapReduce 作业中 Reduce 任务执行块嵌套循环连接的初始 Top-*k* 相似性阈值.

因此,Top-*k* BNLJ 算法共包含 3 次 MapReduce 作业.第 1 次 MapReduce 作业抽样计算 τ 值;第 2 次 MapReduce 作业查找局部的 $S_{\text{Top}k}$;第 3 次 MapReduce 作业从局部 $S_{\text{Top}k}$ 中查找全局 $S_{\text{Top}k}$.由于第 3 次 MapReduce 作业逻辑简单,下面不再深入介绍,仅讨论 Top-*k* BNLJ 算法的第 1 个和第 2 个 MapReduce 作业的算法流程.

2.1 抽样确定τ值

在第 2 个 MapReduce 作业中,每个 Reduce 任务对其获得的数据块分组进行 Top-*k* 相似性连接,并得到局部 $S_{\text{Top}k}$.具体做法是以一个数据块为查询集合,并基于一个事先获得的 Top-*k* 相似性阈值的上界值 τ 在另一个数据块上以嵌套循环的方式进行范围查询.并基于范围查询的结果集进一步缩减 τ 值,以达到不断缩减范围查询搜索空间的目的.由此可见 τ 值的优劣直接决定第 2 个 MapReduce 作业中 Reduce 任务的执行效率,因为其范围查询的初始搜索阈值为 τ.

为了获得一个较小的 τ 值,Top-*k* BNLJ 算法的第 1 个 MapReduce 作业在 Map 阶段让每个 Map 任务以概率 p 对输入数据分片抽样,得到样本直方图集合 S.之后基于同一组 EMD 距离对偶问题的可行解 Φ 计算 S 中各个样本直方图的 *bkey* 值,并将样本直方图按照 *bkey* 值排序.计算每个样本直方图与 *bkey* 值大于自己且离自己最近的样本直方图之间的 EMD 距离,并将所获得的 $|S|-1$ 个 EMD 距离发送给 Reduce 任务,由 Reduce 任务找出所有 EMD 距离中第 k 小的距离值并赋值给 τ.为了避免重复,这里仅计算每个样本直方图与 *bkey* 值大于自己且离自己最近的样本直方图之间的 EMD 距离.根据公式(4)易理解,EMD 距离小的点对,它们的 *bkey* 值之间的差异性也越小.因而上述采样求解过程可快速缩减 τ 的取值.

2.2 查找局部$S_{\text{Top}k}$

在 Map 阶段,每个 Map 任务将输入文件分片中的每个直方图随机地分到一个数据分块中(共有 m 个数据分块).为了保证任意两个数据分块之间都能组队,需要为每个直方图产生 m 个副本并分别发送给 m 个不同的 Reduce 任务,并为来自不同分块的直方图添加不同的标志,以标识其所属的数据分块.公式(5)给出了将一个直方图的 m 个副本分配给 m 个不同的 Reduce 任务的分配逻辑.公式(6)则确定一个直方图的不同副本的标志取值.

$$R_t=\begin{cases}(m_i-1)m_i/2+t, & \text{if } t\in\{1,2,...,m_i\}\\ (t-1)t/2+m_i, & \text{if } t\in\{m_i+1,...,m\}\end{cases} \tag{5}$$

$$T_t=\begin{cases}Left, & \text{if } t\in\{1,2,...,m_i\}\\ Right, & \text{if } t\in\{m_i+1,...,m\}\end{cases} \tag{6}$$

公式(5)和公式(6)中,m_i 表示直方图所属的数据分块号,t 表示直方图的第 t 个副本,R_t 表示第 t 个副本所发送给的 Reduce 任务号,而 T_t 则表示第 t 个副本的标志取值.注意,当 Reduce 任务处理同一数据块间的 Top-*k* 相似性自连接时,Reduce 任务的所有输入数据只有 *Left* 标志;当 Reduce 任务处理两个不同数据块间的 Top-*k* 相似性连接时,则所属分块号较小的数据拥有 *Right* 标志,所属分块号较大的数据拥有 *Left* 标志.

