第 41 卷 第 1 期 计 算 机 工 程 2015 年 1 月
Vol. 41 No. 1 Computer Engineering January 2015

·先进计算与数据处理· 文章编号:1000-3428(2015)01-0054-07 文献标识码:A 中图分类号:TP311.13

受限模糊网络可信近邻查询

高 峻[1],郝忠孝[1,2]

(1. 哈尔滨理工大学计算机科学与技术学院,哈尔滨 150080; 2. 哈尔滨工业大学计算机科学与技术学院,哈尔滨 150001)

摘 要:针对不确定网络环境下的近邻查询问题,给出一种新的解决方法。将不确定网络建模为模糊图,定义模糊图中两点间的可信最短路径距离和可信最短路径期望距离,在可信距离基础上,提出模糊图可信近邻查询概念,并给出网络距离受限条件下的模糊图可信近邻查询算法和即时可信近邻查询算法。算法采用模糊模拟方法降低问题难度,使用网络距离约束缩小搜索空间,运用优先队列快速得到满足精度 ε 要求的可信近邻查询结果。算法的时间复杂度分别为 $O((2r+\Delta r)(e+n\lg n)+h\lg h+\lg n)$ 和 $O(e+(n+1)\lg n)$。理论分析与实验结果表明,可信近邻查询算法能够从模糊角度解决不确定网络环境下的近邻查询问题。

关键词:不确定网络;模糊图;可信距离;可信近邻;模糊模拟;距离约束

中文引用格式:高 峻,郝忠孝. 受限模糊网络可信近邻查询[J]. 计算机工程,2015,41(1):54-60.

英文引用格式:Gao Jun,Hao Zhongxiao. Credible Nearest Neighbor Query in Constraint Fuzzy Network[J]. Computer Engineering,2015,41(1):54-60.

Credible Nearest Neighbor Query in Constraint Fuzzy Network

GAO Jun[1],HAO Zhongxiao[1,2]

(1. College of Computer Science and Technology,Harbin University of Science and Technology,Harbin 150080,China;
2. College of Computer Science and Technology,Harbin Institute of Technology,Harbin 150001,China)

【Abstract】 This paper gives a new method for solving the problem of the nearest neighbor query in uncertain network. The method is that,gives the definition of credible shortest path distance and credible shortest path expectation distance,based on credible distance,puts forward the concept of fuzzy graph credible nearest neighbor query,and proposes fuzzy graph credible nearest neighbor query algorithm and instant credible nearest neighbor query algorithm on the condition of network distance constraint. The algorithm decreases the difficulty of the problem by using fuzzy analog,diminishes search space by using network distance constraint and quickly acquires the result of credible nearest neighbor query according to the precision ε by using the priority queue. The complexity of the algorithm is $O((2r+\Delta r)(e+n\lg n)+h\lg h+\lg n)$ and $O(e+(n+1)\lg n)$. Theory analysis and experimental results show that fuzzy graph credible nearest neighbor query algorithm can solve the problem of the nearest neighbor query in uncertain network at the angle of fuzzy quality.

【Key words】 uncertain network;fuzzy graph;credible distance;credible nearest neighbor;fuzzy analog;distance constraint

DOI:10.3969/j.issn.1000-3428.2015.01.010

1 概述

不确定性数据处理是数据库查询领域的研究热点,目前已取得很多成果,但这些研究多集中在实体数据的不确定性,在现实中覆盖还不够全面。现实中存在实体间关系的不确定性,如路网结点间路径有时关闭,有时开放,多数是根据车流量而变化的介于关闭和开放间的一个不确定量,这时根据结点间最短路径距离得到的最近邻查询结果并不能保证是最有效近邻。这就需要考虑在不确定网络环境下的近邻查询问题。

对于实体数据的不确定性处理,已有很多好的方法,如文献[1]给出 SQL 查询方法,文献[2]给出一种索引方法,文献[3-4]给出 k-近邻查询方法,文献[5]给出一种范围查询分析方法。但这些方法并不能直接用于实体间关系的不确定性处理。

已有文献关注这个问题,其中,文献[6]分析了社会网络环境的不确定性、查询需求并给出处理方

基金项目:黑龙江省自然科学基金资助项目(F200821)。

作者简介:高 峻(1972-),女,副教授、博士研究生,主研方向:时空数据库技术;郝忠孝,教授、博士生导师。

收稿日期:2013-09-24 **修回日期**:2013-12-26 **E-mail**:hustgj@163.com

法。文献[7]给出了移动网络环境使用 k 近邻查询解决概率路径问题的方法。文献[8]表明生物网络环境也有这样的查询需求。文献[9]使用随机理论给出了不确定网络环境下近邻查询处理方法。将不确定网络环境建模为概率加权图,实体间关系的不确定性用固定概率值表示,定义了概率加权图中结点间各种距离,并基于这些距离进行近邻查询。这些工作都将实体间关系的不确定性建模为概率性,但现实中实体间关系的不确定性有时还表现为模糊性,如路网结点间的畅通程度。

模糊集理论是处理不确定性问题的有力工具,已被广泛用于数据类型[10]和运算定义[11]以及空间查询[12]等领域。本文将探讨使用模糊集理论来处理不确定网络环境下的近邻查询问题。

