第 33 卷 第 5 期
2018 年 10 月

天津科技大学学报
Journal of Tianjin University of Science & Technology

Vol. 33 No. 5
Oct. 2018

DOI:10.13364/j.issn.1672-6510.20170260

不确定随机网络 Top-k 最近节点查询算法

连春月，李孝忠，牛浩浩
（天津科技大学计算机科学与信息工程学院，天津 300457）

摘 要：基于机会理论，提出了在非确定数据和不确定数据同时存在条件下的不确定随机网络的 Top-k 最近节点的查询问题. 对一个不确定随机网络，在一定的机会测度下，将节点间的权重建模为节点间的路径长度，根据路径长度寻找距离指定节点最近的 k 个节点. 该算法能有效解决在经验数据和小样本数据混杂情况下的节点查询问题.

关键词：机会理论；Top-k 查询；不确定随机网络；机会测度

中图分类号：TP301.6；O157.5 **文献标志码**：A **文章编号**：1672-6510(2018)05-0068-05

Top-k Nearest Node Query Algorithm for Uncertain Random Networks

LIAN Chunyue，LI Xiaozhong，NIU Haohao
(College of Computer Science and Information Engineering, Tianjin University of Science & Technology, Tianjin 300457, China)

Abstract: Based on the chance theory, this paper proposes a Top-k query algorithm for uncertain stochastic networks with uncertain data. For an uncertain random network, the weight between nodes is modeled as the path length between nodes under a certain chance measure, and the k nodes closest to the specified nodes are searched according to the path length. The algorithm can effectively solve the problem of node query under the mixed situation of experienced data and small sample data.

Key words: chance theory; Top-k query; uncertain random network; chance measure

实际生活中，会遇到许多非确定的现象. 为了研究这类非确定的现象，17 世纪诞生了概率论. 概率论基于大量的历史数据，有效解决了很多统计问题. 但是，有时因为各种原因无法获得足够多的数据，这时使用概率论解决问题就很难办到. 为了解决这类问题，Liu B 于 2007 年提出了不确定理论[1]，并于 2010 年对不确定理论进行重新定义[2]，为小样本数据、甚至是无样本数据的统计提供了新的理论基础.

经过多年研究与实践，不确定理论得到了充分的发展与广泛的应用. 2013 年，高原[3]详细研究了不确定网络的最短路径问题；2014 年，Zhou 等[4]研究了不确定网络的最短路径的逆不确定分布问题；2015 年，Zhou 等[5]给出了不确定网络的最小生成树的路径最优条件.

对于一个复杂网络，某些弧的权重可以通过对历史数据进行统计分析得到，而某些弧由于没有历史数据或者历史数据无效，导致其权重不能通过概率统计得到，只能利用专家的经验数据，得到不确定弧的权重的不确定分布函数.

2013 年，为了解决这类既有非确定因素，又有不确定因素的现象，Liu Y[6]开创了机会理论. 2014 年，Liu B[7]首次将机会理论引入不确定网络，提出不确定随机网络的概念. 2015 年，盛玉红[8]对不确定随机网络的最短路径问题、最小生成树问题和最大流问题进行研究，提出理想机会分布函数的概念并利用其求解了上述问题.

关于非确定性数据的 Top-k 查询问题，学者们提出了各种计算方法，包括 U-topK[9]、U-kRanks[10]，PT-k[11]，Global-topK[12]等. Li 等[13]提出了基于权值参数的排名函数，实现了排名分值与概率平衡. 2016 年，

收稿日期：2017–09–24；**修回日期**：2018–05–08
基金项目：国家自然科学基金资助项目(61603273)；天津市自然科学基金资助项目(16JCYBJC18500)
作者简介：连春月（1993—），女，黑龙江人，硕士研究生；**通信作者**：李孝忠，教授，lixz@tust.edu.cn

郭长友等[14]首次将不确定理论应用到不确定数据的 Top-k 查询计算中.

