无线传感器网络不确定数据 PT-Top k 查询处理技术

毛莺池[1,2] 王 康[1] 任道宁[1] 王久龙[1]

(河海大学计算机与信息学院 南京 211100)[1] (河海大学淮安研究院 淮安 223001)[2]

摘 要 在无线传感器网络现实应用中,感知数据普遍存在不确定性。由于不确定数据引入了概率维度,使得不确定数据查询种类更加丰富,同时也给查询处理带来困难。不确定数据 Top-k 查询是一个典型的不确定数据查询任务。考虑到无线传感器网络查询处理技术对查询响应时间和网络通信消耗的高要求,研究了面向层次聚簇结构的无线传感器网络不确定数据 Top-k 查询处理技术。通过分析不确定数据特点,基于 x-tuple 规则元组模型,采用簇内与簇间的两阶段数据查询处理机制,提出了基于 Poisson 分布的分布式不确定数据 PT-Top k 查询处理近似算法 TPQP。通过实验,从总体通信消耗、与概率阈值 p 相关分析、与排序数 k 相关分析以及数据敏感度分析等方面,说明了 TPQP 算法在通信消耗、查询响应时间上的优越性。

关键词 无线传感器网络,Top-k,层次聚簇,x-tuple 规则,分布式 PT-Top k 查询

中图法分类号 TP393 **文献标识码** A **DOI** 10.11896/j.issn.1002-137X.2014.12.016

Uncertain Data PT-Top k Query Processing in Wireless Sensor Network

MAO Ying-chi[1,2] WANG Kang[1] REN Dao-ning[1] WANG Jiu-long[1]

(College of Computer and Information, Hohai University, Nanjing 211100, China)[1]

(Huaian Research Institute of Hohai University, Huaian 223001, China)[2]

Abstract For the widespread wireless sensor networks applications, due to the quality of sensors and environment factor, the sensor readings are inherently uncertain. With the introduction of the probability dimension in the uncertain data, the query processing technologies for uncertain data become more and more difficult, and the types of uncertain data query have become richer. Uncertain data Top-k query is one of typical query tasks for the uncertain data. Considering the energy consumption and query response time in the wireless sensor network, an uncertain data PT-Top k query processing scheme is studied in a hierarchical structural wireless sensor network. Based on the x-tuple Rule of uncertain data, using intra-cluster and inter-cluster two phases query processing, a distributed Two-Phase PT-Top k Query Processing approximation algorithm (TPQP) was proposed. Finally, the extensive experiment results show that the proposed TPQP can reduce the transmission consumption and query response time in terms of the probability p, the sorted number k, and the data volume.

Keywords Wireless sensor networks, Top-k, Hierarchical cluster structure, x-tuple rule, Distributed PT-Top k query

1 引言

无线传感器网络(Wireless Sensor Networks,WSN)能够实时对各种环境进行监测,协作地感知、采集和处理网络覆盖区域中被感知对象的状态信息,并传输至用户。通常情况,WSN 一般部署在一个范围广阔的区域,但用户有时并不关心整个感知区域的所有数据,可能仅关心感知区域内某一时刻或者某一范围内的环境变化情况,也可能是某一特定目标的活动情况。因此,采用 Top-k 查询获得所需信息。Top-k 查询是无线传感器网络中一种比较典型的查询,要求返回指定区域内感知数据集中 k 个特定数据。

在传统数据库应用中,数据的存在性和精确性均为确定的。然而,在 WSN 的许多现实应用中,感知数据的不确定性普遍存在。如:节点监测精度不高,感知数据自身不精确;由于电池能量的消耗,传感器失效或废弃,产生数据缺失或不正确;无线传感器网络应用环境复杂,受自然环境影响导致数据不精确;网络传输过程中,受到外界信号干扰,导致数据的不确定性[1]。数据的不确定性给传感器网络应用带来很大影响,使得数据不可信,用户不能直接从中获取有用信息。因此,对不确定性数据的 Top-k 查询处理也显得尤为重要。然而,无线传感器网络中感知数据量大,Top-k 查询处理需要传输大量感知数据,消耗了网络能耗。为了减少网络能量消耗,

到稿日期:2013-12-11 返修日期:2014-03-10 本文受国家自然科学基金(61272543),国家科技支撑计划项目(2013BAB06B04),江苏省自然科学基金(BK2012584),中央高校基本业务费资助(2013B06914),河海大学淮安研究院开放基金资助。

毛莺池(1976—),女,副教授,主要研究方向为无线传感器网络、分布计算与并行处理,E-mail:yingchimao@hhu.edu.cn;**王 康**(1989—),男,硕士生,主要研究方向为无线传感器网络、分布计算与并行处理;**任道宁**(1990—),男,硕士生,主要研究方向为分布计算与并行处理;**王久龙**(1991—),男,硕士生,主要研究方向为分布计算与并行处理、数据管理。

延长网络生命周期，提高网络能量的利用效率，研究一个能量有效的不确定数据 Top-k 查询处理技术非常必要。

在不确定数据集上的 Top-k 查询是指一个元组在某个或某些个由不确定数据集中某些元组构成的可能世界中排序在前 k 位，那么这个元组在不确定数据上的 Top-k 概率即为其在所有可能世界中成为 Top-k 的概率之和。目前，不确定数据集上 Top-k 查询语义主要分为 4 种：Uncertain Top-k 查询（U-Top k）[2]、Uncertain k ranks 查询（U-kRanks）[2]、Probabilistic Threshold Top-k 查询（PT-Top k）[3]、Probabilistic k top-k 查询（Pk-Top k）[4]。U-Top k 查询和 U-kRanks 查询对查询结果的排序顺序有着严格要求，Pk-Top k 查询对元组的 Top-k 概率顺序有着一定的要求。而 PT-Top k 查询对结果顺序没有特定要求，而且可以得到更多的信息，更重要的是 PT-Top k 查询对元组的可信度有一定的质量要求，对用户而言，只有满足一定质量要求的数据才是可信的。因此，本文研究无线传感器网络不确定数据 PT-Top k 查询处理技术。

处理不确定数据 PT-Top k 查询最直接的方法就是 Naïve 算法，即打开所有可能世界，按照排序和概率关系求出查询结果。由于可能世界数量级非常大，Naïve 算法是一个低效率算法。文献[5]给出了一个基于 Poisson 分布的 PT-Top k 查询近似算法，该算法可避免打开所有可能世界，高效求得不确定元组 Top-k 概率，但是，此算法仅限于集中式数据库。而无线传感器网络能量受限，将数据全部收集集中处理的方法必消耗大量能量，缩短网络寿命，因此，不能直接应用于无线传感器网络。要解决这个问题，就必须研究提出能耗高效的分布式不确定数据 PT-Top k 查询处理算法。