2　相关概念

模糊集理论在许多实际领域已得到应用。为了度量模糊事件,文献[13]提出了可能性测度,文献[14]提出了必要性测度,之后文献[15]提出了可信性测度。可信性测度被认为是与概率论中的概率测度平行的概念。

定义 1(模糊网络)　模糊网络是指在模糊集理论框架下讨论的不确定网络,即网络实体确定而实体间关系的不确定表现为实体间关系的模糊性。

将模糊网络建模为图,网络实体表示为图中结点,实体间关系表示为图中边,实体间关系是模糊的,则得到模糊图,图中边给出可信值,表示边存在的可信度,即实体间关系的可信度。模糊图与普通图类似,可分为有向图与无向图,加权图与无权图。这里讨论的是无向加权模糊图。下面给出模糊图的形式定义。

定义 2(模糊图)　设 $\tilde{G}=(V,E,W,Cr)$ 表示模糊图,其中,V 表示图的结点集;E 表示图的边集;W 表示边的权重集;Cr 表示边存在的可信度集。$w(e)$ 表示边 e 的权重,$cr(e)$ 表示边 e 存在的可信度。$cr(e)>0$,当且仅当 $e\in E$。

图 1 为模糊图示例。

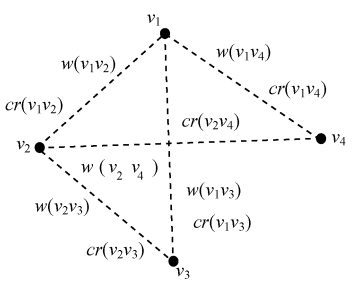

图 1　模糊图示例

设模糊图 $\tilde{G}=(V,E,W,Cr)$,其中,$V=\{v_1,v_2,v_3,v_4\}$,$E=\{v_1v_2,v_1v_3,v_1v_4,v_2v_3,v_2v_4\}$,$W=\{w(v_1v_2),w(v_1v_3),w(v_1v_4),w(v_2v_3),w(v_2v_4)\}$,$Cr=\{cr(v_1v_2),cr(v_1v_3),cr(v_1v_4),cr(v_2v_3),cr(v_2v_4)\}$。

定义 3(样本图)　已知模糊图 $\tilde{G}=(V,E,W,Cr)$,样本图 G 是模糊图 $\tilde{G}$ 的一个实例,E_G 为 G 的边集,则其可信度 $cr(G)=\prod_{e\in E_G}cr(e)$。

图 2 是图 1 所示模糊图 $\tilde{G}$ 的样本图示例,其中,图 2 (a)为样本图 G_1,$E_{G1}=\{v_1v_2,v_1v_4,v_2v_3\}$,$G_1$ 的可信度为 $cr(G)=cr(v_1v_2)\times cr(v_1v_4)\times cr(v_2v_3)=0.3\times0.7\times0.6=0.125$。

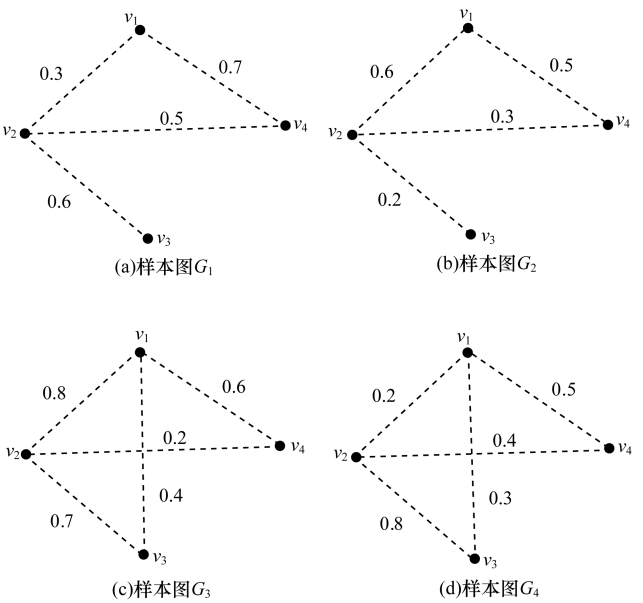

图 2　样本图示例

设模糊图 $\tilde{G}=(V,E,W,Cr)$,G 是 $\tilde{G}$ 的样本图,$v_i,v_j\in V$,将图中两点间的最小权重表达为两点间的最短路径距离。

定义 4　样本图中两点间最短路径距离:$d_G(v_i,v_j)=d$ 表示 v_i,v_j 在 G 中最短路径距离为 d,E_d 为 G 中 v_i,v_j 间最短路径边集,$d_G(v_i,v_j)$ 为 d 的可信度 $cr(d_G(v_i,v_j)=d)=\prod_{e\in E_d}cr(e)$。

