不确定数据库中减小可能世界的RPW-kBest查询

孙平平,刘方爱
(山东师范大学 信息科学与工程学院,山东 济南 250014)

摘 要:不确定数据普遍存在于大量应用之中,如在传感器网络、P2P系统、移动计算及RFID(Radio Frequency IDentification)等,研究者已经提出了多种针对不确定数据库的数据模型,其核心思想都源自于可能世界模型。针对可能世界模型能够演化出数量远大于不确定数据库规模的可能世界实例,文中提出一种减小可能世界的RPW-kBest算法,此算法利用概率和评定条件进行筛选,尽可能将不影响查询结果的数据抛弃,使之在最小的搜索空间内完成查询处理过程,以降低存储开销。实验结果表明,此算法能正确的得到查询结果并显著提高查询效率和降低内存使用。

关键词:不确定数据;可能世界;减小;RPW-kBest算法
中图分类号:TP311 **文献标识码**:A **文章编号**:1673-629X(2011)10-0070-03

A RPW-kBest Query Based on Reduced Possible World in Uncertain Database

SUN Ping-ping, LIU Fang-ai
(School of Information Science and Engineering, Shandong Normal University, Jinan 250014, China)

Abstract: Uncertain data arises from a few important applications. Such as wireless sensor networks, P2P systems, mobile computing and RFID technologh. Many data models have been developed, stemming from the core possible world model. For the possible world models contains a huge number of the possible world instances which is far greater than the volume of the uncertain database, so a high-efficiency RPW-kBest Query be proposed, which reduces the possible world and a lower storage cost. The algorithm computes the bound of the probability and filter the data entries as much as possible, which have no chance to influence the query result and process RPW-kBest queries in a smallest search space. Experiments show that the algorithm can process the queries correctly and efficiently improved query efficiency and little memory usage.

Key words: uncertain data; possible world; reduce; RPW-kBest algorithm

0 引 言

在传统数据库的应用中,数据的存在性和精确性均确定无疑。但近年来随着人们对数据采集和处理技术理解的不断深入,不确定数据[1]逐渐被广泛关注并成为研究的热点问题。现实的许多应用,如数据提取、传感器、RFID和社交网络等,数据的不确定性普遍存在,并扮演着关键角色。

在传统关系数据库中,Top-k查询[2]技术相当受重视。该查询返回k个分值最高的元组[3],能够有效提高在数据量很大时的查询效率。对于确定数据库中,元组的分值就是元组排名的依据,因此Top-k查询的语义是确定的。但在不确定性数据库中,数据的发生引入概率值[4],这就大大增加了查询的难度与复杂性。对于数据属性分值和发生概率的问题,以及如何权衡这二者之间的关系是值得我们深入研究的。

不确定数据查询算法和查询语义是数据库领域近两年的热点。在这一领域,研究者已提出了一些重要算法,如文献[5]定义了概率数据集上的排序查询,并提出了两种结合数据分值和可信值的排序查询算法U-Topk和U-kRank,不确定数据概率Skyline查询[6],K近邻查询[7,8],轨迹查询等。

针对于不确定数据处理的特点,以及在海量的数据[9]中,查询关注的只是很小的结果集,因此检索空间是一个值得研究的问题。文中在此分析的基础上,提出一种新颖的基于减小可能世界的RPW-kBest(Re-

收稿日期:2011-03-10;**修回日期**:2011-06-04
基金项目:国家自然科学基金资助项目(90612003);山东省自然科学基金资助项目(Y2007G11)
作者简介:孙平平(1985-),女,山东日照人,硕士研究生,研究方向为信息管理与数据挖掘技术研究;刘方爱,博士生导师,研究方向为网格计算、网络与网络资源管理。

duced Possible World K-Best)查询算法。以下是文中的主要工作。

1)不确定数据模型中可能世界的数量是元组的指数级,文中权衡不确定数据属性分值和发生概率的大小关系,提出一种过滤策略,过滤掉一些不影响查询结果的元组,以减小可能世界空间,降低存储开销,提高查询效率。