非确定数据的 Top-k 查询在 P2P 系统、电缆铺设等方面已经有了实际应用，但不确定数据和随机数据同时存在的 Top-k 查询方面的研究并不多. 本文基于现有研究成果，将包含不确定数据和随机数据的问题转化为不确定随机网络，并在深度优先遍历的算法上，提出在一定机会测度的情况下，寻找距离某个节点最近的 k 个节点的算法，能有效解决在经验数据和小样本数据混杂情况下的节点查询问题.

1　相关概念

1.1　不确定理论[1]

定义 1　设 Γ 为非空集合，L 是 Γ 上的 σ–代数，任意的 $\Lambda\in L$ 称为一个事件. 如果从 L 到实数集 $\mathbf{R}$ 的集函数 M 满足以下条件：

公理 1（规范性）　对于全集 Γ，有 $M\{\Gamma\}=1$；

公理 2（对偶性）　对于任意的事件 $\Lambda\in L$，有 $M\{\Lambda\}+M\{\Lambda^c\}=1$；

公理 3（次可列可加性）　对于可数的事件序列 $M\{\Lambda\}$，有

$$M\left\{\bigcup_{i=1}^{\infty}\Lambda_i\right\}\leqslant\sum_{i=1}^{\infty}M\{\Lambda_i\}$$

则称集函数 M 为 Γ 上的不确定测度，三元组 (Γ,L,M) 称为一个不确定空间.

为了研究乘积空间上的不确定测度，Liu B 提出了乘积公理[1]：

公理 4　设 (Γ_i,L_i,M_i) 为一列不确定空间，则乘积不确定测度 M 为

$$M\left\{\prod_{i=1}^{\infty}\Lambda_i\right\}=\mathop{\wedge}_{i=1}^{\infty}M_i\{\Lambda_i\}$$

其中 $\Lambda_i\in L_i$，$i=1,2,\cdots$. 称三元组 (Γ,L,M) 为乘积不确定空间，其中 $\Gamma=\Gamma_1\times\Gamma_2\times\cdots$，$L=L_1\times L_2\times\cdots$，$M=M_1\times M_2\times\cdots$.

定义 2　不确定变量 ξ 是从不确定空间 (Γ,L,M) 到实数集 $\mathbf{R}$ 的可测函数，即对任意的 Borel 实数集 B，集合

$$\{\xi\in B\}=\{\gamma\in\Gamma\,|\,\xi(\gamma)\in B\}$$

是一个事件.

不确定变量的定义是由抽象的不确定空间和 Borel 集描述的. 如果仅从定义出发，在理解和应用不确定变量时都会遇到困难，为了更好地理解不确定变量，给出如下不确定分布的概念.

定义 3　设 ξ 为不确定变量，函数

$$\Phi(x)=M(\xi\leqslant x)$$

称为 ξ 的不确定分布.

定义 4　若不确定变量 ξ 具有不确定分布

$$\Phi(x)=\begin{cases}0 & x<a\\ (x-a)/(b-a) & a\leqslant x\leqslant b\\ 1 & x>b\end{cases}$$

其中 a 和 b 为常数，则 ξ 为线性的不确定变量，记为 $L(a,b)$.

定义 5　若对于任意的 $\alpha\in(0,1)$，不确定变量 ξ 的不确定分布 $\Phi(x)$ 的反函数 $\Phi^{-1}(\alpha)$ 存在且唯一，则称 $\Phi(x)$ 为正则分布，称 ξ 为正则不确定变量.

定义 6　若不确定变量 ξ 具有正则分布 $\Phi(x)$，则其反函数 $\Phi^{-1}(\alpha)$ 称为 ξ 的逆不确定分布.

