

面向 RFID 数据流冰川查询的世系追踪模型

刘耀宗 张宏 孟锦 王永利

(南京理工大学计算机科学与技术学院 南京 210094)

摘要 为了获得 RFID 数据流中热门元素以及相关起源的信息,需要对 RFID 数据流进行带起源信息的冰川查询。以 RFID 数据流中单个数据对象的世系追踪为研究对象,分析在海量 RFID 数据流基础上返回极少查询结果的冰山查询执行机理,初步建立一个面向 RFID 数据流冰山查询的数据流世系跟踪原型模型。

关键词 RFID 数据流,世系追踪,冰山查询

Lineage Tracing Model for Iceberg Queries Over RFID Data Stream

LIU Yao-zong ZHANG Hong MENG Jin WANG Yong-li

(College of Computer, Nanjing University of Science and Technology, Nanjing 210094, China)

Abstract In order to get popular elements and the origin of the data in RFID data streams, iceberg queries of RFID data streams with data information on the origin of glaciers query must be made. RFID data lineage to track the application flow needs to know the origin of the data. This paper introduced the data streams lineage tracing techniques, combined with the method of the query RFID data stream, a lineage tracing model for iceberg queries over RFID Data Stream was proposed in this paper.

Keywords RFID data stream, Lineage tracing, Iceberg queries

1 引言

RFID 技术为人们提供了强大的感知、理解和管理物联网世界的能力。RFID 以非视距的方式快速采集数据,能够对带有唯一标识的对象进行高精度的识别,可广泛地用于识别、定位、跟踪、监控物联网中物理对象的应用。

与此同时,RFID 数据采集和管理也提出了许多新的课题:RFID 观测值中包含大量的重复值,需要过滤;RFID 观测值蕴含某种含义,需要转换并聚集成表现为某种数据模型的语义数据^[1];RFID 数据是时态的、流动的、海量的,必须在线处理。因此,将物理 RFID 观测值自动转化为对虚拟世界有用的更为复杂的数据形式,对通用 RFID 数据处理框架具有重要意义^[2],而高层应用到低层数据的反向变换即数据起源的追踪^[3],是更具挑战性的课题。

冰山查询(iceberg queries)的概念是由 Fang 等人^[4]提出来的。冰山查询是在给出大量输入数据元组的情况下,返回极少查询结果的一种查询,特指在一个属性或属性集上计算一个聚集函数,以找出大于某个指定阈值的聚集值。所谓冰山查询是在属性或属性集上计算大于等于给定阈值的聚集函数值。冰山查询也就是对热门元素的聚集查询。RFID 数据流蕴含着丰富的时空信息,对 RFID 数据流进行冰山查询时需要了解热门元素的起源信息,了解这些数据是从哪里来的,

怎么来的。

数据世系追踪(Data Lineage Tracing),也称为数据起源跟踪(Data Provenance Tracing)。主要关注怎样从用户感兴趣的高层视图回溯到导出此视图的原始资源数据的问题。

数据的世系记载了对数据处理的整个历史,包括数据的起源和加于其上的所有后继过程。冰山查询是在给出大量输入数据元组的情况下,返回极少查询结果的一种查询,特指在一个属性或属性集上计算一个聚集函数,以找出大于某个指定阈值的聚集值。与在传统数据仓库中可有可无的地位不同,冰山查询数据世系跟踪是基于 RFID 技术的定位、跟踪应用必需支持的功能,数据分析、数据质量检查等阶段会频繁地对冰山查询的结果进行回溯,其执行效率极大地影响 RFID 数据管理系统响应查询的能力。

基于数据世系跟踪的 RFID 数据流冰山查询的意义在于:冰山查询希望能对单独 RFID 数据条目实现在线跟踪,这为冰山查询提出新的难题。例如,RFID 供应链应用为了能够准确地实时跟踪附有标签的物品,必须在冰山查询结果的基础上对原始数据展开不同层次的推断,这需要与标准的 OLAP 查询一样对单独物品条目的查询进行跟踪、回溯,在合理的时间内解决这些低层(细粒度)查询。