但是，现有的分布式不确定数据 Top-k 查询处理算法还存在不足。1）现有不确定 Top-k 查询处理算法关注簇间数据处理，在簇头节点收集数据阶段，没有给出一个合理的 x-tuple 关系的不确定数据修剪策略，使得将不可能成为查询结果的数据也传输给用户，产生不必要的能耗。2）在簇间数据处理阶段，不确定 Top-k 查询处理算法在执行过程中，仍然存在大量冗余数据被传输到 Sink 节点，也导致网络能量不必要的消耗。3）没有考虑到算法对查询响应延迟的影响。针对以上不足，通过分析不确定数据特点，基于 x-tuple 规则元组模型，采用簇内与簇间两阶段数据查询处理机制，提出基于 Poisson 分布的分布式不确定数据 PT-Top k 查询处理近似算法（Two Phase PT-Top k Query Processing，TPQP）。实验从总体通信开销、与概率阈值 p 相关分析、与排序数 k 相关分析以及数据敏感度分析等方面，来验证 TPQP 算法在通信消耗、查询响应时间上的有效性。

2 相关工作

近年来，开展了许多无线传感器网络 Top-k 查询处理的研究工作[6-13]。TAG[6]基于生成树，在数据路由时，父节点实现对子节点的数据聚合（*Min*，*Max*，*Sum*，*Ave*，*Top-k* 等）操作，减少了 WSN 中的数据传输量，节省了通信消耗。TA 算法[10]基于阈值从查询数据列表中找出满足要求的结果，并将结果合并成一个列表。目前，已有的 TA 改进算法主要依赖于是否允许随机访问列表。文献[13]提出了处理近似结果集的 TA 改进算法，此近似结果集有一定的概率质量保证。文献[11]在数据列表上使用统计方法优化 TA 算法执行。文献[12]提出了排序数据列表，还允许 TA 算法使用先前的查询结果回答查询。

文献[9]将分布式阈值连接算法应用到了 Top-k 查询中，称之为 TJA 算法。TJA 算法是一种在层次拓扑结构的无线传感器网络中执行有效的 Top-k 算法。TJA 算法利用用户给出的相似排序函数找到 k 个最大的结果，通过节点上的额外探测和数据过滤达到减少数据传输量的目的。其至少需要经过两个阶段才能完成最终的 Top-k 查询，而且算法性能在很大程度上受节点中对象排序的影响。当节点中对象的等级排列非常相似时，TJA 算法可以利用较少的数据量传输来完成 Top-k 查询，但是当节点中对象的等级排列差别很大时，将必然导致在第二个阶段后有大量对象是不完整的，因此必然导致第三个阶段中有大量的数据在网络中进行传递，从而算法的性能也大大地降低。

文献[7,14]提出了近似数据聚集技术，其思想是为了平衡数据质量和能量效率。Sliberstein[8]等人提出使用一种基于采样的方法来估算 WSN 中近似 Top-k 查询。文献[15]提出一个模型驱动的方法，其是基于统计的模型技术，用来平衡查询回答的可信度而不是网络中的通信消耗。此外，文献[16,17]研究了无线传感器网络中连续 Top-k 查询。

以上这些研究都是基于确定数据的，在数据库的许多现实应用中，数据往往是不确定的。而上述研究并没有考虑到数据的不确定性，因此其并不适合应用到不确定性数据的查询处理中。本文工作与上述研究最大的不同在于充分考虑了数据的不确定性，研究不确定性数据上的 Top-k 查询处理技术。

Mao Ye[3,18]等人在无线传感器网络不确定数据查询中，引入精确 PT-Top k 查询算法[3]，提出基于聚簇网络拓扑的分布式不确定数据 PT-Top k 查询处理算法 SSB。其根据局部 Top-k 结果和全局 Top-k 结果之间的关系，对簇内数据进行修剪，仅需要一次数据请求就可完成簇内数据查询。在簇内局部数据基础上，簇头节点过滤掉不需要的不确定数据元组，基站执行全局 PT-Top k 查询。在此基础上，进一步提出了处理簇内不确定数据表 T 的方法，簇内节点采用两次数据请求，提出 NSB 和 BB 算法，实现簇内的不确定数据查询[19]。

文献[20]给出了 P2P 网络中分布式不确定数据 Top-k 查询处理方法。对于每个不确定元组中的属性值确定其错误半径 ER。根据 ER 值，不确定元组建模为一个不确定矩形区域。通过构建四叉树索引 UQ-Tree 和全局索引，提高数据查询响应时间。使用空间修剪技术和分布式修剪技术，减少网内数据传输量，节约网络能耗。但是局部数据需要通过中间节点才可以相互通信，交换数据由于节点能量有限，多次通信会造成能量损耗。

因此，本文基于 x-tuple 规则元组模型，采用簇内与簇间两阶段数据查询处理机制，提出基于 Poisson 分布的分布式不确定数据 PT-Top k 查询处理近似算法 TPQP。

3 系统模型与定义

3.1 系统模型

无线传感器网络中通常会有 1 个或多个 Sink 节点，N 个感知节点部署在监测区域，Sink 节点和感知节点可通过多跳

方式通信。为了简化网络模型，对无线传感器网络做出如下假设：

(1)无线传感器网络部署区域 S 为二维平面区域；

(2)网络节点为静态的，所有节点一旦部署不再移动；

(3)网内只有一个 Sink 节点，其存储能力和计算能力不受限制，网内有 N 个具有有限计算能力和通信能力的感知节点；

(4)感知节点仅具有有限的通信半径，Sink 节点可直接与感知节点通信，而感知节点通过多跳方式与 Sink 节点通信；

(5)无线传感器网络是连通的，节点发送数据报文时，节点一跳范围内的邻居节点都可以监听到。

由于无线传感器网络拓扑结构和路由协议不同，因此网内查询的效率和网络能量损耗也不同。本文将采用层次网络拓扑结构，如图 1 所示，使用 TEEN[20] 路由协议执行查询分发和数据传输。

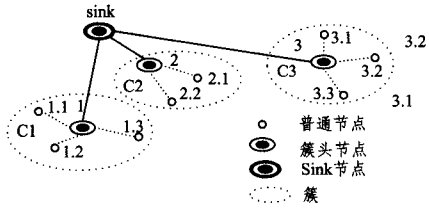

图 1　网络模型

3.2　相关定义

定义 1(不确定元组)　数据表 T 中有 n 条数据元组，若元组 $t_i(1\leqslant i\leqslant n)$ 的值域为 $D=[M]\{\perp\}$，其中 $[M]$ 是一个正实数域，取值概率为 p_i，$\perp$ 表示为空，即不存在，不存在概率为 $1-p_i$，则称此元组为不确定元组。