© 中国科学院软件研究所 http://www.jos.org.cn

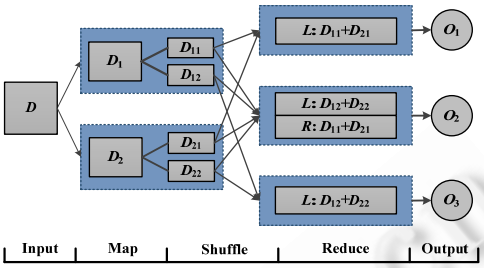


图 1 Top-k BNLJ 算法中第 2 个 MapReduce 作业的数据流图

如图 1 所示,Map 阶段共启动两个 Map 任务,每个 Map 任务将输入文件分片 D_i 划分成 D_{i1} 和 D_{i2} 两部分.由来自于不同 Map 任务的所有 D_{i1} 组成分块 1,所有 D_{i2} 组成分块 2.因而经过 Map 阶段 D 被划分成 2 个分块(即 m=2),需要在 Reduce 阶段启动 $(2\times3)/2=3$ 个 Reduce 任务来分别处理 3 个数据块分组的 Top-k 相似性连接.图中符号 L(*left*)表示左分块,R(*right*)表示右分块,用于区分一个数据块分组中两个不同的数据分块.做自连接的 Reduce 任务中只包含具有 *Left* 标志的数据;在处理两个分块连接的 Reduce 任务中,分块 2 中的数据拥有 *Left* 标志,分块 1 中的数据拥有 *Right* 标志.

在 Reduce 阶段,每个 Reduce 任务首先将输入直方图数据根据 *Left* 和 *Right* 标志分成左数据块和右数据块.其次基于本文第 1 节所述的面向 EMD 距离的索引构建技术在左数据块上构建 L 棵 B^+树索引.然后以右数据块中的每个直方图作为查询点(若 Reduce 任务只有左数据块则以左数据块中的每个直方图作为查询点),以 Top-k 相似性阈值的上界值 τ 为初始相似性阈值,在左数据块的每棵 B^+树的键值空间内做范围查询(查询范围由公式 4 决定),并根据查询结果不断更新 τ 值.由于在左数据块上构建了 L 棵 B^+树索引,因而 L 棵 B^+树会为同一查询对象返回 L 组结果候选集,对各棵树返回的结果候选集取交集即得到约简的结果候选集.为了进一步约简查询候选集的大小,在使用 L 棵 B^+树进行过滤之后,算法进而采用 EMD 距离下界过滤函数 LB_{IM}[13]对结果候选集做进一步过滤,得到最终结果候选集.算法此时需要计算查询对象与候选集里每个直方图之间的 EMD 距离,并根据 EMD 距离计算结果更新 τ 和 S_{Topk}.Reduce 任务最终输出的是局部 S_{Topk} 及相似的概率数据对之间的 EMD 距离,输出数据的形式为$\langle(h_i,h_j),EMD(h_i,h_j)\rangle$.

下面给出定理 1,用于优化数据块自身的 Top-k 相似性自连接.

定理 1. 在执行数据块的 Top-k 相似性自连接时,若 τ 为当前的 Top-k 相似性阈值,则查询对象 Q 在该数据块的 B^+树索引上搜索的键值空间仅为

$$[bkey(Q,\Phi),\tau-bckey(Q,\Phi)] \tag{7}$$

证明:在 B^+树索引中,所有直方图在叶节点上按照 $bkey$ 值从小到大排列.设有直方图 Q_1 和 Q_2,满足 $bkey(Q_1,\Phi)<bkey(Q_2,\Phi)$.

① 若 Q_1 先于 Q_2 做范围查询,则 Q_1 的查询阈值 τ_1 不小于 Q_2 的查询阈值 τ_2,即 $\tau_1\geqslant\tau_2$.当以 Q_2 为查询对象做查询时,若 $EMD(Q_2,Q_1)\leqslant\tau_2$,则有 $EMD(Q_1,Q_2)=EMD(Q_2,Q_1)\leqslant\tau_2\leqslant\tau_1$.即在以 Q_1 为查询对象先做查询时,Q_2 在 Q_1 的结果候选集中,也即 $EMD(Q_1,Q_2)$已被计算过.那么当以 Q_2 为查询对象做查询时就不必重复计算 $EMD(Q_1,Q_2)$,不再需要将 Q_1 加入到查询候选集中.

② 若 Q_2 先于 Q_1 做范围查询,则 Q_2 的查询阈值 τ_2 不小于 Q_1 的查询阈值 τ_1,即 $\tau_2\geqslant\tau_1$.第 1 种情况,当以 Q_1 为查询对象做查询时,若 $EMD(Q_1,Q_2)=EMD(Q_2,Q_1)\leqslant\tau_1\leqslant\tau_2$,即 Q_2 在查询对象 Q_1 的结果候选集中.因为 $EMD(Q_2,Q_1)\leqslant\tau_2$,则在 Q_2 先做查询时,Q_1 也在查询对象 Q_2 的结果候选集中.那么 Q_2 先做查询时不计算 $EMD(Q_1,Q_2)$,留到 Q_1 做查询时再计算.第 2 种情况,当以 Q_1 为查询对象做查询时,若 $EMD(Q_1,Q_2)>\tau_1$,即 Q_2 不会出现在查询对象 Q_1 的结果候选集中,同时点对(Q_1,Q_2)也不会出现在局部 S_{Topk} 中,因其最终会被相似性阈值 τ_1 过滤掉.因此在 Q_2 先做查询时也不需要计算点对(Q_1,Q_2)的 EMD 距离,因其最终不会出现在局部 S_{Topk} 中.