为了揭示 RFID 数据冰山查询的运行机理,需要对如下内容进行详细研究:分析数据流模型下冰山查询与传统静态

本文受国家自然科学基金(60803001),江苏省研究生创新基金(AD20255)资助。

刘耀宗(1975—),男,博士生,主要研究方向为数据流挖掘与网络安全,E-mail: new025@126.com;张宏(1956—),男,教授,博士生导师,主要研究方向为网络安全与人工智能;孟锦(1983—),女,博士生,主要研究方向为网络安全态势评估;王永利(1974—),男,副教授,主要研究方向为 RFID 数据流管理。

关系模型体系下冰山查询执行方法的联系与区别;设计面向 RFID 数据流冰山查询的数据流世系跟踪模型;研究冰山查询经过 RFID 中间件复杂变换前提下,理解 RFID 数据流的数据世系内涵。

2 数据流上冰川查询与世系追踪研究现状

2.1 数据流上的冰山查询研究

在流数据研究领域,热门元素统计一直是倍受关注的问题。热门元素,即出现次数超过用户设定阈值的元素。流数据热门元素,即当前流数据中出现次数超过用户设定阈值的元素。热门元素统计在许多流数据系统中被涉及到,如网络监控就需要统计热门元素。

冰山查询就是对热门元素的聚集查询,流数据长度未知的特点决定了流数据热门元素统计不能采用简单的方法:为每个元素维护一个计数指针;流数据应用的实时性要求决定了热门元素统计必须要在在线统计。综合这些限制,得出一个结论:流数据热门元素统计必然是近似和实时统计。传统热门元素统计方法通常不适用于流数据。

Charikar 等人^[5]为数据流的热门元素统计提出一种计数概要结构。它仅需要通过两次访问流数据,就可以扩展应用到变化热门元素统计。Gibbons 等人^[6]提出两种新的采样方法——压缩采样和计数采样。利用这两种采样技术,可以对数据流中的热门元素进行统计。

在数据流上挖掘出热门元素,文献^[7]根据用户定义的门槛参数 $s \in (0, 1)$,输出在整个数据流中所占比重大于 s 的所有元素。

文献^[8]中提出的定义如下:已知 N 和 ϵ ,设计一种单遍访问算法,用尽可能小的空间来统计最近 N 个流数据元素中的热门元素及其出现次数,最大近似阈值为 ϵ 。最近 N 个元素中的热门元素是指那些出现次数超过 ϵN 的流数据元素。

国内关于数据仓库上的冰山查询的研究文献也较少,文献^[9]给出了在压缩数据仓库中计算 Iceberg Cube 的算法,其性能优于先在压缩数据上计算 Cube 再检查 having 条件的方法。

2.2 数据流上的世系追踪研究

数据世系(又称数据日志)描述一个数据对象精确的历史沿革。其跟踪技术不仅可以回溯一个数据对象的变换过程到其数据源点,还可以验证其清洗、映射规则的正确性。数据世系跟踪技术是斯坦福大学 Whips 项目中 7 个研究方向之一^[10]。

目前数据世系追踪方法有 3 种方法:查询求逆、标注和工作流日志。

2.2.1 查询求逆

查询求逆是在计算时通过对查询或者视图定义进行分析,求逆(执行逆查询)的结果就是视图的世系,由于它是在需要用到数据世系时才进行计算分析,因而又称为“lazy”方式。这种方式主要是在早期把数据世系用于视图维护和更新问题时提出来的。查询求逆计算方法的缺点在于,它不完全适用于复杂查询。假设复杂查询满足一定的条件可以被规范化或

者改写,但实际上并不是所有的查询都是如此;即使满足这些条件求出的数据世系有时也不精确。

2.2.2 标注

在很多情况下,结果数据并不从源数据计算得来,而是通过一些复杂转换或者手工劳动而来,因此,数据世系通过逆查询就不可能计算出来。这时,通常采用的是用标注的方法,即直接在数据上注明其来源。将标注用于数据世系就是在标注中记录一些关于数据出处或者产生过程历史的信息。由于这种方式是在一开始就让数据通过标注携带一些数据世系信息,因而又称为“eager”方式。用标注来存放数据世系信息,在原有数据之外要引入辅助数据-标注,关于标注的组织、管理等有一系列问题有待解决。