定义 2(x-tuple 规则元组)　不确定数据表 T 中有 n 个不确定元组，W 表示 T 中所有不确定元组可构成的可能世界集合，w 是一个可能世界实例。对于 $\forall w\subseteq W$，$\forall t_i\in T$，$\forall t_j\in T$，$(1\leqslant i,j\leqslant n)$，如果存在 $t_i\in w, t_j\notin w$，则称元组 t_i 和 t_j 具有相同 x-tuple 关系，此元组 t_i 和 t_j 称为 x-tuple 规则元组，用 τ 表示。

τ 的存在概率为 $P(\tau\neq\perp)=\sum_{t_i\in\tau}p_i$，不存在的概率为 $P(\tau=\perp)=1-\sum_{t_i\in\tau}p_i$。

从定义 2 可以看出，x-tuple 规则元组是由一些不确定元组构成，而且这些不确定元组存在互斥关系，不会出现在同一可能世界实例中。

本文所描述的 x-tuple 规则元组皆来自相同数据源节点，设传感器节点每次感知数据时，可确定若干数据项，每个数据项都存在确定概率，所有数据项概率和小于等于 1。每个数据项及其概率对应一个元组。因此，相同节点同时产生的多个元组即为 x-tuple 规则元组。

例 1　设某个传感器节点上有一个不确定数据表 T，如表 1 所示，其有 4 个属性，分别是 Sensor_id、Time、Temperature、Probability。Temperature 为感知数据，Probability 为概率。为了方便描述，在 T 中加入 Tuple_id 和 x-Tuple 属性。通过观察可知，数据项 20.1 和 19.8 是传感器节点 s_1 在时间 $Time_1$ 的感知数据，两个数据项所对应的元组具有相同 x-tuple 规则，隶属于 τ_1。同理，数据项 18.5 和 17.6 所对应的元组也具有相同 x-tuple 规则，隶属于 τ_2。

表 1　不确定数据表 T

x-Tuple	Sensor_id	Time	Temperature	Probability
τ_1	s_1	$Time_1$	20.1	0.2
τ_1	s_1	$Time_1$	19.8	0.7
τ_2	s_1	$Time_2$	18.5	0.9
τ_2	s_1	$Time_2$	17.6	0.1

不确定数据表 T 的所有可能世界集合用 W 表示，$w\in W$ 表示一个可能世界实例。根据可能世界实例生成规则[8]，不确定数据表 T 中每个可能世界实例的存在概率为：

$$P(w)=\prod_{\tau\cap w=t_i}p_i\prod_{\tau\cap w=\varnothing}(1-P(\tau)) \tag{1}$$

例 2　以例 1 中表 T 为例，根据可能世界实例存在概率计算式(1)，计算 T 的所有可能世界实例及其存在概率，如表 2 所列。

表 2　不确定数据表 T_3 的可能世界集合

Possible world	Probability
$w_1=\{20.1,18.5\}$	0.18
$w_2=\{20.1,17.6\}$	0.02
$w_3=\{19.8,18.5\}$	0.63
$w_4=\{19.8,17.6\}$	0.07
$w_5=\{18.5\}$	0.09
$w_6=\{17.6\}$	0.01

定义 3(排列顺序，ranking order)　如果不确定数据表 T 由若干元组组成，即 $T=\{t_1,t_2,\cdots,t_n\}$，T 中元组 t_i 在排序函数 f 上的值比 t_j 大，即 $f(t_i)>f(t_j)$，则有 $t_i\prec_f t_j$。

本文以下所使用的排序按照感知数据值大小降序排序。当数据值大小相同时，概率大者排名更高。

定义 4(支配集，Dominant Set，DS[22])　设 T 是不确定数据表，w 是 T 上的一个可能世界，给定元组 $t\in T\wedge t\in w$，t 能否成为可能世界 w 上的 Top-k，依赖于 w 中排序在 t 之前元组数量是否小于 k。所以，元组 t 的支配集可以表示为：

$$DS_t=\{t'\mid t'\in T\wedge t'\prec_f t\} \tag{2}$$

定义 5(修剪上界，Pruning Upper Bound，PUB)　存在一个有序不确定数据表 T，T 中有 n 个元组，$t_i\in T(1\leqslant i\leqslant n)$，$u_i$ 为 t_i 支配集的概率之和，给定正整数 k 和概率阈值 p，当 u_i、k 和 p 满足式(3)时有：

$$\mu_{DS}\geqslant k+\ln\frac{1}{p}+\sqrt{\ln^2\frac{1}{p}+2k\frac{1}{p}} \tag{3}$$

其中，t_i 为不确定数据集 T 上的修剪上界，简称 PUB。

有序不确定数据表 T 中，排序在 PUB 之后的数据集合是不可能成为 PT-Top k 查询结果的，在进行查询处理时，这些数据无需传输至 Sink 节点，减少了通信开销。

定义 6(完备集，Complete Set，CS)　给定不确定数据表 T，$\exists t_i\in T$，$1\leqslant i\leqslant n$，且 $PUB=t_i$，则不确定数据表 T 上的完备集 CS 表示为：

$$CS(T)=\{t\mid t\prec_f t_i\cup t=_f t_i\} \tag{4}$$

由完备集定义和修剪上界定义可知，不确定数据集上的 PT-Top k 查询结果只能出现在完备集 CS 中，不存在例外情况。

定义 7(充足集，Sufficient Set，SS)　给定不确定数据表 T，A 是 T 上的不确定 PT-Top k 查询结果集，A 中有 n 个元组，$\exists t_i\in A$，$\forall t_j\in A$，$1\leqslant i,j\leqslant n$，且 $t_j\neq t_i$。若 $P_{wp-k}(A)>k-p$ 成立，且存在 $t_j\prec_f t_i$，则称元组 t_i 为不确定数据表 T 上的充足集下界(Sufficient Set Lower Bound，SLB)，用 T_{SLB} 表

示。这样，不确定数据表 T 上的充足集 SS 可以表示为：

$$SS(T)=\{t \mid t=_f T_{SLB} \cup t \prec_f T_{SLB}\} \quad (5)$$

定义 8(必须集，Necessary Set，NS) 给定不确定数据表 T，A 是 T 上的不确定 PT-Top k 查询结果集，A 中有 n 个元组，$\exists t_i \in A, \forall t_j \in A, 1 \leqslant i, j \leqslant n$，且 $t_j \neq t_i$。若 $P_{top-k}(A) > k-p$ 不成立，且存在 $t_j \prec_f t_i$，则称元组 t_i 为不确定数据表 T 上的必须集下界(Necessary Set Lower Bound，NLB)，用 T_{NLB} 表示。这样，不确定数据表 T 上的必须集 NS 可以表示为：

$$NS(T)=\{t \mid t=_f T_{NLB} \cup t \prec_f T_{NLB}\} \quad (6)$$

4 两阶段分布式 PT-Top k 查询算法 TPQP

4.1 算法基本思想

为了降低查询过程产生的通信开销，减少查询响应延时，本文提出的两阶段分布式 PT-Top k 查询算法 TPQP 拟采用簇内与簇间两阶段数据查询处理机制实现，其基本思想如下：无线传感器网络采用层次聚簇网络拓扑结构，传感器节点分为簇头节点和簇内节点。每个传感器节点收集感知数据，并将数据存储在本地的由 l 个 x-tuple 规则元组构成的不确定数据表中。当 Sink 节点发出 PT-Top k 查询时，由簇头节点接收查询，将查询请求转发至簇内成员节点，并执行簇内查询。由簇头节点通过多跳方式，将簇内查询的结果返回至 Sink 节点。Sink 节点执行基于 Poisson 分布的 PT-Top k 查询处理近似算法，产生最终查询结果。