2)基于对减小后的可能世界进行查询,提出了 RPW-kBest 查询算法。

3)设计了详细的性能评价实验,结果表明,文中提出的 RPW-kBest 查询算法可以降低存储代价和减小查询时间,提高查询效率,充分地满足实际应用的需求。

1　不确定数据模型

1.1　可能世界模型的概念描述

不确定数据库建模[10]的研究很多,但可能世界模型是应用最广泛的数据模型。一个可能世界是指所有不确定数据分别处于一种状态时的一个组合。可能世界模型包含两个重要的基本概念,一个是置信度,另一个是生成规则,独立和互斥是两种常用的生成规则。

不确定数据库所包含的各元组的任意一种合法组合构成一个可能世界实例,可能世界实例的发生概率是由所包含的元组概率及各元组间的关系所确定的。

1.2　可能世界实例描述

下面通过类似于文献[11]的具体例子来介绍可能世界,并引出文中的研究动机。

在交通系统中,通过雷达自动探测汽车速度,用 OCR(Optical Character Recognition)技术识别车牌号码。由于温度和电压等外界因素都会对雷达造成一定的影响,再者 OCR 可能会因车牌号码不清晰而不能完全正确地识别,所以读取的数据是不一定准确的。表 1 是最近一个小时内读取的汽车速度的一个快照。每个元组包含多个属性,其中"概率"属性是指该元组给出正确信息的概率。由于车速、交通状况等原因,同一辆车不可能在很短的时间间隔内出现在不同的地点。

表 1　不确定数据库

元组	时间	车牌号	地点	速度	概率
t1	9:11	A-777(f1)	L1	160	0.2
t2	9:16	O-321(f2)	L1	150	0.6
t3	9:21	A-777(f1)	L2	130	0.5
t4	9:30	A-777(f1)	L3	90	0.3
t5	9:33	B-567(f3)	L3	70	0.4
t6	9:36	W-621(f4)	L4	35	0.4
t7	9:41	W-621(f4)	L5	25	0.6
t8	9:56	L-111(f5)	L5	20	0.3

因此,Snapshot Readings 中体现元组的独立与互斥。如 t1 ⊕ t3 ⊕ t4,即元组 t1、t3、t4 不能同时发生,但可以同时不发生,类似地,t6 ⊕ t7。

用 PW_j^i 表示可能世界,其中 i 表示可能世界中含有元组的数目,j 表示相应 i 个可能世界中的数量,$N(PW^i)$ 表示含有 i 个元组的可能世界的数目。

考虑元组的互斥不能同时出现在同一个可能世界中,则 $N(PW^0)=C_8^0=1$,$N(PW^1)=C_8^1=8$,$N(PW^2)=C_3^1 * C_2^1 + C_3^1 * C_3^1 + C_2^1 * C_3^1 + C_3^2 = 24$,同理可计算出其它的情况。则可能世界的总数目 $Ns(PW) = N(PW^0) + N(PW^1) + N(PW^2) + N(PW^3) + N(PW^4) + N(PW^5) = 96$。可见仅含 8 个元组的不确定数据库可以演化出数量远远大于其本身数目的可能世界。

2　减小可能世界的策略思想

先滤策略:根据普遍接受的约束条件,先过滤无影响数据源元组,减小可能世界的组成。

假设在一次事故调查中,交警想知道最近一段时间段内行驶速度最快的汽车,则需要查询所有的可能世界的数据库来返回 k 个元组[11]或元组属性,以用来推断最可能肇事车辆。

图 1 所示是一个正态分布的车速概率分布曲线[12]。所罗门得到结论:拥有最低事故率的样本,其速度分布在高于或低于均值的 15% ~20% 的区间内。当速度偏差大于这个范围时,不论其速度高于还是低于均值都出现书顾虑上升的现象。1993 年,Monash 大学交通事故研究中心提出当车速小于均值时,交通事故率只有轻微的增加,即车速低于平均车速时速度离散性与事故率无关。

