基于不确定理论的不确定性数据Top-k查询计算

郭长友[1,2] 郑雪峰[1] 高秀莲[2]
(北京科技大学计算机与通信工程学院 北京 100083)[1] (德州学院 德州 253000)[2]

摘 要 在不确定性数据集中,基于参数化排名函数的Top-k查询研究近年来备受关注。给出了一种新的解决方法,该方法将不确定性数据集中的元组建模为不确定网络,将有序元组的Top-k查询等价转化为相应样本图中边的不确定测度关系,并对样本图依据所包含边的排序位置进行分类,从而将不确定性数据中基于参数化排名函数的Top-k查询等价转换为依Top-k值不同的有限查询。本算法避免了计算所有元组在样本图中的排名不确定测度值,提高了不确定图的Top-k查询计算效率。理论分析和实验结果表明,提出的Top-k查询算法能够从非确定角度解决不确定性数据的Top-k查询计算问题。

关键词 不确定网络,不确定测度,样本图指数,Top-k查询

中图法分类号 TP309.5 **文献标识码** A **DOI** 10.11896/j.issn.1002-137X.2016.3.041

Top-k Query Calculation of Uncertain Data Based on Uncertainty Theory

GUO Chang-you[1,2] ZHENG Xue-feng[1] GAO Xiu-lian[2]
(School of Computer and Communication Engineering, University of Science and Technology Beijing, Beijing 100083, China)[1]
(Dezhou University, Dezhou 253000, China)[2]

Abstract The Top-k query in the uncertain data set based on parametric ranking function has been focused in recent years. This paper gave out a new solution. The tuples of uncertain data set is modeled as uncertain network, Top-k query of the orderly tuples is transformed equivalently into uncertain measure relations of edges in corresponding sample figures, and the sample figures are classified according the ranking position of edge contained in them. So the Top-k query in the uncertain data set based on parametric ranking function is transformed equivalently into different limited query with different Top-k value. The proposed algorithm avoids calculating the ranking uncertain measure values of all tuples in the sample figures, and improves the computation efficiency of Top-k query in uncertain figure. Theoretical analysis and experimental results show that the proposed Top-k query algorithm can solve Top-k query calculation of uncertain data from the uncertainty perspective.

Keywords Uncertain networks, Uncertain measure, Sample graph index, Top-k query

1 引言

在数据获取过程中,由于各种因素的影响,使得数据经常具有不确定性,这类不确定性数据的管理已成为近年研究的重要课题。在不确定性数据处理中,文献[1]分析了社会网络环境的不确定性;文献[2]讨论了移动网络环境下的不确定性问题;文献[3]提出了不确定网络环境下的一种近邻查询处理方法。在数据管理过程中,用户只关心最符合查询条件的数据集。确定数据集上的Top-k查询算法无法直接应用到不确定性数据集合上。已有学者关注此问题,包括U-topk[4-6]、ctypicaltopk[7]等。Li等[8,9]提出基于权值参数的排名函数(Parameterized Ranking Function' Prf);文献[10]提出一种剪枝方法;文献[11]对文献[10]的剪枝方法进行了优化,实现了不确定数据的Top-k查询计算方法。

实际上,上述文献还是用概率、模糊去刻画现实中的实体间关系,而实体间关系有时还表现为主观不确定性,这种主观不确定性既不是随机的也不是模糊的,如云计算网络上各虚拟机在CPU、内存、网络带宽等资源中负载均衡,具有动态迁移性,使得负载在迁移前考虑虚拟机间的连通信度、成本、代价等因素具有很大的主观不确定性。因此建立在模糊理论和概率论基础上的方法都不能从根本上解决上述问题。

对于不确定性数据,因为考虑到数据的不确定因素,所以在对数据元组排名时要综合考虑排序属性值和数据不确定测度间的关系。表1给出了一个某时刻云计算网络中虚拟机使用CPU、内存、硬盘、带宽等资源的记录集,在该记录集的获取工程中因数据丢失、传输延迟和监测不精确等因素,使得该记录集的每条记录都有一个不确定测度值,该记录集是一个不确定性数据集。

到稿日期:2015-10-20 返修日期:2015-11-02 本文受国家自然科学基金(61163025,61370063),北京市重点实验室2012年度阶梯计划项目(Z121101002812005)资助。

郭长友 男,博士生,主要研究方向为网络安全等,E-mail:guochangyouustb@139.com;郑雪峰 男,教授,主要研究方向为计算机系统安全、网络安全等;高秀莲 女,副教授,主要研究方向为图论与组合优化等。

表 1　虚拟机监测记录集

虚拟机 ID	CPU (MIPs)	内存 (GB)	硬盘 (GB)	带宽 (MB/s)	不确定测度
VM1	600	2	120	600	1
VM2	700	1.5	250	150	0.4
VM3	900	1	110	350	0.6
VM4	800	0.5	100	300	0.9

将不确定性数据集中的每一个记录按某属性排序，如表 1 按虚拟机使用的内存(GB)大小降序排列。然后每条数据顺序转化为不确定图上的不确定边，每条数据的不确定测度即为该边在不确定图上的不确定测度。从而对该不确定性数据集的研究等价转换为对不确定图相关属性的研究。