因此,在处理数据块的 Top-k 相似性自连接时,在对 B^+树的键值空间做范围查询时只需探测键值大于自身

© 中国科学院软件研究所 http://www.jos.org.cn

键值且落在范围 $[bkey(Q,\Phi),\tau-bckey(Q,\Phi)]$ 内的直方图,即可保证相似性连接结果的完备性. □

根据定理 1,Reduce 阶段在处理同一数据块的 Top-k 相似性自连接时可进一步优化.查询点在第 1 棵 B^+树上做范围查询时,只查询 $bkey$ 值大于自身的直方图,在其他 B^+树上做正常查询.因多棵 B^+树的查询结果会取交集,所以这种优化可进一步减小候选集的大小.

3 基于数据局部性进行数据划分的并行 Top-k 相似性连接

在 Top-k BNLJ 算法中,每条记录都将产生 m 个副本(m 为数据集的分块数),导致 MapReduce 作业在 Shuffle 阶段需要传输平方级(m^2 个块)的数据,同时还增加了 Reduce 阶段的计算开销.在处理大数据量时,由于带宽等资源的限制,该算法的可扩展性受到很大限制.针对这一情况,本文提出基于数据局部性进行数据划分的并行 Top-k DLPJ 相似性连接算法.该算法在 Shuffle 阶段仅需传输线性级(约 m 个块)的数据量,进而可有效提高 Top-k 相似性连接的执行效率.

本文第 2 节已指出利用 EMD 距离对偶问题的可行解,可将直方图概率数据映射到一维实数空间,并使用 B^+树索引一维空间内的映射值,进而将基于 EMD 距离的相似性查询转换为一维映射空间内的范围查询.由公式(4)可知,以 τ 为相似性阈值,Q 为查询对象的相似性搜索的所有结果直方图在一维实数空间内的映射值 $bkey$ 落在 B^+树的一个合法键值区间内.这说明相似直方图的映射值也相似,即直方图概率数据在一维映射空间内具有良好的数据局部性(data locality).这启发我们利用一维映射空间中数据的局部性对概率数据集进行划分,将概率数据集按映射值 $bkey$ 的大小分成 m 个数据块并交由 m 个 Reduce 任务对数据块进行自连接处理.如图 2(a)所示,在一维映射空间内,键值处于$[bkey_{i-1},bkey_i]$中的直方图被划分到数据分块 m_i 中,又根据公式(4)可知,分块 m_i 中的直方图做范围查询时探测的候选直方图的键值一定落在图 2(a)所示的扩展范围内.因此若将分块 m_i 对应的扩展范围交由一个 Reduce 任务处理,则该 Reduce 任务将执行分块内的相似性自连接并找出对应的局部 $S_{\text{Top}k}$.

Top-k DLPJ 算法共由 3 个 MapReduce 作业构成.第 1 次 MapReduce 作业抽样计算近似分位数和 τ 值,分位数用于数据集的划分,τ 值用于加速相似性连接处理;第 2 次 MapReduce 作业查找局部 $S_{\text{Top}k}$;第 3 次 MapReduce 作业从局部 $S_{\text{Top}k}$ 中查找全局 $S_{\text{Top}k}$.因为第 3 个 MapReduce 作业处理逻辑简单,下面重点介绍前两个 MapReduce 作业.

3.1 抽样确定近似分位数和τ值

由于每个 Reduce 任务负责处理一块数据,因此在按照 $bkey$ 值对概率数据集进行划分时需要考虑每一块数据的处理负载,使它们尽可能均衡.在 Reduce 阶段实现绝对的负载均衡是很困难的,因在执行 Top-k 相似性连接时相似性阈值 τ 是实时更新的,所以无法建立算法执行的精确代价模型.这里采用一种易操作的方法来划分概率数据集,即基于数据集在一维映射空间内的 m–1 个分位数,表示为 $\{bkey_1,...,bkey_{m-1}\}(i<j,bkey_i<bkey_j)$,来划分数据集,使得每个数据分块拥有相等的数据量.这里存在一个基本假设,即数据量相等的数据分块有着近似相等的数据处理负载.通过扫描整个数据集去寻找这 m–1 个分位数并不可行.为此,在 Top-k DLPJ 算法开始执行时先启动一次 MapReduce 作业对数据集进行数据抽样,基于样本集可以获得整个数据集的近似分位数,还可以计算出 Top-k 相似性阈值的上界值 τ(τ 的含义详见第 2 节).