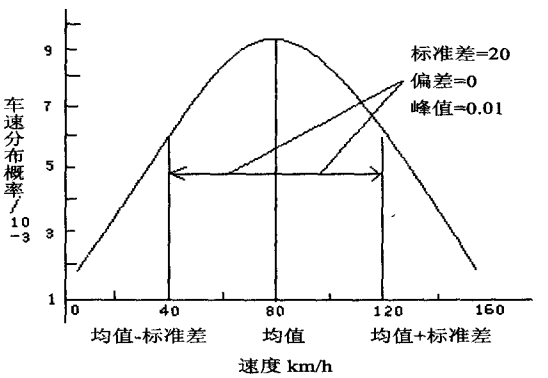

图 1　车速概率分布

则根据上述分析,可看出车速在上述的 3 个不同区段时,相应地对查询结果的影响不同。

文中针对隐含信息提出如下相应的解决方案,以减小可能世界。

a. 当车速高于 120,该元组直接列为可能世界的候选元组。

b. 当车速低于40，概率小于0.2时，直接过滤掉该元组，否则，列为候选元组。

c. 当车速介于两者之间时，不考虑其影响，直接过滤掉此元组。

概率0.1的给出是根据Monash大学交通事故研究中心的结论分析，又概率很小形成的可能世界概率也很小，所以可设定一个小概率作为过滤阈值。

根据上述策略，则不确定数据库过滤后如下。

表2 缩减的不确定数据库

元组	时间	车牌号	地点	速度	概率
t1	9:11	A-777(f1)	L1	160	0.2
t2	9:16	O-321(f2)	L1	150	0.6
t3	9:21	A-777(f1)	L2	130	0.5
t6	9:36	W-621(f4)	L4	35	0.4
t7	9:41	W-621(f4)	L5	25	0.5
t8	9:56	L-111(f5)	L5	20	0.3

可能世界数目为：

$Ns(PW) = N(PW^0) + N(PW^1) + N(PW^2) + N(PW^3) + N(PW^4) = 36$

可以看出在只过滤掉两个元组的情况下，可能世界的数目大大地减少，由此可以降低存储开销，并提高检索速度。

3 RPW-kBest算法

本算法的基本思想是在不确定数据库中的各个元组ti按分值或划分条件高低进行有序排列，因此文中讨论的不确定数据库都已经按分值降序排好的，如图1(按照车速的高低进行降序排列)，查询输出为车牌标识f1～f5。

RPW-kBest算法描述如下。

T[ ]存放元组排列；I[ ]存放元组的车牌标识符；p[ ]存放概率值；C[ ]存放候选元组；Sp[ ]存放元组的速度；As[ ]存放查询结果

1)输入元组根据条件对速度进行计算，过滤元组。

```
for i=0,……,N(元组个数),do
读 Sp[i];
if(40<Sp[i]<=120)过滤掉此元组;
else if(Sp[i]>120)加入后选元组 C[i];
else if(p[i]<0.1)过滤掉此元组;
        else 加入候选元组 C[i];
        i ++;
halt
```

2)由候选元组组成可能世界。

```
for i=0,j=0,……,m, do
读取 C[i];
if 与元组 C[i]互斥的元组未被读取 then
组合成可能世界;
else 与 C[i]互斥的元组被读取 then
        去除此种组合状态;
        j++;//j 为组成可能世界含有的元组数目
    halt
```

3)利用传统的U-Topk查询算法进行查找k个最佳值。

返回k个相应的标识符。

4 实验性能评估

实验环境在一台2.26GHz的Intel处理器，2GB内存的PC机上进行，Windows操作系统，编译环境为jre1.6.0_06。

实验使用业界R-statistical来产生综合数据集实现查询算法。对于一个数据集考察两种算法在该数据集上执行时的运行时间和内存使用情况。本实验结果只对形成可能世界后的查找做了分析比较，U-Topk查询算法在形成可能世界时时间和内存的消耗明显大于RPW-kBest查询算法。