确定图上基于参数化排名函数的 Top-k 查询相关研究成果及基于概率和模糊理论的研究结果也不再适用于不确定图的环境。现实中经常需要解决缺乏或者没有观测数据的问题，此类问题无法用概率论求解事件发生的频率。此时，为了求解问题，不得不依据专家的经验和知识估计事件可能发生的信度。若将信度看成主观概率，则推导出的结果与预期差距很大。为了处理主观不确定性，刘宝碇于 2007 年创立了一个新的数学分支——不确定理论[12-14]，其现已发展为基于规范性、对偶性、次可加性及乘积测度公理系统的一个数学分支。理论和实践都表明，不确定性理论是处理非确定信息的一种非常有效的工具。不确定理论被广泛应用于科学和工程等领域中，解决了许许多多的问题。比如，周健[15]解决了不确定最小生成树问题，高秀莲[16,17]研究了不确定图问题，高欣[18,19]研究了不确定测度的性质，高原[20,21]研究了不确定最短路以及不确定图的直径问题。

我们在前期工作[17]中对不确定图进行了研究，不确定网络图中任意两点间可信距离的精确计算时间复杂度较高，达到指数级。为解决此问题，本文将对其拓展，提出元组位置不确定测度、基于不确定理论的元组参数化排名函数值等概念，并将有序元组的 Top-k 查询等价转化为相应样本图中边的不确定测度关系，同时对样本图依据所包含边的排序位置进行分类，以提高不确定图的 Top-k 查询计算效率。

本文的主要工作在于将不确定性数据集建模为不确定网络，通过计算样本图指数获得每个样本图的不确定测度，来优化待计算元组的不确定测度上界的计算方法，提高在不确定网络样本图中 Top-k 查询过程的效率。理论分析和实验结果表明，文中提出的基于不确定理论的不确定性数据 Top-k 查询计算能够从非确定角度解决不确定网络环境下的 Top-k 查询问题，且符合实际情况。

2　相关概念

2.1　不确定理论

不确定理论由清华大学刘宝碇教授在 2007 年[12]提出并于 2010 年[13]进行了修订，为处理不确定因素提供了一种新的研究方法。其已发展为基于规范性、对偶性、次可加性及乘积测度公理系统的一个数学分支。理论和实践都表明，不确定性理论是处理非确定信息的一种非常有效的工具。

以下介绍一些本文所要用到的不确定理论的概念和结果[13]。

定义 1　假设 Γ 为非空集合，L 是 Γ 上的 σ-代数。任意一个元素 $\Lambda\in L$ 被称为一个事件。如果集函数 $M\{\Lambda\}$ 满足以下 3 条公理，则称其为不确定测度。

公理 1(规范性)　$M\{\Gamma\}=1$。

公理 2(对偶性)　对于任意事件 Λ 都有 $M\{\Lambda\}+M\{\Lambda^c\}=1$，记 Λ^c 为 Λ 的补集。

公理 3(次可列可加性)　对任意可数的事件序列 $\{\Lambda_i\}$，都有 $M\{\bigcup_{i=1}^{\infty}\Lambda_i\}\leqslant M\{\Lambda_i\}$。

三元组 (Γ,L,M) 被称为一个不确定空间。不确定变量 ξ 是指从不确定空间 (Γ,L,M) 到实数集上的一个测度函数。

2009 年刘宝碇教授[22]定义了乘积不确定测度，得到下面的第四条公理。

公理 4(乘积公理)　设 (Γ_k,L_k,M_k) $(k=1,2,\cdots)$ 为不确定空间，则乘积不确定测度 M 是乘积 σ-代数 $L=L_1\times L_2\times\cdots$ 上的不确定测度，满足

$$M\{\prod_{k=1}^{\infty}\Lambda_k\}\leqslant\min_{k\geqslant 1}M\{\Lambda_k\}$$

这里，Λ_k 是 L_k 中的任意闭事件，$k=1,2,\cdots$。

如果对任意的实 Borel 集 $B_1,B_2,\cdots,B_m$，满足

$$M\{\bigcap_{k=1}^{m}(\xi_i\in B_i)\}=\min_{1\leqslant k\leqslant m}M\{\xi_i\in B_i\}$$

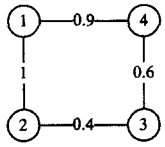
则不确定变量 $\xi_1,\xi_2,\cdots,\xi_m$ 被称为是独立的。

2.2　不确定图的相关概念

定义 2(不确定网络)　不确定网络是指网络实体确定而实体间关系的权值不确定，表现为实体间关系权值的不确定性。

定义 3[17](不确定图)　设 $G=(V,E,M)$ 表示不确定图，其中，V 表示图的结点集，E 表示图的边集，M 表示边存在的不确定测度集。$M(e)$ 表示边 e 存在的不确定测度，其中 $0\leqslant M(e)\leqslant 1$，$M(e)=1$ 表示边 e 一定存在，$M(e)=0$ 表示边 e 一定不存在，$0<M(e)<1$ 称为不确定边。