在 Map 阶段,每个 Map 任务以概率 p 对输入的数据分片进行采样,得到样本直方图集合 S.之后基于同一组 EMD 距离对偶问题的可行解,每个 Map 任务计算每个样本直方图的 $bkey$ 值并发送给同一个 Reduce 任务.在 Reduce 阶段,Reduce 任务将接收到的 bkey 值从小到大排序,找出 $bkey$ 值域上的 m–1 个分位数将样本集合 S 划分成 m 个分块,并将这 m–1 个分位数作为原始数据集的近似分位数.本次 MapReduce 作业在查找分位数的同时,还需求解 Top-k 相似性阈值的上界值 τ.求解过程与 Top-k BNLJ 算法中的求解过程相同,不再赘述.

3.2 查找局部$S_{\text{Top}k}$

3.2.1 Map 阶段

© 中国科学院软件研究所 http://www.jos.org.cn

基于概率数据集在一维映射空间内的 m–1 个近似分位数,将整个概率数据集划分成 m 个近似等大的数据块发给 m 个 Reduce 任务.这样划分数据的一个关键点在于,如何使得与任一直方图 $Q\in m_i$ 之间的 EMD 距离满足 $EMD(Q,P)\leqslant\tau$ 的所有直方图 P 同 m_i 一起被发送给同一个 Reduce 任务,从而确保该 Reduce 任务连接结果是完备的.定理 2 和推论 1 证明了以块 m_i 包含的直方图为查询对象,τ 为 Top-k 相似性阈值的所有范围查询的查询结果的 $bkey$ 值会落在一个确定的 $bkey$ 值区间内,称为数据块 m_i 的扩展范围区间.且该扩展范围区间覆盖块 m_i 所覆盖的 $bkey$ 值区间(如图 2(a)所示).

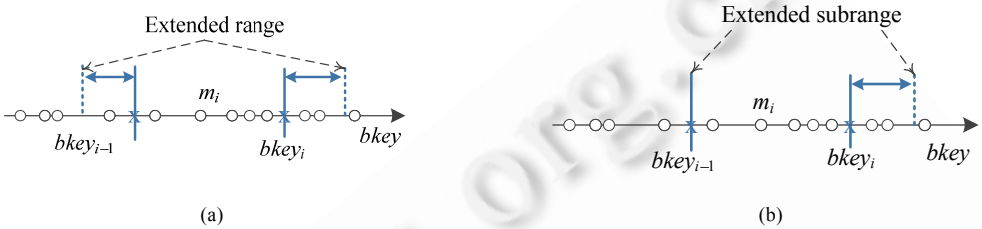


图 2 数据块 m_i 的 $bkey$ 值扩展范围

定理 2. 给定 EMD 距离对偶线性规划问题的一组可行解 Φ 和 Top-k 相似性阈值 τ,数据块 m_i 的扩展范围区间存在确定的上下界.

证明:文献[13]证明对于直方图概率数据 Q,有:

$$bkey(Q,\Phi)+bckey(Q,\Phi)\geqslant\min_i(\phi_i+\pi_i) \tag{8}$$

当相似性阈值 τ 确定后,由公式(8)可推出:

$$\tau-bckey(Q,\Phi)\leqslant\tau-\min_i(\phi_i+\pi_i)+bkey(Q,\Phi)\ .$$

结合公式(4)可知,以 Q 为查询对象 τ 为相似性阈值的范围查询需搜索的 $bkey$ 值区间包含于区间:

$$[\min_i(\phi_i+\pi_i)+bkey(Q,\Phi)-\tau,\tau-\min_i(\phi_i+\pi_i)+bkey(Q,\Phi)] \tag{9}$$

随着查询对象直方图 Q 的 $bkey$ 值增大,公式(9)所示的区间上下界都将增大.故块 m_i 对应的扩展范围区间的下界是 $(\min_i(\phi_i+\pi_i)+bkey_{\min}-\tau)$,上界是 $(\tau-\min_i(\phi_i+\pi_i)+bkey_{\max})$.其中,$bkey_{\min}$ 和 $bkey_{\max}$ 是块 m_i 的左右边界点,即块 m_i 对应的两个分位数(如图 2(a)中的 $bkey_{i-1}$ 和 $bkey_i$). □

由于每个 Reduce 任务都对自己获取的数据块进行自连接处理,因此结合定理 1 的结论,数据块 m_i 对应的扩展范围区间可约简为 $[bkey_{\min},\tau-\min_i(\phi_i+\pi_i)+bkey_{\max}]$(如图 2(b)所示).