图2和图3反映的是在数据集上随k值的不同，传统的U-Topk和RPW-kBest的查询计算时间和内存消耗。

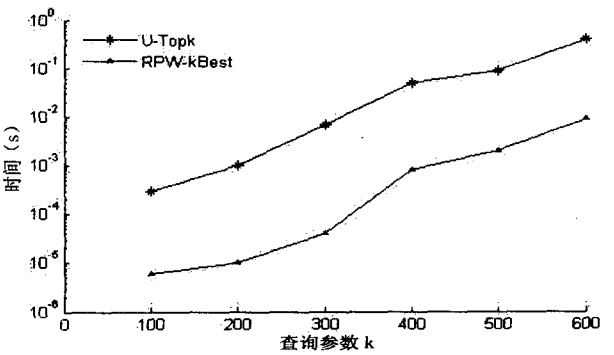

图2 查询计算时间消耗

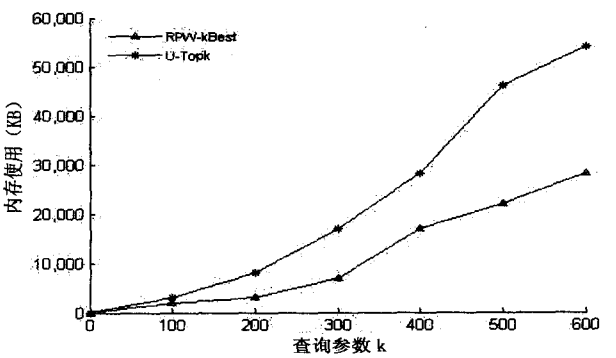

图3 内存使用

实验结果可以看出RPW-kBset查询算法要优于传统的U-Topk算法，查询时间平均小一个数量级以上，在查询参数k增大的情况下，内存使用也明显降低。

(下转第76页)

针对用户评分相对密集的数据，还可以考虑时间差的因素，以进一步提高推荐效果；同时可以考虑文献[12]的方法实现用户属性的加权，以提高聚类的准确性。

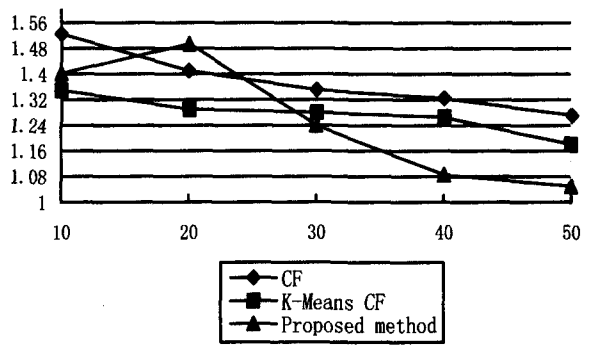

图1 推荐精度的比较

参考文献：

[1] Breese J, Hecherman D, Kadie C. Empirical analysis of predictive algorithms for collaborative filtering[C]//In:Proceedings of the 14th Conference on Uncertainty in Artificial Intelligence (UAI'98). [s. l.]:[s. n.],1998:43-52.

[2] Lee J S,Jun C H,Lee J,et al. Classification-based collaborative filtering using market basket data[J]. Expert System with Applications,2005,29(3):700-704.

[3] Adomavicius G,Tuzhilin A. Toward the next generation of recommender systems:a survey of the state-of-the-art and possible extensions[J] . IEEE Trans on Knowledge and Data Engineering,2005,17(6):734-749.

[4] 李 涛,王建东. 一种基于用户聚类的协同过滤推荐算法[J]. 系统工程与电子技术,2007,29(7): 1178-1182.

[5] 黄国言,李有超. 基于项目属性的用户聚类协同过滤推荐算法[J]. 计算机工程与设计,2010,31(5): 1038-1041.