表 1 的不确定性数据集转换为图 1 的不确定图，其中数字标号 i 表示节点 v_i，边 v_1v_2 表示 vm1 的监测记录，不确定测度为 1。边 v_2v_3 表示 vm2 的监测记录，不确定测度为 0.4。边 v_3v_4 表示 vm3 的监测记录，不确定测度为 0.6。



$$A=\begin{pmatrix}0 & 1 & 0 & 0.9\\ 1 & 0 & 0.4 & 0\\ 0 & 0.4 & 0 & 0.6\\ 0.9 & 0 & 0.6 & 0\end{pmatrix}$$

图 1　不确定图及其邻接矩阵示例

在图 1 中，设不确定图 $G=(V,E,M)$，其中，$V=\{v_1,v_2,v_3,v_4\}$，$E=\{v_1v_2,v_2v_3,v_3v_4,v_4v_1\}$，$M=\{M(v_1v_2),M(v_1v_4),M(v_2v_3),M(v_3v_4)\}$，它有 3 条不确定边 v_2v_3、v_3v_4、v_4v_1，其不确定测度分别为 0.4、0.6 和 0.9。

定义 4[17](样本图)　已知不确定图 $G=(V,E,M)$，样本图 G_k 是不确定图 G 的一个实例，一个有 m 条不确定边的不确定图有 2^m 个样本图。

图 2 是图 1 所示的不确定图 G 的样本图示例，因图 1 中的不确定图有 3 条不确定边，故其有 $2^3=8$ 个样本图。其中，图 2(a)为样本图 G_1，$E_{G_1}=\{v_1v_2\}$，图 2(h)为样本图 G_8，$E_{G_8}=\{v_1v_2,v_2v_3,v_3v_4,v_4v_1\}$。

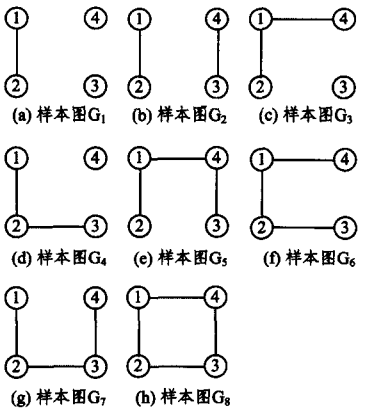


图 2　图 1 的样本图

定理 1[17]　假设 G 是一个 n 阶不确定图，其邻接矩阵为

$$A=\begin{pmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \vdots & \vdots & \cdots & \vdots \\ a_{n1} & a_{n2} & \cdots & a_{nn} \end{pmatrix}$$

如果所有的边都是独立的，则 G 的样本图指数为

$$\rho_k(G)=\begin{cases} \sup\limits_{S_k(G)=1} \min\limits_{1\leqslant i,j\leqslant n} v_{ij}(x_{ij}), & \text{if } \sup\limits_{S_k(G)=1} \min\limits_{1\leqslant i,j\leqslant n} v_{ij}(x_{ij})<0.5 \\ 1-\sup\limits_{S_k(G)=0} \min\limits_{1\leqslant i,j\leqslant n} v_{ij}(x_{ij}), & \text{if } \sup\limits_{S_k(G)=1} \min\limits_{1\leqslant i,j\leqslant n} v_{ij}(x_{ij})\geqslant 0.5 \end{cases}$$

$k=1,2,\cdots,2^m$

其中，X 是 $n\times n$ 不确定对称矩阵，且 X 的主对角线元素是 0，使得

$$X=\begin{pmatrix} x_{11} & x_{12} & \cdots & x_{1n} \\ x_{21} & x_{22} & \cdots & x_{2n} \\ \vdots & \vdots & \cdots & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{nn} \end{pmatrix}$$

这里 $x_{ji}=x_{ij}$，其值为 0 或 1，v_{ij} 的取值为

$$v_{ij}(x_{ij})=\begin{cases} a_{ij}, & \text{if } x_{ij}=1 \\ 1-a_{ij}, & \text{if } x_{ij}=0 \end{cases}, i,j=1,2,\cdots,n$$

图 2(a)为样本图 G_1，$E_{G_1}=\{v_1v_2\}$，根据定理 1 可以计算出 $\sup\limits_{S(G_1)=1} \min\limits_{1\leqslant i,j\leqslant n} v_{ij}(x_{ij})=0.1<0.5$，故样本图 G_1 的不确定测度为 0.1。图 2(b)为样本图 G_2，$E_{G_2}=\{v_1v_2,v_3v_4\}$，根据定理 1 可以计算出 $\sup\limits_{S(G_2)=1} \min\limits_{1\leqslant i,j\leqslant n} v_{ij}(x_{ij})=0.1<0.5$，故样本图 G_2 的不确定测度为 0.1。图 2(c)为样本图 G_3，$E_{G_3}=\{v_1v_2,v_1v_4\}$，根据定理 1 可以计算出 $\sup\limits_{S(G_3)=1} \min\limits_{1\leqslant i,j\leqslant n} v_{ij}(x_{ij})=0.4<0.5$，故样本图 G_3 的不确定测度为 0.4。