例 1　根据定义 4—定义 6，线性不确定变量 $L(a,b)$ 的逆不确定分布为

$$\Phi^{-1}(\alpha)=(1-\alpha)a+\alpha b$$

1.2　机会理论[6]

定义 7　设 $(\Gamma,L,M)\times(\Omega,A,Pr)$ 是一个机会空间，Θ 是 $L\times A$ 的一个事件，那么事件 Θ 的机会测度为

$$Ch\{\Theta\}=\int_0^1 Pr\{\omega\in\Omega\,|\,M\{\gamma\in\Gamma\,|\,(\gamma,\omega)\in\Theta\}\geqslant x\}\mathrm{d}x$$

定义 8　从机会空间 $(\Gamma,L,M)\times(\Omega,A,Pr)$ 到实数集 $\mathbf{R}$ 的可测函数 ξ 称为不确定随机变量，即对于任意的 Borel 实数集 B，集合

$$\{\xi\in B\}=\{(\gamma,\omega)\in\Gamma\times\Omega\,|\,\xi(\gamma,\omega)\in B\}$$

是 $L\times A$ 中的一个事件.

定义 9　设 $f:R^n\to R$ 是一个可测函数，$\xi_1,\xi_2,\cdots,\xi_n$ 是机会空间 $(\Gamma,L,M)\times(\Omega,A,Pr)$ 上的一列不确定随机变量，那么对任意的 $(\gamma,\omega)\in\Gamma\times\Omega$，$\xi=f(\xi_1,\xi_2,\cdots,\xi_n)$ 是由

$$\xi(\gamma,\omega)=f(\xi_1(\gamma,\omega),\xi_2(\gamma,\omega),\cdots,\xi_n(\gamma,\omega))$$

所确定的一个不确定随机变量.

定义 10　$\eta_1,\eta_2,\cdots,\eta_m$ 是一系列独立的随机变量，概率分布分别为 $\Psi_1,\Psi_2,\cdots,\Psi_m$，$\tau_1,\tau_2,\cdots,\tau_n$ 是一列不确定变量，不确定分布分别为 $\Upsilon_1,\Upsilon_2,\cdots,\Upsilon_n$，那么不确定随机变量

$$\xi=f(\eta_1,\eta_2,\cdots,\eta_m,\tau_1,\tau_2,\cdots\tau_n)$$

有一个机会分布

$$\Phi(x)=\int_{R^m}F(x,y_1,y_2,\cdots,y_m)\mathrm{d}\Psi_1(y_1)\Psi_2(y_2)\cdots\Psi_m(y_m)$$

其中，对任意的实数 $y_1,y_2,\cdots,y_m$，$F(x,y_1,y_2,\cdots,y_m)$ 是不确定变量 $f(\eta_1,\eta_2,\cdots,\eta_m,\tau_1,\tau_2,\cdots\tau_n)$ 的不确定分布，

它可由其反函数 $F^{-1}(\alpha;y_1,y_2,\cdots,y_m)=f(y_1,y_2,\cdots,y_m,\Upsilon_1^{-1}(\alpha),\Upsilon_2^{-1}(\alpha),\cdots,\Upsilon_n^{-1}(\alpha))$ 决定，条件是 f 为对于 $\tau_1,\tau_2,\cdots,\tau_n$ 的单增函数.

例 2　设 η 是一个随机变量，其概率分布为 Ψ，τ 是一个不确定变量，其不确定分布为 Υ. 那么，$\xi=\eta+\tau$ 是一个不确定随机变量，ξ 的机会分布为

$$\Phi(x)=Ch\{\xi\leqslant x\}=\int_{-\infty}^{+\infty}M\{y+\tau\leqslant x\}\mathrm{d}\Psi(y)=\int_{-\infty}^{+\infty}\Upsilon(x-y)\mathrm{d}\Psi(y)$$

其中，y 是随机变量 η 的任一实现.

2　不确定随机网络下的 Top-k 查询算法

2.1　不确定随机网络

N 是节点集合，U 是不确定弧的集合，R 是随机弧的集合，W 是不确定权重和随机权重的集合，那么四元组 (N,U,R,W) 被称为一个不确定随机网络.