在簇内数据查询处理阶段，簇内节点接收到查询请求，根据概率阈值 p 和排序数 k 在其不确定数据表上执行 PT-Top k 查询，当满足查询算法终止执行条件时，将满足查询要求的不确定元组发送至簇头节点。簇头节点将所有收集到的不确定元组进行排序，找出排序最高的不确定元组作为局部阈值，并转发给其簇内成员节点。簇内节点将本地不确定数据表上大于此局部阈值的所有不确定元组发送给簇头节点，有效地过滤掉部分冗余数据。

对于簇间数据查询处理阶段而言，目的是过滤不会成为全局不确定数据表 PT-Top k 查询结果的部分元组。簇头节点将收集到的所有不确定元组降序排序，并执行 PT-Top k 查询处理近似算法，计算出查询结果。此查询结果是各个簇头节点局部不确定数据表上的查询结果，而局部不确定数据表上的查询结果并不一定能够成为全局不确定数据查询结果，存在数据冗余。考虑到在查询结果之外可能存在一个高概率元组(此元组排序在查询结果之后，且概率为 1)，能够对查询结果产生影响，为此，将查询结果分为两类：受到影响查询结果和不受影响查询结果。将查询结果排序最低元组的感知数据项作为查询结果下界。相应地，查询结果下界分为受到影响查询结果下界和不受影响查询结果下界。在 Sink 节点，分别确定所有受到影响查询结果下界的最小值和不受影响查询结果下界的最大值。再比较此最小值和最大值，将其中较大值记为全局下界，并发送至簇头节点。簇头节点将感知数据项大于全局阈值的不确定元组发送至 Sink 节点。最终，Sink 节点将收集的不确定元组按降序排序，执行 PT-Top k 查询近似算法，得到最终的查询结果。

4.2 簇内查询处理算法 CSB

设在一个簇 C_i 内有 n 个簇内节点 S_i。节点中存储采集到的不确定数据，在接收到查询请求后，将数据发送至簇头节点，因此，簇可以被看作一个局部分布式数据库。

每个簇内成员节点 S_i 维持一张降序排列的不确定数据表 T_{Si}，通过计算 T_{Si} 上的修剪上界 PUB 找出其完备集 $CS(S_i)$。簇头节点只要接收到能够有效计算簇内不确定数据 PT-Top k 查询结果的数据即可。但是，若簇内节点将其局部数据表上的完备集 $CS(S_i)$ 直接传输至簇头节点，必存在不可能成为最终查询结果的数据，簇内节点传输冗余数据，导致不必要的通信开销。考虑采用簇内两次数据请求策略，实现簇内数据修剪。基于完备集 CS 和修剪上界 PUB 定义，提出基于完备集的簇内查询处理算法(Complete Set-Based，CSB)。

设 $L_{PUB}(S_i)$ 和 C_{PUB} 分别表示簇内节点和簇头节点的修剪上界 PUB。第一次数据请求时，簇内节点仅将局部不确定表 T_{Si} 上的修剪上界 $L_{PUB}(S_i)$ 发送给簇头节点，簇头节点从接收到的簇内节点 $L_{PUB}(S_i)$ 中找出合适的上界作为 C_{PUB}，并广播给其簇内成员。第二次数据请求时，簇内节点将簇内全局上界 C_{PUB} 作为阈值，将其完备集 $CS(S_i)$ 的全部数据发送至簇头节点。如何确定合适 C_{PUB} 是簇内数据处理算法 CSB 的关键。根据完备集定义，簇头节点选择最大的 $CS(S_i)$ 作为簇内修剪上界 C_{PUB}。算法步骤如下，伪代码如图 2 所示。

(1)簇内节点计算其局部不确定数据集上的修剪上界 $L_{PUB}(S_i)$，并发送至簇头节点。

(2)簇头节点接收所有簇内节点的 $L_{PUB}(S_i)$，选择其中最大值 $MAX\{L_{PUB}(S_i), i=1,2,\cdots,n\}$ 作为簇内全局修剪上界 C_{PUB}，并广播至簇内所有节点，请求排序在 C_{PUB} 之前的元组。

(3)簇内节点接收 C_{PUB} 后，将排序在 C_{PUB} 之前的数据元组集合发送至簇头节点。

(4)簇头节点接收所有簇内节点发送的数据，并存储在其不确定数据表 T_C 中。

```
(1)For each Si in Ci;
(2)     LPUB(Si);
(3)     Send LPUB(Si) to Cluster node;
(4)CPUB<--MAX{LPUB(Si),i=1,2,…,n} ;
(5)Broadcast CPUB to each Si;
(6)Waiting data C'pub from Si;
  //C'pub表示排序在CPUB之前的数据元组集合
(7)Send C'pub to Cluster node;
(8) Save data in TC;
```

图 2 簇内 CSB 算法伪代码

图 3 给出了簇内查询处理算法 CSB 的示意图。给定 3 个簇内节点 S_1, S_2, S_3，其降序排列的不确定数据表分别是 T_{S1}, T_{S2}, T_{S3}，完备集分别是 $CS(S_1), CS(S_2), CS(S_3)$，修剪上界 PUB 分别为 $L_{PUB}(S_1), L_{PUB}(S_2), L_{PUB}(S_3)$。可知，$C_{PUB}=L_{PUB}(S_3)=MAX\{L_{PUB}(S_1), L_{PUB}(S_2), L_{PUB}(S_3)\}$，只有簇内节点 S_3 发送完整的完备集 $CS(S_3)$，另外两个节点 S_1, S_2 发送部分数据。

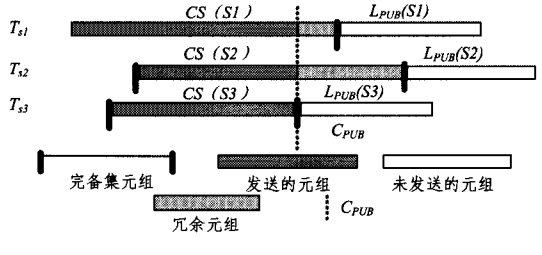

图 3 簇内 CSB 算法数据处理

根据簇内查询处理算法 *CSB* 的执行过程可知，通过簇头和簇内节点两次通信，簇内节点进行数据修剪，相对于直接发送簇内节点所有数据集上的完备集 *CS*，大大减少了数据传输量，降低了通信能耗。