3　基于 Prf 语义的不确定数据 Top-k 查询

3.1　参数化排名函数

定义 5　n 个不确定性数据元组组成的集合 $T=\{t_1,t_2,\cdots,t_j,\cdots,t_n\}$ 中，对于任一元组 $t_j=(s_j,m_j)$，s_j 为 T 中的排序属性，m_j 为元组存在的不确定测度，其中 $t_1,t_2,\cdots,t_j,\cdots,t_n$ 已按属性值降序排列，集合 T 的可能世界样本空间 $W=\{W_1,W_2,\cdots,W_m\}$。记 $M(W)$ 为可能世界 W 的不确定测度，记 $M(i,j)$ 为元组 t_j 的位置不确定测度，表示包含元组 t_j 的所有可能世界中，属性值排在第 i 位的不确定测度最大值，即

$$M(i,j)=\max_{t_j\in W_k,W_k\in W} M(W_k)$$

定义 6　n 个不确定性数据元组组成的集合 $T=\{t_1,t_2,\cdots,t_j,\cdots,t_n\}$ 中，$t_1,t_2,\cdots,t_j,\cdots,t_n$ 已按属性值降序排列，元组的参数化排名函数值 $\gamma(t_j)$ 定义为 $\gamma(t_j)=\sum\limits_{1\leqslant i\leqslant j}\omega(i)M(i,j)$，其中，$\omega(i)$ 是权值参数，$M(i,j)$ 是元组 t_j 的位置不确定测度。

3.2　参数化排名函数的计算方法

若计算所有元组的 Prf 值，再取前 k 个结果，算法复杂度较高，达到指数级。通过对 Prf 查询分析，得出如下定理。

定理 2　在不确定性数据集 $T=\{t_1,t_2,\cdots,t_j,\cdots,t_n\}$ 中，设 $t_1,t_2,\cdots,t_j,\cdots,t_n$ 已按属性值降序排序，任一元组 t_j 的 $\gamma(t_j)$ 只与排在它之前的元组 $t_1,t_2,\cdots,t_{i-1}$ 的位置不确定测度有关。

证明：根据定义 6，元组 t_j 的 Prf 函数值：

$$\gamma(t_j)=\sum_{1\leqslant i\leqslant j}\omega(i)M(i,j)$$

计算元组 t_{j+1} 的 $\gamma(t_j)$，因包含元组 t_j 的任一 W 加入元组 t_{j+1} 之后，元组 t_j 的排名位置不变，即 $M(i,j)$ 不变，因此元组 t_j 的 Prf 值 $\gamma(t_j)$ 不变。

定义 7　n 个不确定性数据元组组成的集合 $T=\{t_1,t_2,\cdots,t_j,\cdots,t_n\}$ 中的任意一元组 $t_j=(s_j,m_j)$ 在各个位置上的位置不确定测度 $M(i,j)$ 之和为该元组的存在不确定测度 M_j，即

$$\sum_{1\leqslant i\leqslant j}M(i,j)=M_j$$

定义 8　n 个不确定性数据元组组成的集合 $T=\{t_1,t_2,\cdots,t_j,\cdots,t_n\}$ 中，$t_1,t_2,\cdots,t_j,\cdots,t_n$ 已按属性值降序排列，对于其任一个元组 $t_j=(s_j,m_j)$，各位置不确定测度 $Ms(i,j)$ 表示前 j 个元组中在可能世界中属性值排为第 i 位的不确定测度之和，即 $Ms(i,j)=\sum\limits_{1\leqslant l\leqslant j}M(i,l)$。记剩余未计算的任一元组 t_x 在所有可能世界中排在第 i 位的最大位置不确定测度为 $M_{\max}(i,j)$。

定理 3　在数据集 $T=\{t_1,t_2,\cdots,t_j,\cdots,t_n\}$ 中，对于 $\forall i,j$ $(1\leqslant i\leqslant j\leqslant n)$，都有 $Ms(i,j)\leqslant 1$。

证明：由可能世界的定义知，任一可能样本世界中，每个位置上的不确定测度和小于等于所有可能样本世界不确定测度和，所有元组排在第 i 位的不确定测度 $Ms(i,n)\leqslant 1$，所以对于所有的 i,j，都有 $Ms(i,j)\leqslant Ms(i,n)\leqslant 1$。

根据定理 2，Top-k 剪枝计算方法的主要步骤如下：

(1)计算已按某属性值排序的前 k 个元组的 Prf 值。

(2)将前 k 个元组作为 Top-k 集合的初值，记 min 为 Top-k 集合的最小值。记 Tlist 为已计算的前 j 个元组集。

(3)计算下一元组 tnext 的 Prf 值。

若 $Prf>min$，则更新当前 Top-k 集合，计算 min；把 tnext 加入到 Tlist。

(4)τ=SUPER(Tlist);//任一未计算元组的 Prf 值上界通过已计算元组集来求得。

(5)若 $\tau>min$，则执行(3)。

若算法未结束，则取 k 个未计算的较大元组进行计算。记任意一个未计算元组 t_x 的 Prf 值上界为 τ(注：元组 t_x 排在第 j 位，$x>j$)。

根据定义 3，

$$\begin{aligned}\gamma(t_x)&=\sum_{1\leqslant l\leqslant x}\omega(l)M(l,x)\\&=\sum_{1\leqslant l_1\leqslant j}\omega(l_1)M(l_1,x)+\sum_{j+1\leqslant l_2\leqslant x}\omega(l_2)M(l_2,x)\\&\leqslant\sum_{1\leqslant l_1\leqslant j}\omega(l_1)M_{\max}(l_1,j)+\sum_{j+1\leqslant l_2\leqslant n}\omega(l_2)M(l_2,x)\end{aligned}$$