在图 2(a)所示样本图 G_1 中,设 v_1,v_2 和 v_2,v_3 间距离均为 1,则 v_1,v_3 间最短路径距离 $d_{G_1}(v_1,v_3)=2$,其边集 $E_2=\{v_1v_2,v_2v_3\}$,$d_{G_1}(v_1,v_3)=2$ 的可信度为:

$$cr(d_{G_1}(v_1,v_3)=2)=cr(v_1v_2)\times cr(v_2v_3)=0.3\times0.6=0.18$$

在图 2(b)所示样本图 G_2 中,设 v_1,v_2 和 v_2,v_3 间距离均为 1,则 v_1,v_3 间最短路径距离 $d_{G_2}(v_1,v_3)=2$,其边集 $E_2=\{v_1v_2,v_2v_3\}$,$d_{G_2}(v_1,v_3)=2$ 的可信度为:

$$cr(d_{G_2}(v_1,v_3)=2)=cr(v_1v_2)\times cr(v_2v_3)=0.6\times0.2=0.12$$

在图 2(c)所示样本图 G_3 中,设 v_1 和 v_3 间距离为 1,则 v_1,v_3 间最短路径距离 $d_{G_3}(v_1,v_3)=1$,其边集 $E_1=\{v_1v_3\}$,$d_{G_3}(v_1,v_3)=1$ 的可信度为:

$$cr(d_{G_3}(v_1,v_3)=1)=cr(v_1v_3)=0.4$$

在图 2(d)所示样本图 G_4 中,设 v_1 和 v_3 间距离为 1,则 v_1,v_3 间最短路径距离 $d_G(v_1,v_3)=1$,其边集 $E_1=\{v_1v_3\}$,$d_{G_4}(v_1,v_3)=1$ 的可信度为:

$$cr(d_{G_4}(v_1,v_3)=1)=cr(v_1v_3)=0.3$$

定义 5 模糊图中两点间最短路径距离:模糊图 $\tilde{G}$ 中 v_i,v_j 间最短路径距离记为 $d(v_i,v_j)$,$d(v_i,v_j)$ 为 d 的可信度为:

$$cr(d(v_i,v_j)=d)=\sum_G cr(d_G(v_i,v_j)=d)$$

设图 2 所示为模糊图 $\tilde{G}$ 的所有样本图,其中满足 v_1,v_3 间最短路径距离 $d_G(v_1,v_3)=2$ 的样本图为图 2(a)、图 2(b),则模糊图 $\tilde{G}$ 中 v_1,v_3 间最短路径距离 $d(v_1,v_3)=2$ 的可信度为:

$$cr(d(v_1,v_3)=2)=cr(d_{G_1}(v_1,v_3)=2)+$$
$$cr(d_{G_2}(v_1,v_3)=2)=0.12+0.18=0.3$$

满足 v_1,v_3 间最短路径距离 $d_G(v_1,v_3)=1$ 的样本图为(c)(d),则模糊图 $\tilde{G}$ 中 v_1,v_3 间最短路径距离 $d(v_1,v_3)=1$ 的可信度为:

$$cr(d(v_1,v_3)=1)=cr(d_{G_3}(v_1,v_3)=1)+$$
$$cr(d_{G_4}(v_1,v_3)=1)=0.4+0.3=0.7$$

定义 6 可信最短路径距离:v_i,v_j 间可信最短路径距离 $d_C(v_i,v_j)=\arg(\max_d cr(d(v_i,v_j)=d))$。

根据上述假设与计算可知,图 1 所示模糊图 $\tilde{G}$ 中 v_1,v_3 间可信最短路径距离为:

$$d_C(v_1,v_3)=\arg(\max_d cr(d(v_1,v_3)=d))=$$
$$\arg\max_d\{cr(d(v_1,v_3)=2),cr(d(v_1,v_3)=1)\}=$$
$$\arg\max_d\{0.3,0.7\}=1$$

定义 7 可信最短路径期望距离:v_i,v_j 间可信最短路径期望距离 $d_{CE}(v_i,v_j)=\sum_d d\times cr(d(v_i,v_j)=d)$。

在图 1 所示模糊图 $\tilde{G}$ 中,v_1,v_3 间最短路径期望距离为:

$$d_{CE}(v_1,v_3)=\sum_d d\times cr(d(v_1,v_3)=d)=$$
$$2\times cr(d(v_1,v_3)=2)+1\times cr(d(v_1,v_3)=1)=$$
$$2\times 0.3+1\times 0.7=1.3$$

3 模糊图的存储方法

模糊图 $\tilde{G}=(V,E,W,Cr)$ 中的 V,E,W 相对稳定,而 Cr 则变化较大,因此在存储时分开存储,用指针相连。使用邻接表对模糊图进行表示。模糊图 $\tilde{G}=(V,E,W,Cr)$ 的邻接表表示由一个包含 $|V|$ 个列表的数组所组成,其中每个列表对应于 V 中的一个顶点。对于每一个 $v_i\in V$,邻接表 l_i 包含所有满足条件 $v_iv_j\in V$ 的结点 v_j。邻接表形式为 $l_i:<v_j,w(v_iv_j),pt_1,pt_2,next>$。其中,$l_i$ 表示结点 v_i 的邻接表;v_j 表示 v_i 的邻接点;$w(v_iv_j)$ 表示 v_iv_j 的权重,在图中简记为 w_{ij};pt_1 为指向 v_iv_j 具体抽象数据类型所在页指针,而抽象数据类型的前后分别设置指针指向结点邻接表所在页;pt_2 为指向邻接表中动态部分存储位置的指针,其存储格式为三元数组,数组第一个位置表示样本图号,数组第 2 个位置表示 v_j 和 v_i 在样本图中是否邻接的逻辑值,若在样本图中两结点邻接,则逻辑值为 1,否则逻辑值为 0,数组第 3 个位置表示 v_iv_j 在样本图中的可信度。$next[v_j]$ 为指向 v_j 的后继元素指针,若 $next[v_j]$ 指向 *null*,则 v_j 没有后继元素,它是尾。