在本次 MapReduce 作业的 Map 阶段,map 函数首先基于第 1 个 MapReduce 作业得到的 Top-k 相似性阈值的上界值 τ,以及第 1 个 MapReduce 作业使用的那一组 EMD 距离对偶问题的可行解 Φ 确定每个数据块的 $bkey$ 值扩展范围区间.其次基于 Φ 计算每个直方图概率数据的 $bkey$ 值,从而根据 m–1 个 $bkey$ 值的分位数确定该直方图的所属数据块号 m_i,并将该直方图的一个副本发送到块 m_i 对应的 Reduce 任务.map 函数随后还需判断该直方图是否属于其它分块的扩展范围区间,若属于则发送一个副本到对应的 Reduce 任务(容易理解最后一个数据分块的扩展范围区间与该数据块覆盖的 $bkey$ 值区间相同).基于一个较优的 τ 值所得到的各分块的扩展范围区间和各分块所覆盖的原 $bkey$ 值区间之间的差异很小,即只有少量直方图概率数据需要被重复发送,从而保证 Top-k DLPJ 算法在 Shuffle 阶段具有线性级的数据传输量.

3.2.2 Reduce 阶段

由于本文讨论的 EMD 距离是一种距离测度,因此 EMD 距离具有三角不等性,即对于任意直方图概率数据 P_1,P_2 和 P_3,满足:

$$\left|EMD(P_1,P_2)-EMD(P_2,P_3)\right|\leqslant EMD(P_1,P_3)\leqslant EMD(P_1,P_2)+EMD(P_2,P_3)\ .$$

利用 EMD 距离的三角不等性可进一步约简相似性连接处理中范围查询的候选集.如图 3 所示,P_2 和 P_3 都在查询点 P_1 的搜索范围 R_{P1} 内.若 P_1 先做查询,并且 $EMD(P_1,P_2)$和 $EMD(P_1,P_3)$都在查询处理中被计算过,则在

© 中国科学院软件研究所 http://www.jos.org.cn

P_2 做查询时,若满足$|EMD(P_1,P_2)-EMD(P_1,P_3)|>\Gamma$,则由 EMD 距离的三角不等性可知 $EMD(P_2,P_3)>\Gamma$,即说明 P_3 可从查询点 P_2 的结果候选集里去除掉.

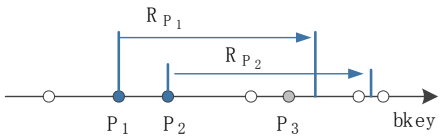


图 3 三角不等式过滤

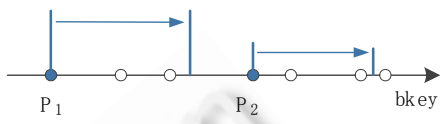


图 4 无效查询记录的删除

在本次 MapReduce 作业的 Reduce 阶段,所有 Reduce 任务都执行数据块的 Top-k 相似性自连接.Reduce 任务将接收到的数据分成左右两个数据块.其中左数据块对应概率数据集的某一分块的数据,右数据块由左数据块对应的扩展范围内的所有数据构成.随后在右数据块上构建 B^+树索引,并以左数据块内的数据为查询对象在 B^+树索引上执行范围查询.在查询过程中,为了有效地利用 EMD 距离的三角不等性进行过滤,一方面左数据块中的直方图需按照 *bkey* 值从小至大的顺序做范围查询;另一方面需要记录之前范围查询中计算过的直方图概率数据对以及它们之间的 EMD 距离,即维护查询记录.为了降低查询记录的维护代价,倘若当前查询对象直方图的 *bkey* 值已经超出之前某查询对象直方图 Q_i 的搜索范围,则可删除和 Q_i 相关的查询记录.以图 4 为例,在以 P_2 为查询对象做范围查询时,即可删除 P_1 的查询记录,因为其对于 P_2 及 P_2 之后的查询对象直方图已无法提供有效的过滤信息了.