例 3　图 1 是有 4 个节点的不确定随机网络. 其中，随机权重 η_{AB}、η_{CD} 的概率分布分别为 Ψ_{AB}、Ψ_{CD}. 不确定权重 τ_{BC} 具有正则的不确定分布 Υ_{BC}. 各边的权重的分布函数见表 1.

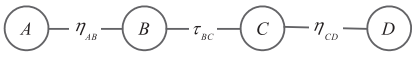

图 1　4 节点不确定随机网络

Fig. 1　Uncertain random network with 4 nodes

表 1　图 1 网络中各边的权重的分布函数

Tab. 1　Distribution function of the edge's weight of figure 1

边	分布函数	边	分布函数
(A,B)	$U(4,8)$	(C,D)	$U(7,9)$
(B,C)	$L(6,12)$		

注：U 为均匀概率分布，L 为线性不确定分布.

Ψ_{AB}，Ψ_{CD} 的概率分布函数分别为

$$\Psi_{AB}(y_{AB})=\begin{cases}0 & y_{AB}<4\\ (y_{AB}-4)/4 & 4\leqslant y_{AB}<8\\ 1 & y_{AB}\geqslant 8\end{cases}$$

$$\Psi_{CD}(y_{CD})=\begin{cases}0 & y_{CD}<7\\ (y_{34}-7)/4 & 7\leqslant y_{CD}<9\\ 1 & y_{CD}\geqslant 9\end{cases}$$

Υ_{BC} 的不确定分布为

$$\Upsilon_{BC}(\tau)=\frac{\tau}{6}-1$$

则权重的机会分布函数为

$$\Phi(x)=\int_4^8\int_7^9\Upsilon(x-y_{AB}-y_{CD})\mathrm{d}\Psi_{AB}(y_{AB})\mathrm{d}\Psi_{CD}(y_{CD})$$

计算可得

$$\Phi(x)=\begin{cases}0 & x<20\\ (x-20)/6 & 20\leqslant x<26\\ 1 & x\geqslant 26\end{cases}$$

假设机会测度 $\Phi(x)=0.95$，计算得权重 $x=25.7$.

2.2　Top-k 最近节点查询算法

给定不确定随机网络 G，节点间的权重的机会分布函数为 Φ，则节点间的路径长度机会分布函数 $D=\Phi$，即将节点间的权重建模为节点间的路径长度. 基于节点间的路径长度，提出以下的不确定随机网络 Top-k 最近节点查询算法.

输入：不确定随机网络 $G=(N,U,R,W)$，选择的最近节点个数 k，初始节点 p，机会测度 α.

输出：当机会测度为 α 时，与 p 最近的 k 个节点，以及对应路径.

具体算法如下：

(1) 计算所有节点间 i,j 间的路径，初始节点为：$i=1$，$j=2$，并将 i 标记为已访问；

(2) 从 N 中取出非 i 节点 k，若两点间有路径，将 k 标记为已访问；

(3) 若 $k=j$，将所有已访问节点组成的路径放入路径集合 D 中，回溯至上一个已访问节点，重复步骤(2)；否则，重复步骤(2)；

(4) 若 $k\geqslant n$，$j++$，重复步骤(1)—(3)，直至 $j=n$；

(5) 若 $j>n$，$j--$，$i++$，重复步骤(1)—(4)，直至 $i=n$；

(6) 计算 D 中所有路径的机会分布函数；

(7) 计算当机会测度为 α 时，点 p 到所有节点的路径长度；

(8) 找到距离点 p 最近的 k 个点及对应路径，算法结束.