4.3 簇间查询处理算法 SNSB

采用簇内查询处理算法 CSB，修剪了簇内不确定数据，减少了数据传输量。但是，CSB 仅仅是根据元组排序和存在概率对数据修剪，并没有考虑到不确定数据元组 Top-k 概率与概率阈值的关系。由于不确定元组的存在概率大于等于不确定元组 Top-k 概率，对于排序比较低的不确定元组，即使其存在概率很大，最后得到的 Top-k 概率也可能非常低，甚至不满足概率阈值 p 的要求。若将所有簇头节点存储的不确定数据都发送给 Sink 节点，则仍然存在数据冗余，因此，基于定义的充足集 *SS* 和必须集 *NS*，提出一个簇间分布式不确定数据 PT-Top k 查询处理近似算法（Sufficient or Necessary Set-Based，SNSB）。

簇间数据处理时，簇头节点 C_i 首先将其局部数据表 T_G 上的充足集下界 $SLB(C_i)$ 和必须集下界 $NLB(C_i)$ 传输给 Sink 节点。由 Sink 节点根据所有簇头节点的 $SLB(C_i)$ 和 $NLB(C_i)$ 排序高低，找出一个全局下界 G_{LB}。G_{LB} 作为阈值发送给所有簇头节点，簇头节点将排序在阈值 G_{LB} 之前的数据发送给 Sink 节点。如何确定全局下界 G_{LB} 是簇间查询处理算法 *SNSB* 的关键。根据 *SS* 和 *NS* 定义，即当 *NS* 之外加入一个排序低于 *NLB* 的高概率元组时，其可以转化为 *SS*，而 *SS* 不受影响。因此，Sink 节点选择 $G_{LB}=\mathrm{MAX}\{\mathrm{MAX}\{SLB(C_i)\},\mathrm{MIN}\{NLB(C_i)\}\}$，$i=1,2,\cdots,m$ 作为全局下界 G_{LB}。算法步骤如下，伪代码如图 4 所示。

(1)簇头节点 C_i 在其局部不确定数据集 T_G 上，通过执行集中式近似 PT-Top k 查询算法，计算不确定数据集的必须集下界 $NLB(C_i)$ 和充足集下界 $SLB(C_i)$，并将其发送至 Sink 节点。

(2)Sink 节点接收所有簇头节点的 $NLB(C_i)$ 和 $SLB(C_i)$，选择 $G_{LB}=\mathrm{MAX}\{\mathrm{MAX}\{SLB(C_i)\},\mathrm{MIN}\{NLB(C_i)\}\}$ 作为全局下界，并将 G_{LB} 发送至所有簇头节点。

(3)簇头节点根据接收的 G_{LB}，将其不确定数据集 T_G 上排序高于 G_{LB} 的不确定数据元组集传输给 Sink 节点。

(4)Sink 节点接收簇头节点发送的数据，存储在其不确定数据表 T_{Sink} 中，接着，Sink 节点将 T_{Sink} 中所有元组按排序函数降序排列，执行基于 Poisson 分布的不确定数据 PT-Top k 近似查询处理算法[3]。

```
(1)For each T_Ci in C_i;
(2)     集中式-PT-Top k(T_Ci);
(3)     Send {NLB(C_i),SLB(C_i)} to Sink;
(4)  G_LB=MAX{MAX{SLB(C_i)},MIN{NLB(C_i)}};
(5)Send G_LB to each C_i;
(6)Waiting data G'_LB from C_i;
//G'_LB表示 T_Ci 上排序高于 G_LB的不确定数据元组集合
(7)Send G'_LB to Sink;
(8)Save data in T_sink;
(9)T_Sink in descending order;
(10)Poisson-PT-Top k(T_sink);
```

图 4　簇间数据处理算法 SNSB 伪代码

图 5 说明了簇间数据处理算法 SNSB 的过程，簇头节点分别为 C_1，C_2，C_3，C_4，其局部数据集的充足集下界中 *SLB* 最大者为 $SLB(C_2)$，*NLB* 最小者为 $NLB(C_4)$，而 $SLB(C_2)>NLB(C_4)$，所以，全局不确定数据表查询结果下界 $G_{LB}=SLB(C_2)$。

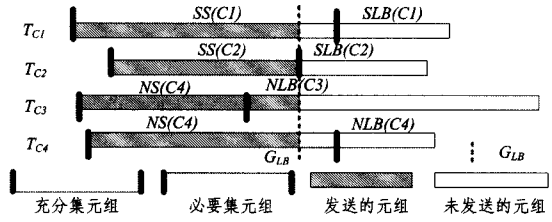

图 5　簇间数据处理算法 SNSB

从 SNSB 算法执行过程可知，Sink 节点需要两次数据请求，收集到期望的不确定数据元组。虽然，两次数据请求增加了网络通信消耗，但是，减少了查询数据的冗余度，通过实验验证，总体上可以有效地减少网络通信消耗。

4.4 算法分析

4.4.1 正确性分析

通过分析元组在不确定数据集上 Top-k 概率的计算过程可知，不确定元组的 Top-k 概率受到其在所有可能世界排序的影响，也就是说，受到其支配集影响。根据支配集 *DS* 定义及 Top-k 概率计算过程，分析两阶段分布式 PT-Top k 查询算法 TPQP 的正确性。

给定一个不确定元组 $t\in T(i)\wedge t\in T(j)$，$T(i)$ 和 $T(j)$ 是两个不确定数据表，$T(i)\neq T(j)$，$DS_{T(i)}(t)$ 是元组 t 在 $T(i)$ 上的支配集，$DS_{T(j)}(t)$ 是元组 t 在 $T(j)$ 上的支配集，如果存在 $DS_{T(i)}(t)\subseteq DS_{T(j)}(t)$，则有 $P_{T(i),topk}(t)\leqslant P_{T(j),topk}(t)$。

根据基于 Poisson 分布的 PT-Top k 近似查询分析结果可知[13]，

$$P_{topk}(t)=P(t)\frac{\Gamma(k,\mu)}{(k-1)!}$$

其中，μ 是元组 t 支配集的概率之和。μ 是一变量，$P_{topk}(t)$ 随着变量 μ 的增大而减小。又因为 $DS_{T(i)}(t)\subseteq DS_{T(j)}(t)$，$\mu_{T(i)}\leqslant\mu_{T(j)}$，有 $P_{T(i),topk}(t)\leqslant P_{T(j),topk}(t)$。