图 1、图 2 所示的模糊图的样本图的表示形式如图 3 所示。

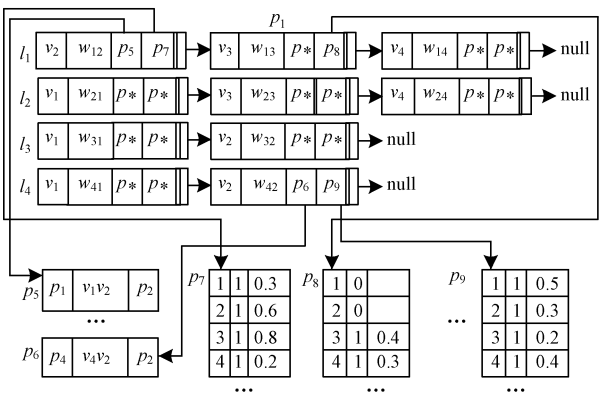

图 3 模糊图的表示形式

以结点 v_1 为例。v_1 有 3 个邻接点 v_2,v_3 和 v_4,则其邻接表 l_1 包括形式相同的 3 项。以第一项为例,其形式为 $<v_2,w_{12},p_5,p_7,next>$,其中,$v_2$ 是 v_1 的邻接点;w_{12} 表示 v_1v_2 的权重;p_5 为边 v_1v_2 的具体抽象数据类型所在页,而 v_1v_2 的具体抽象数据类型的前后的 p_1 和 p_2 分别为结点 v_1 和 v_2 的邻接表所在页;p_7 为边 v_1v_2 的动态信息所在页,$next[v_2]$ 指向后继元素 v_3。p_7 中第一行 $<1,1,0.3>$ 为样本图 G_1 中 v_1v_2 的动态信息,第一位置的 1 表示样本图为 G_1,第二位置的 1 表示在样本图 G_1 中 v_1v_2 连接,第三位置的 0.3 表示在样本图 G_1 中 v_1v_2 连接的可信度,其他行同理,分别表示样本图 G_2,G_3 和 G_4 中 v_1v_2 的动态信息。

给定一个新的结点 v_k,不妨设其邻接点为 $v_i\cdots v_j$,过程 Adjacent-insert 将 v_k 加入到邻接表中。

Adjacent-insert (L,v_k):

(1)建立 v_k 邻接表 l_k,包含项 $<v_i,w_{ki},p_*,p_*,next>\cdots<v_j,w_{kj},p_*,p_*,next>$;

(2)对邻接表 $l_i\cdots l_i$,增加项 $<v_k,w_{k*},p_*,p_*,next>$,原尾部指针指向新加项,而 $next[v_k]$ 指向 *null*;

给定一个已有结点 v_k,不妨设其邻接点为 $v_i\cdots v_j$,过程 Adjacent-delete 将从邻接表中删除 v_k。

Adjacent-delete (L,v_k):

(1)删除 v_k 邻接表 l_k;

(2)在邻接表 $l_i\cdots l_i$ 中,若项 $<v_k,w_{k*},p_*,p_*,next>$ 位于邻接表头部,则直接删除;若其位于邻接表中部,则将其前序结点的 *next* 指针指向其后继结点,删除此项;若其位于邻接表尾部,则将其前序结点的 *next* 指针指向 *null*,删除此项。

4　可信距离计算

模糊图中可信距离的精确计算代价较高,需在所有可能样本图中计算点间的最短路径距离和点间路径存在的可信度,复杂度达到指数级。这里使用对模糊系统进行抽样实验的模糊模拟技术进行近似计算。模糊模拟技术没有随机模拟技术成熟,没有大数定理来保证抽样结果的精确度,但可将模糊理论与模糊模拟结合使用来得到满意的近似值。

定义 8　设 $\xi_1,\xi_2,\cdots,\xi_r$ 是独立同分布的模糊变量,有相同有限期望值。若 $\forall\varepsilon>0,\lim\limits_{i\to\infty}Cr\{|\xi_i-\xi|\geqslant\varepsilon\}=0$,则称模糊变量序列 $\{\xi_i\}$ 依可信性收敛到 ξ[16]。

定义 9　设 $\xi_1,\xi_2,\cdots,\xi_r$ 是独立同分布的模糊变量,有相同有限期望值。若 $\lim\limits_{i\to\infty}E[|\xi_i-\xi|]=0$,则称模糊变量序列 $\{\xi_i\}$ 依均值收敛到 ξ[16]。

4.1　可信最短路径距离计算

根据定义 8 可通过过程 Compute-$d_C(q,v)$ 得到可信最短路径距离的近似值。

Compute-$d_C(q,v)$:

(1)在模糊图中均匀抽取 r 个样本,计算每个样本图中 v_i,v_j 两点间最短路径距离 d 及其可信度,得到模糊图中 v_i,v_j 两点间最短路径距离为 d 的近似可信度:

$$cr_r(d(v_i,v_j)=d)=\sum_{l=1}^{r}cr(d_{G_l}(v_i,v_j)=d)$$

(2)在模糊图中均匀抽取 $r+\Delta r$ 个样本,计算每个样本图中 v_i,v_j 两点间最短路径距离 d 及其可信度,得到模糊图中 v_i,v_j 两点间最短路径距离为 d 的近似可信度:

$$cr_{r+\Delta r}(d(v_i,v_j)=d)=\sum_{l=1}^{r+\Delta r}cr(d_{G_l}(v_i,v_j)=d)$$

(3)若对精度 $\varepsilon>0$

$$|cr_{r+\Delta r}(d(v_i,v_j)=d)-cr_r(d(v_i,v_j)=d)|\leqslant\varepsilon$$

则 $cr_{r+\Delta r}(d(v_i,v_j)=d)$ 即作为模糊图中 v_i,v_j 两点间最短路径距离为 d 的近似可信度。否则令 $r=r+\Delta r$,转(2)。

(4)比较 v_i,v_j 两点间最短路径距离的近似可信度,其中使近似可信度最大的 d 值即为 v_i,v_j 两点间可信最短路径距离 $d_C(q,v)$ 的近似值。

定理 1　设模糊图中两点间最短路径距离的可信度是收敛的,则对任意 $\varepsilon>0$,当 $r\to\infty$ 时,有:

$$|cr_{r+\Delta r}(d(v_i,v_j)=d)-cr_r(d(v_i,v_j)=d)|\leqslant\varepsilon$$

证明:设模糊图中两点间最短路径距离的可信度收敛到 c。则根据定义 8 知,$\lim cr_r(d(v_i,v_j)=d)=c$,即对任意 $\frac{\varepsilon}{2}>0$,$r\to\infty$ 时,$|cr_r(d(v_i,v_j)=d)-c|\leqslant\frac{\varepsilon}{2}$。而:

$$\begin{aligned}&|cr_{r+\Delta r}(d(v_i,v_j)=d)-cr_r(d(v_i,v_j)=d)|=\\&|\{cr_{r+\Delta r}(d(v_i,v_j)=d)-c\}-\{cr_r(d(v_i,v_j)=d)-c\}|\\&|cr_{r+\Delta r}(d(v_i,v_j)=d)-c|+|cr_r(d(v_i,v_j)=d)-c|\leqslant\\&\frac{\varepsilon}{2}+\frac{\varepsilon}{2}=\varepsilon\leqslant=\varepsilon\end{aligned}$$

证毕。

在模糊模拟理论中,因模糊变量收敛,所以模拟结果应与其极限值接近,但其极限值无法实际得到,因此使用两次模拟结果的差值来作为模拟结束的条件。

根据定理 1 和模糊模拟理论可知,若对任意 $\varepsilon>0$,$|cr_{r+\Delta r}(d(v_i,v_j)=d)-cr_r(d(v_i,v_j)=d)|\leqslant\varepsilon$,则 $cr_{r+\Delta r}(d(v_i,v_j)=d)$ 即可作为模糊图中 v_i,v_j 两点间最短路径距离为 d 的精度为 ε 的近似可信度。

假设模糊图的基图结点数为 n,边数为 e。抽取 r 个样本,每个样本图中,v_i,v_j 两点间最短路径距离的计算使用 Dijkstra 算法,时间复杂度为 $O(e+n\lg n)$,计算模糊图中 v_i,v_j 两点间最短路径距离为 d 的近似可信度的时间复杂度为 r,即第一步的时间复杂度为 $O(r(e+n\lg n))$,第 2 步同理,时间复杂度为 $O((r+\Delta r)(e+n\lg n))$,第 3 步时间复杂度为常数,第 4 步若设 v_i,v_j 两点间最短路径距离的可能值有 h 个,则其时间复杂度 $O(h\lg h))$,即可信最短路径距离计算的时间复杂度为 $O((2r+\Delta r)(e+n\lg n)+h\lg h)$。

4.2　可信最短路径期望距离计算

根据定义 9 可通过过程 Compute-$d_{CE}(q,v)$ 得到可信最短路径期望距离的近似值。过程 Compute-$d_{CE}(q,v)$ 与 Compute-$d_C(q,v)$ 类似。

定理 2 设模糊图中两点间可信最短路径期望距离是收敛的,则对 $\varepsilon>0$,当 $r\to\infty$ 时,有:

$$|d_{CE}(v_i,v_j)_{r+\Delta r}-d_{CE}(v_i,v_j)_r|\leqslant\varepsilon$$

定理 2 的证明与定理 1 的证明类似。同理,若对任意 $\varepsilon>0$,有 $|d_{CE}(v_i,v_j)_{r+\Delta r}-d_{CE}(v_i,v_j)_r|\leqslant\varepsilon$,则 $d_{CE}(v_i,v_j)_{r+\Delta r}$ 即可作为模糊图中 v_i,v_j 两点间可信最短路径期望距离的精度为 ε 的近似值。

可信最短路径期望距离计算的时间复杂度也为 $O((2r+\Delta r)(e+n\lg n)+h\lg h)$。这个关于结点数和边数的多项式近似解法比精确指数解法可行,且精度可以根据需要调节。