又根据定义 7，

$$\sum_{1\leqslant l\leqslant x} M(i,x)=M(t_x)$$

且假设权值参数单调递减，若

$$\exists m,1\leqslant m<j,\ \sum_{1\leqslant l\leqslant m} M_{\max}(l,j)\leqslant M(t_x)$$

且

$$\sum_{1\leqslant l\leqslant m} M_{\max}(l,j)+M_{\max}(m+1,j)>M(t_x)$$

那么

$$\gamma(t_x)\leqslant \sum_{1\leqslant l\leqslant m}\omega(l)M_{\max}(l,j)+\omega(m+1)(M(t_x)-\sum_{1\leqslant l\leqslant m}M_{\max}(l,j))=\tau$$

否则

$$\gamma(t_x)\leqslant \sum_{1\leqslant l\leqslant m}\omega(l)M_{\max}(l,j)+\omega(j+1)(M(t_x)-\sum_{1\leqslant l\leqslant j}M_{\max}(l,j))=\tau$$

然后计算最大位置不确定测度 $M_{\max}(l,j)$。假设已计算得出前 j 位元组的 Prf 值及 j 个元组的 $\{Ms(1,j),Ms(2,j),\cdots,Ms(m,j),\cdots,Ms(j,j)\}$，根据定理 3，$Ms(m,j)\leqslant 1$，因此元组 t_x 在所有可能世界中排在第 i 位的最大位置不确定测度和为

$$M_{\max}(i,j)=1+M_t(i,j)$$

且 t_x 的存在不确定测度最大值为 $M(t_x)=1$。该未计算元组的不确定信度上界是通过已计算元组在各个位置上的位置不确定测度和得到的。

4 云计算环境下的不确定性数据 Top-k 查询

云计算(Cloud Computing)是一种正在蓬勃发展的商业计算模式。在这种模式下，云计算用户能够按需获取计算力、存储空间和信息服务[23]。各大 IT 企业纷纷推出各种云计算平台，然而云计算编程模式始终是云服务商必须面对的问题，它是用户考虑是否选用云计算平台的首要因素。简单地将传统的串行编程算法完全移植到云平台下无效。

云计算是在分布式网络的基础上发展而来的，相比分布式网络，它的分布性、动态性以及虚拟化程度都达到了前所未有的高度。从云计算自身视角来看，在云计算中，用户很难有效掌握自己数据存储在哪些节点、应用运行在哪些节点以及所使用的虚拟资源由哪些实际节点提供等信息。

4.1 云计算环境下的分布式计算框架

云计算环境下的计算框架的两个主要实例是批处理系统 Map-Reduce[24]和实时处理系统，如 Apache Spark[25]、Twitter Storm 和 Yahoo S4[26]。上述系统均以(Key,Value)方式来处理数据。

MapReduce 的数据处理过程如图 3 所示，主要分为映射阶段(Map 函数)和化简阶段(Reduce 函数)。在映射阶段，首先将一个大任务划分成了 M 个子任务(即 M 个分段)，每个分段都有键值对〈Key,Value〉。其次，在每个 Map 操作中，该片段的键值对〈Key,Value〉输入是通过执行用户根据实际需要而编写的 Map 函数来实现的。最后，输出一个中间态的键值对，将中间态的 Key 排序，将中间态 Key 的个数作为 Reduce 任务的数目。在 Reduce 阶段，运行用户根据实际需要编写的 Reduce 函数，最终得到用户需要的结果输出[27,28]。现在的云计算模型(如 Map-Reduce、Hadoop、Spark、Dryad/DryadLINQ)都可以在集群上使用程序实现大规模静态数据和高速实时数据的有效分析。

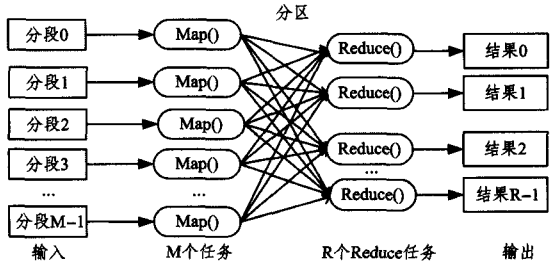


图 3 MapReduce 处理数据流程

目前，基于 MapReduce 技术的应用研究有很多，文献[29]提出了一系列在 Map-Reduce 框架下高效处理 Skyline 的查询算法。文献[30]给出了改进的 Apriori 算法在 Hadoop 中的 MapReduce 编程模型上的执行流程。

4.2 云计算环境下基于 MapReduce 的不确定性数据 Top-k 查询

从定义 4 可看出，在计算样本可能世界 W 的不确定测度及不确定性元组集的位置不确定测度时，都非常适合用并行程序设计来实现。考虑到单机环境的处理速度、处理数据规模等因素，将本文所提算法应用到云计算环境下的 Map-Reduce编程框架下实现。