对于无线传感器网络而言，在簇头节点的不确定数据表中，元组 t 的支配集包含于全局不确定数据表上元组 t 的支配集。只要将元组 t 在各个簇头上的支配集传输给 Sink 节点，那么其 Top-k 概率只会更加精确。TPQP 算法在收集数据时，都会考虑到将不确定元组的支配集 *DS* 完全收集，因此，无线传感器网络分布式不确定数据 PT-Top k 查询处理技术 TPOP 算法是正确的。

4.4.2 通信开销分析

根据层次型网络拓扑结构的特点，估算分布式 PT-Top k 查询处理算法的数据包大小。设监测区域部署了由 $N+M$ 个传感器节点构成的无线传感器网络，其中有 M 个簇头节点，簇内普通感知节点数量为 N。S_q 表示发送一次查询请求的通信开销，S_b 表示发送一条上下界消息的通信开销，S_t 表示一条数据元组的节点通信消耗。$|CS(S_i)|$ 表示簇内节点 S_i 发送的不确定数据元组数量，$|SNS(C_j)|$ 表示簇头节点 C_j 上传不确定数据元组的数量。H_j 表示簇头节点 C_j 到 Sink 节点的跳数。

根据 CSB 算法的执行过程可知，其需要簇头和簇内节点两次通信才能实现簇内数据修剪。一次为发送查询请求，一

次为发送 C_{PUB}。此外，簇内节点向簇头节点发送不确定数据元组。因此，簇内查询处理算法 CSB 的通信开销为：

$$C_{CSB}=C_{query}+C_{C\_bound}+C_{tuple}$$
$$=M\times S_q+(M+N)\times S_b+\sum_{i=1}^{N}|CS(S_i)|*S_t$$

执行簇间查询处理算法 SNSB 时，仅有簇头节点发送下界信息与不确定数据元组，因此，簇间查询处理算法 SNSB 的通信开销为：

$$C_{SNSB}=C_{sink\_bound}+C_{tuple}$$
$$=\sum_{j=1}^{M}S_b\times H_j+\sum_{j=1}^{M}|SNS(C_j)|\times S_t\times H_j$$
$$=\sum_{j=1}^{M}(S_b+|SNS(C_j)|\times S_t)\times H_j$$

分布式不确定数据 Top-k 查询算法 TPQP 算法是由簇内查询处理算法 CSB 和簇间查询处理算法 SNSB 构成。因此，TPQP 算法产生的通信开销为：

$$Cost=C_{CSB}+C_{SNSB}$$
$$=M\times S_q+(M+N)\times S_b+\sum_{i=1}^{N}|CS(S_i)|*S_t+$$
$$\sum_{j=1}^{M}(S_b+|SNS(C_j)|\times S_t)\times H_j$$

5 实验验证

5.1 实验环境

实验采用仿真软件 MatLab 7.8.0(R2009)进行仿真。监测区域部署了由一个 Sink 节点和 625 个传感器节点构成的无线传感器网络，并采用层次聚簇网络拓扑结构。其中，节点被分为 30 个簇，即有 30 个簇首节点与 595 个普通节点。簇首节点根据其通信半径大小采用单跳与多跳的方式和 Sink 节点通信。实验不考虑噪声干扰对节点通信的影响，假定节点通信状况良好。

无线传感器网络不确定数据 TP-Top k 查询处理算法的优劣影响网络通信开销与查询响应时间。实验将对 TPQP 算法、BB 算法以及 SSB 算法在网络通信开销与查询响应时间两方面的性能指标进行比较与分析。同时，为了说明提出的 TPQP 算法对数据精确度是在可控范围内，采用文献[6]给定的基于 Poisson 分布的集中式 PT-Top k 查询近似算法数据精确度模型，来分析 TPQP 算法的平均绝对误差率。实验参数具体如表 3 所列。

表 3 实验参数设置

参数名	默认值	可选值
元组排序 k	5	(3,10)
概率阈值 p	0.5	(0.3,0.7)
簇头节点数量	30	
单位元组长度(byte)	32	(20,100)
查询消息长度(byte)	8	
下界消息长度(byte)	8	
节点数据传输率(byte/s)	10^5	
执行计算时间(s)	0.0001	
近似算法 Top-k 时间(s)	0.003	
节点延迟(s)	0.01	

5.2 实验数据

实验所用数据集是从 2004 年 2 月 28 日到 2004 年 4 月 5 日，由部署在英特尔伯克利研究实验室(Intel Berkeley Research Lab)的 54 个传感器节点所收集到的数据，称为 Intel Lab data[15]。此数据集通过 TinyOS 网内查询处理系统每 31 秒收集一次感知数据，共计 143M。数据集有 4 个感知属性，分别是：temperature、humidity、light、voltage，实验以各个元组中的 temperature 属性作为查询对象，并添加 x-tuple_id 和 probabilistic 属性将此数据集扩展为 x-tuple 规则元组的不确定数据集。其中，probabilistic 是元组的存在概率 p，表示元组以概率 p 存在于数据库中。具有相同 x-tuple_id 的元组，属于同一 x-tuple，且概率和小于等于 1。

5.3 结果分析

5.3.1 网络能耗分析

图 6 显示了当 $k=5$，$p=0.5$ 时，集中式不确定数据 TP-Top k 查询算法、BB 算法、SSB 算法及 TPQP 算法在执行 TP-Top k 查询时产生的网络通信开销。从图 6 可以看出集中式查询算法产生的网络通信需要的网络通信开销非常高，这也说明了集中式查询处理方法不适合于无线传感器网络。相对于集中式查询算法，BB 算法、SSB 算法与 TPQP 算法都能够有效地减少网络通信开销，比率分别为 87.35%、86.05%、87.51%。由此，可以得出 BB 算法、SSB 算法及 TPQP 算法在减少网络通信开销方面相差无几。

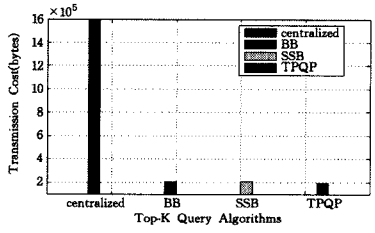

图 6 4 种方法整体数据通信开销对比图

5.3.2 概率阈值 p 对查询的影响分析

执行 BB 算法、SSB 算法、TPQP 算法产生的通信开销与概率阈值 p 之间的关系图如图 7 所示。从图 7 可以看出，随着概率阈值 p 的增大，3 种算法所产生的网络通信开销逐渐减少。一方面，TPQP 算法和 BB 算法产生的通信开销要小于 SSB 算法。另一方面，当概率阈值 p 较小时，TPQP 算法和 BB 算法产生的通信开销相差不大，但是，随着 p 的增大，TPQP 算法产生的通信开销明显少于 BB 算法。因此，仅从查询处理产生的通信开销的角度考虑，TPQP 算法比 BB 算法和 SSB 算法更为有效。