4 实验结果与分析

本节使用大规模真实概率数据集对本文提出的 Top-k BNLJ 算法和 Top-k DLPJ 算法进行实验评估.实验使用的集群由 20 台机器组成.一台机器担任主控机(master),其余 19 台机器担任工作机(slaver).每台机器的配置是 Intel(R) Core(TM) i3 CPU(3.10GHz)、8G 内存,运行内核为 2.6.32 的 Linux 操作系统.集群上部署的 MapReduce 框架是 Hadoop 0.20.2 版本.算法实现语言为 C++,使用 Hadoop 针对 C++语言开发的 Hadoop Pipes 接口.

实验使用的真实数据集是基于国内著名的 C2C 交易平台淘宝网上抓取的 640 万张商品图片集生成的.基于该图片集的 Top-k 相似性连接可用于检测海量图片中的潜在副本,从而及时发现商品图片盗用情况.通过提取每张商品图片归一化后的灰度直方图即得到包含多达 640 万条直方图概率数据的概率数据集,大小为 14.4G.灰度直方图的原始维度为 256 维,即包含 256 个数据桶,标识不同的灰度等级.为了测试并行算法的可扩展性,将概率数据集分割成大小分别为 20 万、40 万、80 万、160 万、320 万和 640 万的概率数据集.为了测试并行算法在处理不同维度直方图概率数据时的可扩展性,生成了直方图维度分别为 8、16、32、64、128 和 256 维的 6 个数据集(通过将原 256 维直方图相邻的若干维度累加得到),每个数据集均包含 160 万条概率数据.

以下实验通过改变概率数据集所包含的数据个数 n、Top-k 相似性连接的参数 k、Reduce 任务的数量 r 以及灰度直方图的维度 d,分别测试了 Top-k BNLJ 算法和 Top-k DLPJ 算法在运行时间及 Shuffle 阶段传输数据量上的差异.上述参数的默认设置是:n=160 万,k=200,r=28,d=256.算法对数据集进行抽样时采用的抽样概率是 p=1%.多次实验证实以 1%的概率对数据集进行抽样即可获得一个较理想的 Top-k 相似性阈值的上界值 τ.

(1) 改变数据集大小 n.

图 5 和 6 分别展示了数据集大小对算法执行时间和 Shuffle 阶段的数据传输量的影响.易见随着数据集的增大,Top-k BNLJ 算法的执行时间和数据传输量急剧增长,而 Top-k DLPJ 算法的执行时间和数据传输量则增长缓慢,且在处理大小为 640 万的概率数据集时 Top-k DLPJ 算法最快比 Top-k BNLJ 算法快了近 6.9 倍,说明 Top-k DLPJ 算法具有更好的数据可扩展性.这是因为一方面 Top-k BNLJ 算法需要传输平方级的数据量,而 Top-k DLPJ 算法仅需传输线性级的数据量;另一方面传输数据量的降低也减少了 Reduce 阶段的计算代价.同时,由于 Top-k DLPJ 算法还使用了基于 EMD 距离的三角不等式过滤技术,因此能进一步提升算法的执行效率.

© 中国科学院软件研究所 http://www.jos.org.cn

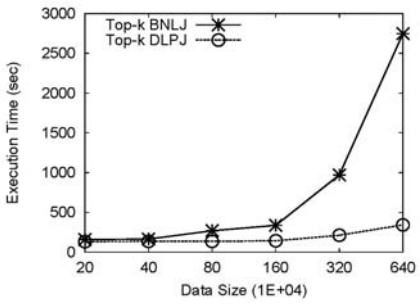


图 5 数据集大小对执行时间的影响

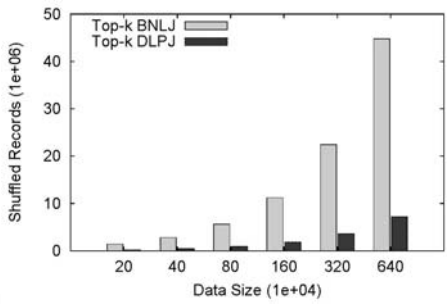


图 6 数据集大小对 Shuffle 阶段传输数据量的影响

(2) 改变 Top-*k* 相似性连接的参数 *k*.

图 7 展示了 *k* 值变化对算法执行时间的影响.随着 *k* 值的增加,需要求精计算的概率数据对的数目也增加了,因而两种算法的执行时间都出现缓慢增长.但由于 Top-*k* DLPJ 算法具有更小的数据传输量并使用了三角不等式过滤技术,因而比 Top-*k* BNLJ 算法的执行效率更高,在不同 *k* 值的情况下,Top-*k* DLPJ 算法平均比 Top-*k* BNLJ 算法快了 1.3 倍.