5 可信近邻查询

定义 10 可信 k 近邻查询:已知模糊图 $\tilde{G}=(V,E,W,Cr)$,查询对象 $q\in V$,可信 k 近邻查询给出与 q 的可信最短路径距离(或可信最短路径期望距离)最小的 V 中 k 个对象,即对 $\forall v\in V$,有:

$$NN_C=\{v_{i_m}\mid d_C(q,v_{i_m})\leqslant d_C(q,v),m=1,2,\cdots,k\}$$

近邻查询是要找到与查询对象距离较近的目标对象,若结果与查询对象距离较远,则失去意义,因此这里仅考虑距离受限的近邻查询,从而缩小搜索空间。

算法思想:算法中设置了一结点集 NN_C,从查询点 q 到 NN_C 中结点的可信最短路径距离均已确定。算法反复计算访问结点 v 与查询点 q 的可信最短路径距离,若小于 NN_C 中结点的可信最短路径距离,则将 v 加入 NN_C 中。

以可信最短路径距离为基础的可信 k 近邻查询算法描述如算法 1 所示。

算法 1 可信 k 近邻查询算法

输入 模糊图 $\tilde{G}=(V,E,W,Cr)$,查询点 $q\in V$,样本数初值 r,样本数增量 Δr,近邻数 k,距离约束 D,精度 ε

输出 *k-NN* 查询结果集 NN_C

(1) $NN_C=\phi$;
(2) 设 $d_{C\text{-max}}$ 为 NN_C 中结点的最大可信最短路径距离;
(3) 从 q 出发执行 Dijkstra 算法直到达到距离约束 D;
(4) for 每个访问的结点 $v\in V$ do
(5) if $|NN_C|<k$ then $NN_C=NN_C\cup\{v\}$;
(6) else 调用过程 Compute-$d_C(q,v)$ 得到 $d_C(q,v)$;
(7) if $d_C(q,v)<d_{C\text{-max}}$ then $NN_C=NN_C\cup\{v\}$;
(8) else 舍去;
(9) return NN_C。

定理 3 模糊图 $\tilde{G}=(V,E,W,Cr)$,查询点 $q\in V$。对该图运行可信 k 近邻查询算法,则在算法终止时,对在距离约束 D 范围内的所有 $v\in V$,有:

$$NN_C=\{v_{i_m}\mid d_C(q,v_{i_m})\leqslant d_C(q,v),m=1,2,\cdots,k\}$$

证明:初始化:初始时,NN_C 为空集,显然成立。

保持:$d_{C\text{-max}}$ 为 NN_C 中结点与 q 的最大可信最短路径距离,当 NN_C 为空集时,$d_{C\text{-max}}$ 为无穷大。NN_C 中元素个数未达到 k 个时,则将 v 直接加入 NN_C,否则比较 v 与 q 的 $d_C(q,v)$ 和 $d_{C\text{-max}}$。若小于 $d_{C\text{-max}}$,则放入 NN_C,大于 $d_{C\text{-max}}$,则舍去。这就使 NN_C 中结点是当前已访问结点中与查询点 q 可信最短路径距离最小的 k 个结点。

终止:在终止时,所有与 q 的距离小于预先给定距离 D 的结点,均已访问,且与 $d_{C\text{-max}}$ 比较完毕。因此,对在距离约束 D 范围内的所有 $v\in V$,有:

$$NN_C=\{v_{i_m}\mid d_C(q,v_{i_m})\leqslant d_C(q,v),m=1,2,\cdots,k\}$$

证毕。

算法分析:算法 1 中的步骤(4)~步骤(6)是每个访问结点与查询点 p 的可信最短路径距离及其可信度的计算过程,其复杂度为 $O((2r+\Delta r)(e+n\lg n)+h\lg h)$,步骤(7)~步骤(9)是利用优先队列找到 k 个近邻,其时间复杂度最坏为 $O(\lg(n))$,即算法 1 的时间复杂度为 $O((2r+\Delta r)(e+n\lg n)+h\lg h+\lg n)$。

以可信最短路径期望距离为基础的近邻查询算法,这里称为算法 2(略),与算法 1 过程相同,只是步骤(6)中可信最短路径期望距离及其可信度的计算与算法 1 不同,其正确性证明与时间复杂度也与算法 1 相同。

在现实应用中,还有即时可信近邻的查询需求,这种需求想知道某时的可信近邻情况,而不想知道模糊网络的整体情况,针对这种需求,给出算法 3。

算法 2 即时可信近邻查询算法

输入 模糊图 $\tilde{G}=(V,E,W,Cr)$,查询点 $q\in V$,近邻数 k,距离约束 D

输出 *k-NN* 查询结果集 NN_C

(1) $NN_C=\phi$;
(2) 抽取某一时刻样本图;
(3) 从 q 出发执行 Dijkstra 算法直到达到距离约束 D;
(4) for 每个访问的结点 $v\in V$ do
(5) 计算 $d(q,v)=d,cr(d(q,v)=d)=\prod_{e\in E_d}cr(e)$;
(6) if $cr(d(q,v)=d)>cr(d(q,t)=d),t\in NN_C$ then
(7) $NN_C=NN_C\cup\{v\}$;
(8) return NN_C。