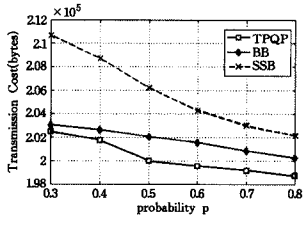

图 7 概率值 p 对 3 种算法数据通信开销的影响

BB 算法、SSB 算法以及 TPQP 算法的查询响应时间与概率阈值 p 之间的关系如图 8 所示。从图 8 可知，随着概率阈值 p 的逐渐增大，3 种查询处理算法的查询响应时间逐渐减少。TPQP 算法要优于 BB 算法和 SSB 算法。其原因：1) TPQP 算法以 Poisson 分布 PT-Top k 查询处理近似算法为基础，而 BB 算法和 SSB 算法以 PT-Top k 查询处理精确算法为基础，相比 PT-Top k 查询处理方法的执行效率，基于 Poisson 分布的 PT-Top k 查询处理近似算法执行效率更高。2) TPQP 算法在簇内就执行数据修剪，使得簇内传输数据量少于 BB 算法和 SSB 算法，在一定程度上导致 TPQP 算法在执

行效率上优于 BB 算法和 SSB 算法。

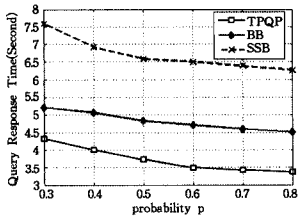

图 8　概率值 p 对 3 种算法查询响应时间的影响

图 9 展现了执行 TPQP 算法时，查询结果的 Top-k 概率绝对平均误差率与概率阈值 p 之间的关系。总体上说，随着概率阈值 p 增大，查询结果的 Top-k 概率误差率也随着减小。从图 9 可以看出，$p=0.3$ 时，误差率最大，其原因是采用基于 Poisson 近似查询算法计算得到的 Top-k 概率会产生误差，查询结果数据量越多误差越大。当 k 不变，$p=0.3$ 时，查询结果的数据量是最多的，导致其查询结果的误差率相对较大。但是，其平均绝对误差率在可接受范围，因此，认为查询结果是有效的。$p=0.8$ 时，TPQP 算法的绝对平均误差率接近于 0。因为能够满足这种高概率的数据较少，而且查询结果在整个数据集中的排序较前，所以其查询结果的概率和精确查询相差无几。

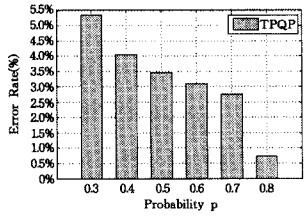

图 9　与 p 相关的 TPQP 算法查询精度

综上，在满足不确定数据 TP-Top k 查询精度的情况下，当概率阈值 p 可变时，相对于 BB 算法和 SSB 算法，TPQP 算法产生的通信消耗最少，查询响应延迟时间最小。

5.3.3　*排序数 k 对查询的影响分析*

执行 BB 算法、SSB 算法、TPQP 算法产生的通信开销和查询响应时间分别与最高排序 k 之间的关系如图 10 和图 11 所示。从图 10 中可以判断出，随着 k 的增大，TPQP 算法在减少通信开销方面优于 BB 算法和 SSB 算法。通过观察图 11 可知，随着 k 值增大，BB 算法、SSB 算法以及 TPQP 算法的查询响应时间也随之增大。但是，TPQP 算法的查询响应时间增速明显小于 BB 算法和 SSB 算法。因此，从查询响应时间方面看，TPQP 算法优于 BB 算法和 SSB 算法。

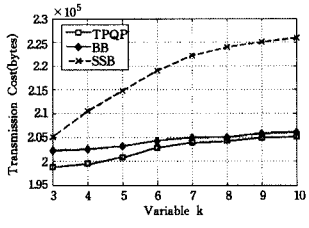

图 10　k 值对 3 种算法数据通信开销的影响

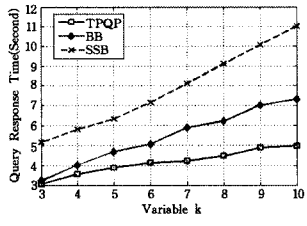

图 11　k 值对 3 种算法查询响应时间的影响

图 12 给出了执行 TPQP 算法时，查询结果的 Top-k 概率平均绝对误差率与最高排序 k 之间的关系。总体上说，随着 k 值的逐渐增大，TPQP 算法查询结果的 Top-k 概率平均绝对误差率也逐渐增大。从中观察到，$k=10$ 时，误差率最大，其原因是使用基于 Poisson 近似查询算法计算得到的 Top-k 概率存在误差，查询结果数据量越多误差也越大。当 p 不变，$k=10$ 时，查询得到的数据数量最多，导致其查询结果误差率相对较大。反之，当 $k=3$ 时，查询的数据很小，而且查询的结果在整个数据集中的排序很高，TPQP 算法的绝对平均误差率最小。

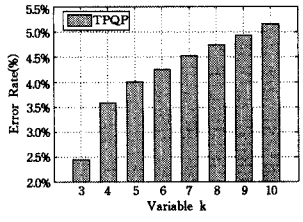

图 12　与 k 相关的 TPQP 算法查询精度

综上，在满足不确定数据 TP-Top k 查询精度的情况下，当排序 k 可变时，相对于 BB 算法和 SSB 算法，TPQP 算法可以在较少的通信开销的情况下，降低查询响应时间。

5.3.4　*数据敏感度查询分析*

图 13 和图 14 分别反映了 BB 算法、SSB 算法及 TPQP 算法查询所需的通信开销和查询响应时间与单位元组大小之间的关系。从图 13 和图 14 可知，随着单位元组不断增大，执行 3 种算法所需要的通信开销和查询响应时间也随之增大。另一方面，执行 SSB 算法产生的通信开销增长最快，BB 算法次之，TPQP 算法最慢。TPQP 算法的查询响应时间也明显小于 BB 算法和 SSB 算法。

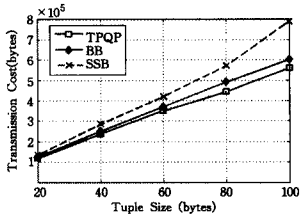

图 13　元组大小对 3 种算法数据通信开销的影响

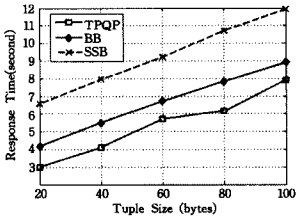

图 14　元组大小对 3 种算法查询响应时间的影响

图 15 说明了执行 TPQP 时，算法查询结果的 Top-k 概率平均绝对误差与单位元组大小之间的关系。可以看出，随着单位元组逐渐增大，TPQP 算法查询结果的 Top-k 概率的平均绝对误差并没有明显的变化，平均绝对误差率始终保持在一个可接受范围内。也就是说，无论元组大小如何变化，TPQP 算法的查询精度是可以保证的。