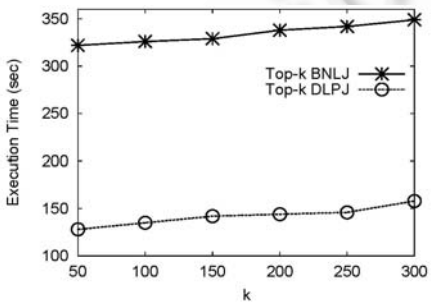


图 7 不同 *k* 对执行时间的影响

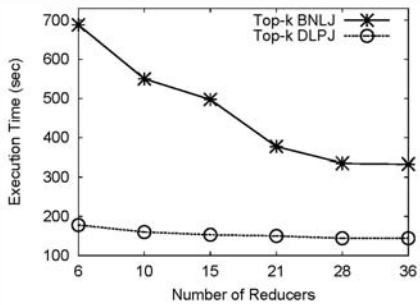


图 8 Reducer 数量对执行时间的影响

(3) 改变 Reduce 任务的数量 *r*.

图 8 和图 9 分别展示了不同 Reduce 任务数 *r* 对 Top-*k* BNLJ 算法和 Top-*k* DLPJ 算法的执行时间和 Shuffle 阶段数据传输量的影响.如图 8 所示,随着 Reduce 任务数量 *r* 的增加,连接算法的并行度也随之增大,因而两种算法的执行时间都出现不同程度的降低,且在启动 36 个 Reduce 任务时,Top-*k* DLPJ 算法至少比 Top-*k* BNLJ 算法快了 1.3 倍.然而,当 *r* 由 28 增加至 36 时,两种算法的执行时间并没有明显下降.这是因为,Reduce 任务增大虽然带来并行执行度大的好处,但也增加了主控机对工作机的通信管理开销.因而实际应用中需要在 Reduce 任务数量和并行系统管理代价之间做权衡.如图 9 所示,随着 Reduce 任务数量 *r* 的增加,Top-*k* BNLJ 算法在 Shuffle 阶段传输的数据量呈线性增长,而 Top-*k* DLPJ 算法传输的数据量并无明显变化.这是因为对于 Top-*k* BNLJ 算法,*r* 增大意味着数据集的数据分块数 *m* 也随之增大,满足 $m(m+1)/2=r$,因而数据传输量 $o(mn)$呈线性增长的趋势.而对于 Top-*k* DLPJ 算法,无论 *r* 为何值,其数据传输量约为 $o(n)$.图 8 和图 9 说明了 Top-*k* DLPJ 算法对处理大数据集具有更好的可扩展性.

(4) 改变直方图概率数据的维度 *d*.

图 10 展示了直方图概率数据的维度 *d* 对两种算法执行时间的影响.随着 *d* 增大,两种算法的执行时间都出现不同程度地增长.特别地,Top-*k* DLPJ 算法的执行时间增长较为缓慢,且在处理 256 维的概率数据时,Top-*k* DLPJ 算法最快比 Top-*k* BNLJ 算法快了 1.3 倍.这是因为对于相同的 Reduce 任务数,Top-*k* DLPJ 算法中每个 Reduce 任务所需处理的数据量较小,因而在处理高维概率数据时具有更好的算法可扩展性.

© 中国科学院软件研究所 http://www.jos.org.cn

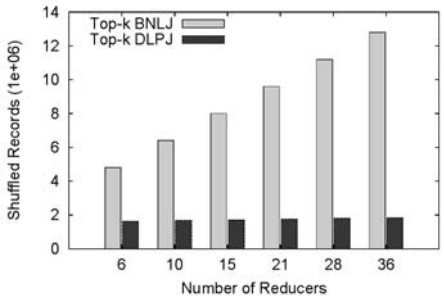


图 9 Reducer 数量对 Shuffle 阶段传输数据量的影响

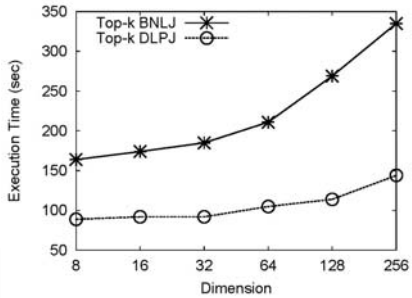


图 10 不同维度对执行时间的影响

5 总 结

本文首次讨论了大规模概率数据上基于 EMD 距离的 Top-*k* 相似性连接问题.形式化定义了基于 EMD 距离的 Top-*k* 相似性连接,提出了两种基于 MapReduce 框架的并行算法.利用 EMD 距离对偶问题的特性,首先提出基于块嵌套循环连接的并行 Top-*k* 相似性连接算法作为基本解决方案;进而改进数据划分策略,提出了基于数据局部性进行数据划分的并行 Top-*k* 相似性连接算法.在大规模真实概率数据集上进行大量实验验证,结果证明了本文提出的基于数据局部性进行数据划分的并行 Top-*k* 相似性连接算法的高效性和处理大数据集时的良好的可扩展性.