算法 3 的正确性证明与算法 1 同理。算法 3 中步骤(3)~步骤(5)是计算每个访问结点与查询点的距离及其可信度,其时间复杂度为 $O(e+n\lg n)$,步骤(6)、步骤(7)行是访问结点与当前近邻对象可信度的比较,其时间复杂度为 $O(\lg(n))$,即算法 3 的时间复杂度为 $O(e+(n+1)\lg n)$。

6　实验结果分析

实验在 2.0 GHz 双核处理器、1 GB 内存、Windows XP 平台上用 Visual C++ +6.0 实现。实验数据使用 2 个路网,一个是人工合成网 N_1,使用模拟器产生 100 000 个平面点,随机连接点形成边,对边在[1,10]区间范围内赋权重,对边在[0,1]区间范围内赋可信度,点的最大出度设为 10。另一个是实际路网 N_2,从 Digital Chart of World(DCW)获得,包含 430 274 个结点、594 104 条边。将边的可信度设为边的通过时间的倒数。使用本文所述的存储结构对两个网络进行存储,页面大小为 4K。

实验分为 2 个部分,第 1 部分测试可信距离的特征,第 2 部分测试各个参数对基于可信距离的近邻查询算法的影响。

实验 1　测试可信距离的分布情况。为估计可信距离分布,在 2 个实验数据网中分别抽取 300 个样本,每个样本中遍历 300 个结点累积距离,距离的分布结果如图 4 所示。从图 4 可知,2 个实验网络情况相似,2 个距离函数有相似分布,且与确定图中最短路径距离函数的分布相似。

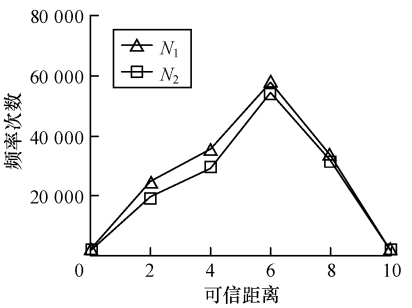

图 4　可信距离分布

实验 2　测试可信距离的收敛性与样本数的关系。在 2 个网络中分别以 300 个样本的结果为基准,计算不同样本数时近似距离的均方差。结果如图 5 所示,从图 5 可知随着样本数的增加,均方差趋于 0,可信距离依样本数收敛。

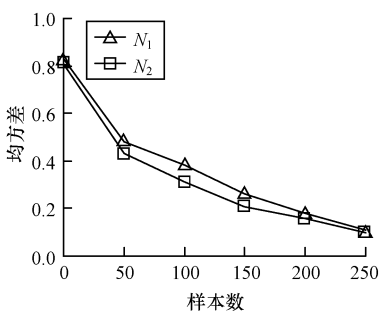

图 5　可信距离与样本数关系

实验 3　测试 k 值与基于可信距离的近邻查询算法性能的关系。因可能成为近邻查询结果的是近邻查询过程中 Dijkstra 算法访问的结点,故以访问结点数与图的总结点数的比值为主要衡量标准,来评估近邻算法的的性能。图 6 是样本数为 200 时,访问结点数与 k 值的关系结果图,从图 6 可知随着 k 值增加,访问结点数缓慢增加,近邻查询算法的性能随之下降。

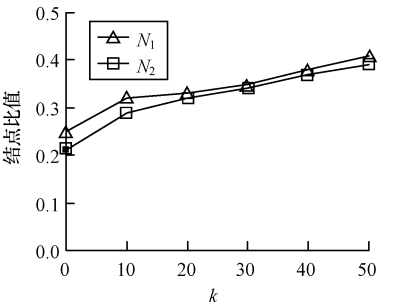

图 6　k 值与算法性能关系

实验 4　测试样本数与基于可信距离的近邻查询算法性能的关系。图 7 是 $k=20$ 时,访问结点数与样本数的关系结果,从图 7 可知随着样本数增加,访问结点数初期也急剧增加,但样本数达到一定数值后,访问结点数渐趋平稳,没有出现性能极剧恶化的情况。

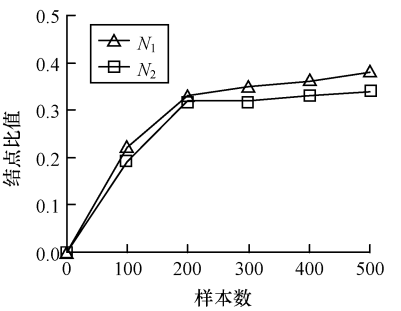

图 7　样本数与基于可信距离的近邻查询算法性能关系

距离约束 D 和精度 ε 分别影响访问结点数和样本数,从而间接影响算法性能。随着距离约束 D 的增大和精度 ε 的提高,算法性能下降。

7　结束语

针对不确定网络空间近邻查询问题,将不确定网络建模为模糊图,给出模糊图中可信最短路径距离与可信最短路径期望距离定义,以可信距离为度量,提出可信近邻查询概念,并给出距离受限条件下的可信近邻查询算法和即时可信近邻查询算法。算法使用取样方法进行近似计算,使时间复杂度为指数级的问题在多项式时间内解决,并可根据实际需

要进行精度调节。理论分析与实验结果表明算法可行且性能稳定。下一步的研究是同时考虑不确定网络的随机性与模糊性,提高不确定网络的描述能力,使不确定网络环境下的近邻查询更有效。

参考文献

[1] Cheng R, Kalashniko V D V, Prabhakar S. Querying Imprecise Data in Moving Object Environments[J]. IEEE Transactions on Knowledge and Data Engineering, 2004,16(9):1112-1127.