结合云计算环境下的批处理系统 MapReduce 的特点，提出云计算环境下基于 MapReduce 的不确定性数据 Top-k 查询算法。

算法流程为：首先可以将按某属性值排序预处理过后的不确定性数据集分成多份，然后交由多个 Map 任务进行处理。每个 Map 任务负责执行所分得数据片上的 Top-k 计算，然后由一个 Reduce 求得已计算元组的 Top-k 计算，并判断是否已求得所有数据集的 Top-k 查询，若没有，则取下一批元组集，直到算法结束。

Map 函数：计算所分得数据片上的元组集在指定属性上的最大值 *Max-Value*，将指定属性上排序前 m 的数据元组发送给 Reduce 任务。

Reduce 函数：求得所有 Map 任务发送的指定属性值上最小元组的上确界，记为 *supre-min*。在 Reduce 中利用本文所提算法求得属性值在区间[*supre-min*,*Max-Value*]中的最大元组集。若算法不满足终止条件，则将本次作业的 *supre-min* 作为下次作业的 *Max-Value*。

5 实验与结果分析

针对以上算法进行大量实验来验证本文所提出的不确定数据 Top-k 查询计算的执行效率及其稳定性。由于目前在不确定网络研究中，尚没有基于不确定理论下的位置不确定测度等条件约束的 Top-k 查询算法，本文进行如下两组实验分析。第一组实验主要考察提出的查询方法在不同规模及不同约束条件下的不确定性数据集上的执行效率。第二组实验基于 Apache Spark 和 Map-Reduce 分布式计算来实现所提出的查询算法，在不同规模的数据集、不同不确定测度期望、不同 Top-k 值查询下的运行时间等方面进行实验分析。第一组实验数据在 Windows 7 操作系统、1.6GHz 处理器和 4GB 内存条件下，采用 Visual C++6.0 编程环境和人工模拟不确定网络实现。

表 2　实验使用人工模拟的不确定性数据集

元组集名称	元组数
T_1	9000
T_2	3500
T_3	700

实验使用定义 4 中定义的不确定网络，使用表 2 所列的 3 个不确定记录集。假定数据集中元组已按某属性降序排列，每个元组的不确定测度为随机生成的[0,1]间的实数。实验部分测试在不同元组集上不同 Top-k 值查询对不确定性数据 Top-k 查询算法的影响。

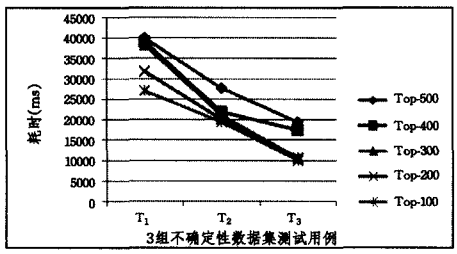


图 4　Top-k 的查询结果

实验 1　测试不同元组集对不确定性数据 Top-k 查询算法的影响，Top-k 的查询结果如图 4 所示。从图 4 可知，不同元组集对不确定性数据 Top-k 查询算法有相似分布，随着不确定性数据集中元组数量的变化而相应变化。

第二组实验的硬件环境是 8 台 IBM 服务器，服务器配置为 2 Intel Quad-Core Xeon E5420 2.5GHz/EM64T CPU，12MB L2 Cache，8GB PC2-5300 CL5 ECC DDR2 667MHz 主存，146G HDD * 3。操作系统：带有 Eclipse 的 64 位 Centos Linux 6.4。实验过程中，使用表 2 所列的 3 组不确定性数据集。表 3 中不同规模的数据集来源于国内某著名电商服务平台，在每个数据记录上添加一个[0,1]区间的随机值作为元组存在不确定测度。

表 3　实验使用真实的大规模不确定性数据集

元组集名称	元组数
T_1	100000
T_2	50000
T_3	10000

实验使用定义 4 中定义的不确定网络。假定数据集中元组已按某属性降序排列，每个元组的不确定测度为随机生成的[0,1]间的实数。实验部分测试在 T_1、T_2、T_3 不同元组集上本文所提算法在元组集具有不同不确定测度期望、不同 Top-k 值查询时的运行时间的变化。

实验 2　表 2 所列的 T_1、T_2、T_3 3 个不同规模的不确定性元组集在不同的不确定测度期望值下，不确定性数据 Top-k 查询算法的运行时间如图 5 所示。从图 5 可知，不确定性元组集所具有的不确定测度期望值对不确定性数据 Top-k 查询算法有相似分布，随着不确定性数据集中元组的存在不确定测度变化而相应变化。

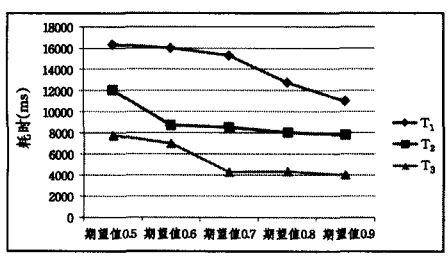


图 5　不同测度期望值下 Top-k 查询分析

在此实验中，k 取 100，由图 5 可看出，当元组存在不确定测度的期望越大，本文提出的算法所用的运行时间越少。分析认为，由元组存在的不确定测度定义及定理 2 可知，本文所提算法所用时间会因元组存在的不确定测度值的增大而减少。