算法中步骤(2)—(3)部分的代码如下：

```
function FindPath(matrix, startNode, endNode) {
  result[nPos] = startNode.key; //将当前节点放入结果集
  Mark[startNode.key] = true; //标记为已访问
  nPos++;
  while(nPos != 0) {
    var tempVal = result[nPos–1]; //获取结果集最后
```

```
一个元素
if(tempVal = = endNode.key){//如果当前节点为结束节点,将结果集中的节点放入路径结果集中
  for(let j = 0; j < nPos; j + + ){
    resultPath[pathNum][j] = result[j];
  }
  nPos- -; //回溯至目标节点的上一个节点
  result[nPos] = 0;
  pathNum + + ; //新增路径数目
  Mark[endNode.key] = false;
  break;
};
while(startNode.flag < matrix.length){
  if(matrix[tempVal][startNode.flag] = = 1){
    if(Mark[startNode.flag] = = false){
      var tempNode = new Node();
      tempNode.key = startNode.flag;
      tempNode.flag = 0;
      FindPath(matrix, tempNode, endNode);
    }
  }
  startNode.flag + + ;
}
if(startNode.flag = = matrix.length){
  nPos- -;
  startNode.flag = 0;
  Mark[startNode.key] = false;
  break;
}
}
```

为了更好地理解算法，用例 4 进行简要说明.

例 4　有 5 个节点的不确定随机网络，如图 2 所示. 其中 τ_{AB}、τ_{BD}、τ_{DE} 是不确定权重，且分别具有正则的不确定分布 Υ_{AB}、Υ_{BD}、Υ_{DE}；η_{AE}、η_{BC}、η_{CD} 是随机权重，分别具有概率分布 Ψ_{AE}、Ψ_{BC}、Ψ_{CD}，网络中各边的权重的分布函数见表 2. 求距离节点 A 最近的 3 个节点.

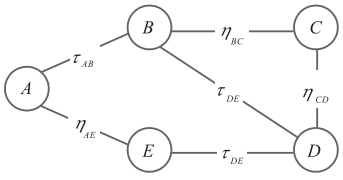

图 2　5 节点不确定随机网络

Fig. 2　Uncertain random network with 5 nodes

表 2　图 2 网络中各边的权重的分布函数

Tab. 2　Distribution function of the edge's weight of figure 2

边	分布函数	边	分布函数
(A,B)	$L(2,7)$	(A,E)	$U(3,6)$
(B,D)	$L(5,8)$	(B,C)	$U(6,8)$
(D,E)	$L(7,12)$	(C,D)	$U(5,11)$

注：U 为均匀概率分布；L 为线性不确定分布.

能够得到任意两点间的路径长度机会分布函数

$$D(x)=\int_0^{+\infty}\cdots\int_0^{+\infty}F(x;y_{ij},(i,j)\in R)\prod_{(i,j\in R)}\mathrm{d}\Psi_{ij}(y_{ij})$$

其中 $F\left(x,y_{ij},(i,j)\in R\right)$ 由它的逆不确定分布确定：

$$F^{-1}(\alpha;y_{ij},(i,j)\in R)=f(c_{ij},(i,j)\in U\bigcup R)$$

当机会测度 $\alpha=0.9$ 时，计算出任意两点间的最小路径长度，计算结果见表 3.

表 3　$\alpha=0.9$ 时节点间路径长度

Tab. 3　Path length between nodes when $\alpha=0.9$

节点	路径长度				
	A	*B*	*C*	*D*	*E*
A		6.5	13.5	14.2	5.7
B	6.5		7.8	7.7	9.33
C	13.5	7.8		10.4	13.7
D	14.2	7.7	10.4		12.4
E	5.7	9.33	15.75	11.5	

当机会测度 $\alpha=0.8$ 时，计算出任意两点间的最小路径长度，计算结果见表 4.

表 4　$\alpha=0.8$ 时节点间路径长度

Tab. 4　Path length between nodes when $\alpha=0.8$

节点	路径长度				
	A	*B*	*C*	*D*	*E*
A		6	13	12.4	5.4
B	6		7.6	7.4	8.83
C	13	7.6		9.8	15.67
D	12.4	7.4	9.8		11
E	5.4	8.83	15.67	11	

所以，当机会测度 $\alpha=0.9$，$k=3$ 时，距离节点 A 最近的 3 个点分别为 E、B、C. 当机会测度 $\alpha=0.8$，$k=3$ 时，距离节点 A 最近的 3 个点分别为 E、B、D.