[6] Deneubourg J L,Goss S,Franks N,et al. The dynamics of collective sorting:Robot-like ants and ant-like robots[C]//Proceedings of the First international Conference on Simulation of Adaptive haviour: From Animals to Animals J. Cambridge, MA:MIT Press,1991:356-365.

[7] 杨 燕,张昭涛. 基于阈值和蚁群算法结合的聚类方法[J]. 西南交通大学学报,2006,41(6):719-742.

[8] 马 良,朱 刚,宁爱兵. 蚁群优化算法[M]. 北京:科学出版社,2008.

[9] Aggarwal C C. On the effects of dimensionality reduction on high dimensional similarity search[C]//Proceedings of the 20th ACM SIGMOD-SIGACT-SIGART. Symposium on Principles of Database Systems. [s. l.]:[s. n.],2001:256-266.

[10] 王明文,陶红亮. 双向聚类迭代的协同过滤推荐算法[J]. 中文信息学报,2008,7(22):61-65.

[11] Sarwar B,Karypis G, Konstan J,et al. Item-Based collaborative filtering recommendation algorithms[C]//In:Proceedings of the 10th International World Wide Web Conference. [s. l.]:[s. n.],2001:285-295.

[12] 李玲娟,李 冰. 一种基于特征加权的蚁群聚类新算法[J]. 计算机技术与发展,2010,20(8):67-70.

(上接第72页)

5 结束语

文中将不确定数据中形成的可能世界进行了缩减，以此为基础进行k个最佳结果查询。通过实验表明文中提出的RPW-KBest算法显著地提高了查询效率，减小内存消耗。由于概率查询算法依然面临很多挑战和尚待解决的问题，因此，下一步的工作综合分析现有查询算法的优缺点，在查询算法上进行深入研究。

参考文献：

[1] 周傲英,金澈清,王国仁,等. 不确定性数据库管理技术研究综述[J]. 计算机学报,2009,32(1):1-16.

[2] Soliman M A, Ilyas I F, Chang KevinChen-Chuan. Top-k Query Processing in Uncertain Databases[C]//2007 IEEE 23rd International Conference on Data Engineering. [s. l.]:[s. n.],2007:15-20.

[3] 孙永佼,王国仁. P2P环境中不确定数据Top-k查询处理算法[J]. 计算机研究与发展,2009,46(S):280-286.

[4] R'e C,Dalvi N,Suciu D. Efficient Top-k Query Evaluation on Probabilistic Data[C]//IEEE 23rd International Conference on Data Engineering. [s. l.]:[s. n.],2007:15-20.

[5] Lian Xiang,Chen Lei. Top-k Dominating Queries in Uncertain Database[C]//Data Engineering. ICDE 2007. IEEE 23rd International Conference. [s. l.]:[s. n.],2007.

[6] Pei J,Jiang B,Lin X,et al. Probalitistic kyline on uncertain data[C]//Proceeding of the 33rd international conference on very large databases. Vienna,Austria:[s. n.],2007.

[7] Huang Y, Chen C, LEEC. Continuous k-nearest neighbor query for moving objects with uncertain velocity[J]. Geoinformatica,2009,13(1):1-25.

[8] 周 帆,李树全,肖春静. 不确定数据Top-k查询算法[J]. 电子测量与仪器学报,2010,30(10):2605-2609.

[9] 韩希先,杨东华,李建中. TKEP: 海量数据上一种有效的Top-K查询处理算法[J]. 计算机学报,2010,33(8):1405-1418.

[10] 周 逊,李建中,石胜飞. 不确定数据上两种查询的分布式聚集算法[J]. 计算机研究与发展,2010,47(5):762-771.

[11] 刘德喜,万常选,刘喜平. 不确定数据库中基于x-tuple的高效Top-k查询处理 算法[C]//第26届中国数据库学术会议论文集(A辑). [出版地不详]:[出版者不详],2009:15-18.

[12] 吴义虎,武志平. 基于平均车速和车速标准差的路段安全方法[J]. 公路交通科技,2008,25(3):139-142.