实验 3　表 2 所列的 T_1、T_2、T_3 3 个不同规模的不确定性元组集在元组集的不确定测度期望值为 0.8 的情况下，不确定性数据 Top-k 查询在不同 Top-k 值查询时的结果如图 6 所示。从图 6 可知，不同规模的不确定性元组集在不同 Top-k 值查询时所用的时间有相似分布，随着 Top-k 值变化而相应变化。由图 6 可看出，Top-k 值越大，所用的运行时间越多，符合真实情况。

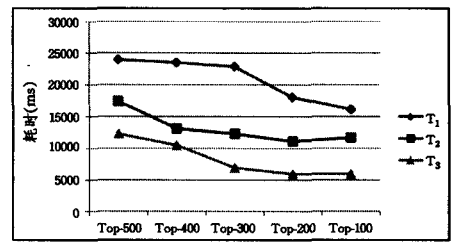


图 6　不同 Top-k 值查询的运行时间

结束语　本文给出一种新的不确定性数据集上的 Top-k 查询问题解决方法。通过将不确定性数据集中的元组建模为不确定网络中的边，将有序元组的 Top-k 查询等价转化为相应样本图中边的不确定测度关系，并对样本图依据所包含边的排序位置进行分类，根据参数化排名函数的特性，该算法避免了计算所有元组在样本图中的排名不确定测度值，提高了不确定图的 Top-k 查询计算效率。

参 考 文 献

[1] Adar E, Re C. Managing Uncertainty in Social Networks[J]. IEEE Data Engineering Bulletin, 2007, 30(2): 15-22

[2] Ghosh J, Ngo H, Yoon S, et al. On a Routing Problem Within Probabilistic Graphs and Its Application to Intermittently Connected Networks[C]// Proceedings of INFOCOM'07. [S. l.]: IEEE Press, 2007: 216-222

[3] Potamias M, Bonchi F, Gionis A, et al. Nearest-neighbor Queries in Probabilistic Graphs[EB/OL]. (2009-10-21). http://www.cs.bu.edu

[4] Soliman M A'Ilyas I F. Ranking with Uncertain Scores[C]// Proc of the 25th IEEE International Conference on Data Engineering. Shanghai, China, 2009: 317-328

[5] Hua Ming, Pei Jian, Liu Xue-min. Ranking Queries on UncertainData[J]. The International Journal on Very Large Data Bases, 2011, 20(1): 129-153

[6] Jestes J, Cormode G, Li Fei-fei, et al. Semantics of Ranking Queries for Probabilistic Data[J]. IEEE Trans on Knowledge and Data Engineering, 2011, 23(12): 1903-1917

[7] Ge Tingj-ian, Zdonik S, Madden S. Top-k Queries on Uncertain Data: On Score Distribution and Typical Answers[C]// Proc of the ACM SIGMOD International Conference on Management of Data. Providence, USA, 2009: 375-388

[8] Li Jian, Saha B, Deshpande A. An Unified Approach to Ranking in Probabilistic Databases[J]. The VLDB Journal, 2011, 20(2): 249-275

[9] Li Jian, Deshpande A. Ranking Continuous Probabilistic Datasets[J]. Proceedings of the VLDB Endowment, 2010, 3(1/2): 638-649

[10] Wang Chong-hai, Yuan Li-yan, You Jia-huai, et al. On Pruning

for Top-k Ranking in Uncertain Databases[J]. Proceedings of the VLDB Endowment,2011,4(10):598-609

[11] Lu Xin,Chen Hua-hui,Dong Yi-hong,et al. Top-k Query Calculations on Uncertain Dataset under MapReduce Framework[J]. Pattem Recognition and Aitificial Intelligence,2013(7):695-704 (in Chihese)
卢鑫,陈华辉,董一鸿,等. MapReduce 框架下的不确定数据 Top-k 查询计算[J]. 模式识别与人工智能,2013(7):695-704

[12] Liu B. Uncertainty Theory(2nd edition)[M]. Springer-Verlag, Berlin,2007

[13] Liu B. Uncertainty Theory: A Branch of Mathematics for Modeling Human Uncertainty[M]. Springer-Verlag,Berlin,2010

[14] Liu Bao-ding. Uncertainty Distribution and Independence of Uncertain Processes[J]. Fuzzy Optimization and Decision Making, 2014,13(3):259-271

[15] Zhou Jian,Chen Lu,Wang Ke. Path Optimality Conditions for Minimum Spanning Tree Problem with Uncertain Edge Weights,International Journal of Uncertainty[J]. Fuzziness and Knowledge-Based Systems,2015,23(1):49-71

[16] Gao X L. Uncertain relations on a finite set and their properties [J]. Pure and Applied Mathematics Journal,2014,3(1):13-19

[17] Gao X L,Gao Y. Connectedness Index of Uncertainty Graphs, International Journal of Uncertainty[J]. Fuzziness and Knowledge-Based Systems,2013,21(1):127-137

[18] Gao X. Some properties of continuous uncertain measure,International Journal of Uncertainty[J]. Fuzziness and Knowledge-Based Systems,2009,17(3):419-426

[19] Gao X,Gao Y,Ralescu D. On Liu's Inference Rule for Uncertain Systems,International Journal of Uncertainty[J]. Fuzziness and Knowledge-Based Systems,2010,18(1):1-11

[20] Gao Y,Yang L X,et al. On Distribution Function of the Diameter in Uncertain Graph[J]. Information Sciences,2015,296(1): 61-74