当机会测度变化时，查询到的最短路径节点也可能发生变化，所以需要根据现实需求来指定机会测度，以得到预期结果.

3　结　语

本文提出了不确定随机网络的 Top-k 最近节点查询问题，并使用机会理论求解该问题，设计不确定随机网络在一定机会测度条件下的 Top-k 查询算

法. 此算法能正确求解不确定随机网络的 Top-k 查询问题，在网络节点数目较小、不确定随机分布较简单时的效率较高；一旦网络节点数目众多、网络非常复杂时，则计算的时间复杂度会较高. 如何对算法进行改进，以提高算法的计算效率是下一步的研究方向.

参考文献：

[1] Liu B. Uncertainty Theory[M]. 2nd ed. Berlin：Springer Verlag，2007.

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

[3] 高原. 不确定图与不确定网络[D]. 北京：清华大学，2013.

[4] Zhou J，Yang F，Wang K. An inverse shortest path problem on an uncertain graph[J]. Journal of Networks，2014，9(9)：2353–2359.

[5] Zhou J，Chen L，Wang K. Path optimality conditions for minimum spanning tree problem with uncertain edge weights[J]. International Journal of Uncertainty，Fuzziness and Knowledge-Based Systems，2015，23(1)：49–71.

[6] Liu Y. Uncertain random variables：A mixture of uncertainty and randomness[J]. Soft Computing，2013，17(4)：625–634.

[7] Liu B. Uncertain random graph and uncertain random network[J]. Journal of Uncertain Systems，2014，8(1)：2592–2599.

[8] 盛玉红. 不确定随机网络优化[D]. 北京：清华大学，2015.

[9] Soliman M A，Ilyas I F，Chang K C C. Top-k query processing in uncertain databases[C]//Proceedings of the 23rd IEEE International Conference on Data Engineering. Piscataway：IEEE，2007：896–905.

[10] Lian X，Chen L. Probabilistic ranked queries in uncertain databases[C]//Proceedings of the 11th International Conference on Extending Database Technology. New York：Association for Computing Machinery，2008：511–522.

[11] Hua M，Pei J，Liu X. Ranking queries on uncertain data[J]. The International Journal on Very Large Databases，2011，20(1)：129–153.

[12] Zhang X，Chomicki J. On the semantics and evaluation of Top-k queries in probabilistic databases[C]//Proceedings of the 24th IEEE International Conference on Data Engineering. Piscataway：IEEE，2008：556–563.

[13] Li J，Saha B，Deshpande A. An unified approach to ranking in probabilistic databases[J]. The VLDB Journal，2011，20(2)：249–275.

[14] 郭长友，郑雪峰，高秀莲. 基于不确定理论的不确定性数据 Top-k 查询计算[J]. 计算机科学，2016，34(3)：225–230.

责任编辑：常涛

(上接第 13 页)

[132] Ariyanti D，Mills L，Dong J，et al. $NaBH_4$ modified TiO_2：Defect site enhancement related to its photocatalytic activity[J]. Materials Chemistry and Physics，2017，199：571–576.

[133] Wang W，Lu C，Ni Y，et al. Enhanced visible light photoactivity of{001} facets dominated TiO_2 nanosheets with even distributed bulk oxygen vacancy and Ti^{3+}[J]. Catalysis Communications，2012，22：19–23.

[134] Wu Z，Cao S，Zhang C，et al. Effect of bulk and surface defects on the photocatalytic performance of size-controlled TiO_2 nanoparticles[J]. Nanotechnology，2017，28(27)：275706.

[135] Kong M，Li Y，Chen X，et al. Tuning the relative concentration ratio of bulk defects to surface defects in TiO_2 nanocrystals leads to high photocatalytic efficiency[J]. Journal of the American Chemical Society，2011，133(41)：16414–16417.

责任编辑：常涛