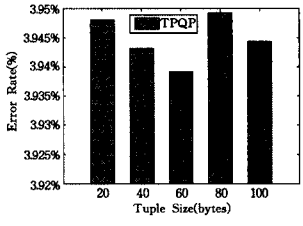

图 15　元组大小对 TPQP 算法平均绝对误差率的影响

综上，无论元组大小如何变化，TPQP 算法的通信开销与查询响应时间优于 BB 算法和 SSB 算法，同时 TPQP 算法可以保证查询精度。

结束语　针对现有的分布式不确定数据 Top-k 查询处理

算法还存在不足，通过分析不确定数据特点，基于 x-tuple 规则元组模型，采用簇内与簇间的两阶段数据查询处理机制，提出基于 Poisson 分布的分布式不确定数据 PT-Top k 查询处理近似算法 TPQP，以达到减少查询相应时间、降低网络开销的目的。仿真实验从总体通信开销、与概率阈值 p 相关分析、与排序数 k 相关分析以及数据敏感度分析等方面，验证了 TPQP 算法在通信消耗、查询响应时间上的有效性。

另一方面，TPQP 算法也有不足之处，其仅适合 x-tuple 规则元组来自同一数据源的情况，当 x-tuple 规则元组来自不同数据源时，TPQP 算法并不适用。同时，概率阈值 p 对查询的影响分析缺乏相应的理论分析，这将在后续工作中展开。

参 考 文 献

[1] Hu C A, Fan L W, Mao Y M. HPDBSCAN: Efficient clustering algorithm for processing uncertain data[J]. Computer Engineering and Design, 2013, 34(3): 1044-1049

[2] Liu X, Yang D N, Ye M, et al. U-skyline: A new skyline query for uncertain databases [J]. IEEE Transactions on nowledge and Data Engineering, 2013, 25(4): 945-960

[3] Ye M, Lee W C, Lee D L, et al. Distributed processing of probabilistic top-k queries in wireless sensor networks [J]. IEEE Transactions onKnowledge and Data Engineering, 2013, 25(1): 76-91

[4] Wang Y, Yu J. A Top-k Query Algorithm on Uncertain Streaming Data [J]. Journal of Computational Information Systems, 2013, 9(13): 5273-5279

[5] Soliman M A, Ilyas I F, Chang K C C. Top-k query processing in uncertain databases[C]// Proceedings of the 23nd International Conference on Data Engineering, Istanbul, Turkey. IEEE, New York, 2007

[6] Nasridinov A, Park Y H. Optimal Aggregator Node Selection in Wireless Sensor Networks [J]. ICCA 2013, ASTL, 2013, 24: 37-39

[7] Sharaf A, Beaver J, Labrinidis A, et al. Balancing energy efficiency and quality of aggregate data in sensor networks[J]. VLDB Journal, 2004, 13(4): 384-403

[8] Silberstein A S, Braynard R, Ellis C, et al. A sampling-based approach to optimizing top-k queries in sensor networks [C]// Proc. International Council for Open and Distance Education. 2006: 68

[9] Zeinalipour-Yazti D, Vagena Z, Gunopulos D, et al. The Threshold Join Algorithm for Top-k Queries in Distributed Sensor Networks[C]// DMSN'05 Proceedings of the 2nd International Workshop on Data Management for Sensor Networks. 2005: 121-1

[10] Fagin R, Lotem A, Naor M. Optimal aggregation algorithms for middleware [C]// Proceedings of Special Interest Group on Management of Data. 2001: 23-33

[11] Bast H, Majumdar D, Schenkel R, et al. Io-top-k: Index-aceess optimized top-k query processing[C]// Very Large Data Base. 2006: 475-486

[12] Das G, Gunopulos D, Koudas N, et al. Answering top-k queries using views[C]// Very Large Data Base. 2006: 451-462

[13] Theobald M, Weikum G, Schenkel R. Top-k query evaluation with probabilistic guarantees[C]// Very Large Data Base. 2004: 648-659

[14] Han Q, Mehrotra S, Venkatasubramanian N. Energy efficient data collection in distributed sensor environments[C]// Proc. Institute of Electrical and Electronics Engineers. 2004: 590-597

[15] http://berkeley.intel-research.net/labdata

[16] Ye M, Lee W, Lee D, et al. Distributed Processing of Probabilistic Top-k Queries in Wireless Sensor Networks [J]. IEEE Transactons on Knowledge and Data Engineering, 2013, 25(1): 76-91

[17] Li J, Saha B, Deshpande A. A unified approach to ranking in probabilistic databases[J]. Pro. Very Large Data Base, 2009, 2(1): 502-513

[18] Ye M, Liu X, Lee W C, et al. Probabilistic Top-k Query Processing in Distributed Sensor Networks[C]// Proc. International Council for Open and Distance Education. 2010

[19] Li J, Saha B, Deshpande A. A unified approach to ranking in probabilistic databases[J]. Pro. Very Large Data Base, 2009, 2(1): 502-513

[20] Sun Yong-jiao, Yuan Ye, Wang Guo-ren. Top-k query processing over uncertain data in distributed environments [C]// Proc. Springer Science Business Media. 2011

[21] Manjeshwar A, Agrawal D P. TEEN: A protocol for enhanced efficiency in wireless sensor network[C]// The 15th Parallel and Distributed Procesing Symp. Sam Francisco: Institute of Electrical and Electronics Engineers Computer Society, USA, 2001

[22] Hua M, Pei J, Zhang W, et al. Ranking queries on uncertain data: a probabilistic threshold approach[C]// Proc. Special Interest Group on Management of Data. 2008

(上接第 47 页)

[9] Timmins J, Hone A, Stibor T, et al. Theoretical advances in artificial immune systems[J]. Theoretical Computer Science, 2008, 403: 11-32

[10] Yang J, Liu X J, Li T, et al. Distrubited agents model for intrusion detection based on AIS [J]. Knowledge-Based Systems, 2009, 22: 115-119

[11] Sobh T S, Mostafa W M. A cooperative immunological approach for detecting network anomaly [J]. Applied Soft Computing, 2011, 11: 1275-1283

[12] Powers S T, He J. A hybrid artificial immune system and self organising map for network intrusion detection [J]. Information Sciences, 2008, 178: 3024-3042

[13] Meisel M, Pappas V, Zhang L. A taxonomy of biologically inspired research in computer networking [J]. Computer Networks, 2010, 54: 901-916

[14] Visconti A, Tahayori H. Artificial immune system based on interval type-2 fuzzy set paradigm [J]. Applied Soft Computing, 2011, 11: 4055-4063

[15] Tsai C F, Hsu Y F, Lin C Y, et al. Intrusion detection by machine learning: a review [J]. Expert Systems with Applications, 2009(36): 11994-12000