[21] Gao Y. Shortest Path Problem with Uncertain Arc Lengths [J]. Computers and Mathematics with Applications,2015,296 (1):61-74

[22] Liu B. Some research problems in uncertainty theory[J]. Journal of Uncertain Systems,2009,3(1):3-10

[23] Luo Jun-zhou, Jin Jia-hui, Song Ai-bo, et al. Cloud computing: architecture and key technologies[J]. Journal on Communications,2011,32(7):3-21(in Chinese)
罗军舟,金嘉晖,宋爱波,等. 云计算:体系架构与关键技术[J]. 通信学报,2011,32(7):3-21

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

[25] Zaharia M, Chowdhury M, Franklin M J, et al. Spark: cluster computing with working sets[C]//Proceedings of the 2nd USENIX Conference on Hot Topics in Cloud Computing. 2010:10

[26] Neumeyer L, Robbins B, Nair A, et al. S4: Distributed stream computing platform[C]// 2010 IEEE International Conference on Data Mining Workshops (ICDMW). IEEE,2010:170-177

[27] Wieder A,Bhatotia P,Post A,et al. Brief announcement:modelling MapReduce for optimal execution in the cloud[C]// Proceedings of the 29th ACM SIGACT-SIGOPS Symposium on Principles of Distributed Computing. New York, NY, USA: ACM,2010:408-409

[28] Zheng Q. Improving MapReduce fault tolerance in the cloud[C]// 2010 IEEE International Symposium on Parallel & Distributed Processing,Workshops and Phd Forum (IPDPSW). Atlanta, GA:IEEE,2010:1-6

[29] Ding Lin-lin,Xin Jun-chang,Wang Guo-ren,et al. Efficient Skyline Query Processing of Massive Data Based on Map-Reduce [J]. Chinese Journal of Computers,2011,34(10):1785-1796(in Chinese)
丁琳琳,信俊昌,王国仁,等. 基于 Map-Reduce 的海量数据高效 Skyline 查询处理[J]. 计算机学报,2011,34(10):1785-1796

[30] Li Ling-juan,Zhang Min. Research on Algorithms of Mining AssociationRule under Cloud Computing Environment[J]. Computer Technology and Development,2011,21(2):43-46(in Chinese)
李玲娟,张敏. 云计算环境下关联规则挖掘算法的研究[J]. 计算机技术与发展,2011,21(2):43-46

[31] Zeng Qing-sen, Huang Xian-ying. Fast Data Mining Algorithm Based on FP-tree[J]. Journal of Chongqing Institute of Technology(Natural Science),2009,23(10):72-76(in Chinese)
曾庆森,黄贤英. 基于 FP-tree 的快速数据挖掘算法[J]. 重庆工学院学报(自然科学版),2009,23(10):72-76

(上接第 224 页)

[3] Neumann T, Weikum G. The RDF-3X engine for scalable management of RDF data[J]. The VLDB Journal,2010,19:91-113

[4] Wang Yan,Tian Cui-hua,Zhu Shun-zhi,et al. RDF Data Index Method Based on Association of SPARQL Query Twis[J]. Journal of Xiamen University(Natural Science),2014,53(3):322-329(in Chinese)
王琰,田翠华,朱顺痣,等. 基于 SPARQL 查询小枝关联的 RDF 数据索引方案[J]. 厦门大学学报(自然科学版),2014,53(3): 322-329

[5] Weiss C, Karras P, Bemstein A. Hexastore: sextuple indexing for semantic web data anagementl [C]//Proceedings of the 34rd International Conference on Very Large Data Bases. New York: ACM,2008:1008-1019

[6] Dong Shu-jian,Wang Jing-bin. HMSST:An efficient algorithm for SPARQL query[J]. Computer Science,2014,41(S2):323-326,336(in Chinese)
董书琳,汪璟玢. HMSST:一种高效的 SPARQL 查询优化算法[J]. 计算机科学,2014,41(S2):323-326,336

[7] Zeng Chao-yu,Li Jin-xiang. Redis application in cache system [J]. Microcomputer&Its Applications,2013,12:11-13(in Chinese)
曾超宇,李金香. Redis 在高速缓存系统中的应用[J]. 微型机与应用,2013,12:11-13

[8] Gao X,Fang X. High-Performance Distributed Cache Architecture Based on Redis[C]//Proceedings of the 9th International Symposium on Linear Drives for Industry Applications,Volume 1. Springer Berlin Heidelberg,2014:105-111

[9] Guo Y, Pan Z, Heflin J. LUBM: A benchmark for OWL knowledge base systems[J]. Web Semantics: Science, Services and Agents on the World Wide Web,2005,3(2):158-182

[10] Huang H,Liu C. Selectivity estimation for SPARQL graph pattern[C]//Proceedings of the 19th international conference on World Wide Web. ACM,2010:1115-1116

[11] Liu L,Yin J,Gao L. Efficient Social Network Data Query Processing on MapReduce[C]//Proc of the 5th ACM workshop. New York:ACM,2013:27-32

[12] Kim H S,Ravindra P,Anyanwu K. From SPARQL to MapReduce:The journey using a nested TripleGroup algebra[J]. Proc. of the VLDB Endowment,2011,4(12):1426-1429