[2] 庄 毅. ISU-Tree:一种支持概率 k 近邻查询的不确定高维索引[J]. 计算机学报,2010,33(10):1934-1941.

[3] Yi Ke, Li Feiei, Kollios G, et al. Efficient Processing of Top-k Queries in Uncertain Databases with X-relations[J]. IEEE Transactions on Knowledge and Data Engineering, 2008,20(12):1669-1682.

[4] 高 峻,郝忠孝. 受限网络移动对象的概率最近邻查询[J]. 计算机工程,2013,39(7):26-30.

[5] 陈逸菲,秦小麟. NU2RA:一种路网中不确定移动对象范围查询分析方法[J]. 计算机研究与发展,2010,47(6):1060-106.

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

[7] 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.1.]:IEEE Press,2007:216-222.

[8] Asthana S, King O D, Gibbons F D, et al. Predicting Protein Complex Membership Using Probabilistic Network Reliability[J]. Genome Research, 2004, 14(1): 1170-1175.

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

[10] Schneider M. Fuzzy Topological Predicates, their Properties, and their Integration into Query Languages[C]//Proceedins of the 9th ACM International Symposium on Advances in Geographic Information Systems. New York, USA:[s.n.],2001:212-221.

[11] Tang X, Kainz W. Analysis of Topological Relations between Fuzzy Regions in a General Fuzzy Topological Space[C]//Proceedings of Canadian Geomatics Conference. Ottawa, Canada:[s.n.],2002:114-129.

[12] Zheng Kai, Fung Pui Cheong, Zhou Xiaofang. K-Nearest Neighbor Search for Fuzzy Objects[C]//Proceedings of 2010 ACM SIGMOD International Conference on Management of Data. Indiana, USA: IEEE Press, 2010: 333-345.

[13] Zadeh L A. Fuzzy Sets as a Basis for a Theory of Possibility[J]. Fuzzy Sets and Systems, 1978, (1): 3-28.

[14] Zadeh L A. Mathematical Frontiers of the Social and Policy Sciences[M]. Boulder, USA: Westview Press,1979.

[15] Liu Baoding, Liu Yankui. Expected Value of Fuzzy Variable and Fuzzy Expected Value Models[J]. IEEE Transactions on Fuzzy Systems,2002,10(4):445-450.

[16] Liu B. Uncertainty Theory: An Introduction to Its Axiomatic Foundations[M]. Berlin, Germany: Springer-Verlag,2004.

编辑 索书志

(上接第 53 页)

参考文献

[1] Kobayashi N, Inui K, Matsumoto J, et al. Collecting Evaluative Expressions for Opinion Extraction[C]//Proceedings of IJCNLP'05. Berlin, Germany: Springer, 2005:596-605.

[2] Li Zhuang, Feng Jing, Zhu Xiaoyan. Movie review Mining and Summarization[C]//Proceedings of the 15th ACM International Conference on Information and Knowledge Management. [S.1.]:ACM Press,2006:43-50.

[3] Hu Mingqing, Liu Bing. Mining Opinion Features in Customer reviews[C]//Proceedings of the 19th National Conference on Artifical Intelligence. San Jose, USA:AAAI Press,2004:755-760.

[4] Pang Bo, Lee L. Opinion Mining and Sentiment Analysis[J]. Foundations and Trends in Information Retrieval,2008,2(1/2):1-135.

[5] Pang Bo, Lee L, Vaithyanathan S. Thumbs Up? Sentiment Classification Using Machine Learning Techniques[C]//Proceedings of ACL'02. [S.1.]: Association for Computational Linguistics,2002:79-86.

[6] 徐琳宏,林鸿飞,杨志豪. 基于语义理解的文本倾向性识别机制[J]. 中文信息学报,2007,21(1):96-100.

[7] 朱嫣岚,闵 锦,周雅倩,等. 基于 hownet 的词汇语义倾向计算[J]. 中文信息学报,2006,(1):14-20.

[8] Nasukawa T, Yi J. Sentiment Analysis: Capturing Favorability Using Natural language processing[C]//Proceedings of the 2nd International Conference on Knowledge Capture. Sanibel Island,USA:ACM Press,2003:70-77.

[9] 刘宏哲,须 德. 基于本体的语义相似度和相关度计算研究综述[J]. 计算机科学,2012,39(2):8-13.

[10] Turney P D, Littman M L. Measuring Praise and Criticism: Inference of Semantic Orientation from Association[J]. ACM Transactions on Information Systems,2003,21(4):315-346.

[11] 朱 杰,刘功申,陈 卓. 中文文本倾向性分类技术比较研究[J]. 信息安全与通信保密,2010,(4):56-58.

[12] Makhoul J, Kubala F, Schwartz R, et al. Performance Measures for Information Extraction[C]//Proceedings of DARPA'99. [S.1.]:IEEE Press,1999:249-252.

编辑 索书志