References:

[1] Trajcevski G, Wolfson O, Hinrichs K, Chamberlain S. Managing uncertainty in moving objects databases. ACM Trans. on Database Systems (TODS), 2004,29(3):463−507.

[2] Wang DZ, Franklin MJ, Garofalakis M, Hellerstein JM. Querying probabilistic information extraction. In: Proc. of the VLDB Endowment, 2010,3(1):1057−1067.

[3] Deshpande A, Guestrin C, Madden SR, Hellerstein JM, Hong W. Model-Based approximate querying in sensor networks. The VLDB Journal, 2005,14(4):417−443.

[4] Rubner Y, Tomasi C, Guibas LJ. The earth mover's distance as a metric for image retrieval. Int'l Journal of Computer Vision, 2000, 40(2):99−121.

[5] Lakshmanan LVS, Leone N, Ross R, Subrahmanian VS. ProbView: A flexible probabilistic data base system. ACM Trans. on Database Systems, 1997,22(3):419−469.

[6] Benjelloun O, Sarma AD, Halevy A, Widom J. ULDBs: Databases with uncertainty and lineage. In: Proc. of the 32nd Int'l Conference on Very Large Data Bases. 2006. 953−964.

[7] Kim Y, Shim K. Parallel Top-*K* similarity join algorithms using MapReduce. In: Proc. of IEEE 28th Int'l Conf. on Data Engineering (ICDE). 2012. 510−521. [doi: 10.1109/ICDE.2012.87]

[8] Grauman K, Darrell T. Fast contour matching using approximate earth mover's distance. In: Proc. of the 2004 IEEE Computer Society Conf. on Computer Vision and Pattern Recognition. 2004. 220−227. [doi: 10.1109/CVPR.2004.1315035]

[9] Li NH, Li TC, Venkatasubramanian S. *t*-Closeness: privacy beyond *k*-anonymity and *l*-diversity In: Proc. of the IEEE 23rd Int'l Conf. on Data Engineering (ICDE). 2007. 106−115. [doi: 10.1109/ICDE.2007.367856]

[10] Assent I, Wenning A, Seidl T. Approximation techniques for indexing the earth mover's distance in multimedia databases. In: Proc. of the 22nd Int'l Conf. on Data Engineering (ICDE). 2006. [doi: 10.1109/ICDE.2006.25]

[11] Wichterich M, Assent I, Kranen P, Seifl T. Efficient EMD-based similarity search in multimedia databases via flexible dimensionality reduction. In: Proc. of the 2008 ACM SIGMOD Int'l Conf. on Management of data. New York: ACM Press, 2008. 199−212. [doi: 10.1145/1376616.1376639]

[12] Ruttenberg BE, Singh AK. Indexing the earth mover's distance using normal distributions. In: Proc. of the VLDB Endowment. 2011,3(5):205−216.

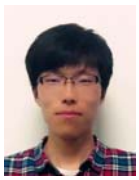
© 中国科学院软件研究所 http://www.jos.org.cn

[13] Xu J, Zhang ZJ, Tung AK, Yu G. Efficient and effective similarity search over probabilistic data based on earth mover's distance. The Int'l Journal on Very Large Data Bases, 2012,21(4):535−559.

[14] Dean J, Ghemawat S. MapReduce: Simplified data processing on large clusters. Communications of the ACM, 2008,1(51): 107−113.

[15] Apache Hadoop. http://hadoop.apache.org

[16] Papadimitriou CH, Steiglitz K. Combinatorial optimization: Algorithms and complexity. New York: Dover Publications, 1998. 67−70.



雷斌(1989－),男,陕西韩城人,硕士生,主要研究领域为数据并行处理.
E-mail: binlei_neu@163.com

谷峪(1981－),男,博士,副教授,CCF 会员,主要研究领域为空间数据管理,图数据管理.
E-mail: guyu@ise.neu.edu.cn

许嘉(1984－),女,博士生,CCF 会员,主要研究领域为概率数据管理.
E-mail: xujia@ise.neu.edu.cn

于戈(1962－),男,博士,教授,博士生导师,CCF 高级会员,主要研究领域为数据库理论和技术,分布与并行系统.
E-mail: yuge@mail.neu.edu.cn

© 中国科学院软件研究所 http://www.jos.org.cn