现有的系统由于是没有存储标注的,因此首要问题就是如何创建或者获取标注,即标注从何而来。在有些系统中提供了相应的工具集帮助用户创建标注,但是从数据世系的角度来说,RFID 数据流本身就蕴含了适于自动获取的时空信息,怎样为 RFID 数据世系自动标注或手动标注,目前还没有人论述过这个问题。

2.2.3 工作流日志

工作流日志只是基于消息层次,没有足够的语义信息,即使收集到了,对于重塑工作流得到原始数据也很难,所以在实际的操作过程中工作流日志往往起到辅助作用,作为其他两种方式的补充。

有关流数据分析与管理的研究是目前国际数据库研究领域的一个热点^[7]。在过去 30 多年中,尽管传统数据库技术发展迅速且得到了广泛应用,但是它不能处理在诸如电信、传感器网络、金融领域等应用中所生成的一种新型数据,即流数据。流数据的特点是数据持续到达,且速度快、规模宏大,而对流数据世系的研究刚刚起步,最新的研究并不多见^[11]。

流数据的世界追踪与管理有它自身的特点^[11]:首先,通常流数据的到达速度很快,对流数据进行在线标注和保存会给系统的运行带来严重的负担;其次,流数据应用只能单遍扫描数据,这就限制了流数据世系的形式;再次,流数据算法执行都在内存中进行,而维护世系是很占空间的任务;另外,流数据世系应该能够响应用户的连续查询和即时查询;最后,流数据常常是时间序列的类型,一般采用滑动窗口模型处理,即当流数据过期了但是其世系不一定过期。综上所述,在流数据应用场景下进行世系追踪和管理是具有挑战性的任务。必须采用新的技术和方法才能解决流数据的世界追踪与管理问题。

在知道流数据的特定模式信息前提下,可以通过标注数据实例来表示流数据的世界,标注信息可以是源数据项,也可以是源数据项的一个链接。对小规模的简单应用,标注源数据项可能是充分的,而源数据的链接使得元数据信息独立于数据类型和数据大小,因此更适合在流数据场景下应用^[11]。

由于在流数据模式上增加了标注信息表示流数据的世界,因此可以通过数据流管理系统(Data Stream Management System, DSMS)提供的接口进行在线的世界查询^[11]。M. Blount 等^[12]提出了 TVC(Time-Value-Centric)模型以解决

面向由医疗传感器形成的流数据的世系查询。

3 基于世系的 RFID 数据流冰山查询体系结构

3.1 RFID 数据流的数据世系实例

RFID 数据具有不确定性、冗余性、蕴含时空性等特点,由于来源可能不可访问或访问的代价高昂,因此在 RFID 数据仓库中查询数据在世系信息非常困难。解决此问题的一种方法是使用辅助信息找到被请求个体的来源,而不是直接对海量 RFID 数据进行钻取。然而,在 RFID 场景下定义辅助信息提出了许多研究问题,比如,需要何种辅助信息?多少辅助信息?

与常规静态数据库上的世系追踪相比,RFID 数据面临的世系追踪问题更为复杂、世系追踪查询代价更为高昂。RFID 阅读器的大量部署会导致海量标签信息涌入系统,因此必须解决 RFID 数据的喷涌问题,目前公认的方法是在 RFID 数据管理系统中设置数据连接器,包括 RFID 中间件、事件处理与内存数据 cache^[2]。原始数据经过复杂中间过程,增加了追踪的难度。在推导规则不确定性的前提下,使用世系跟踪发现适合物化的推导规则集,是 RFID 数据世系追踪的难点。

RFID 数据流的起源数据由记录号、标签号、强度(Intension,阅读器与标签之间的识别距离,一般分为 32 个等级)、阅读器号与时间戳组成。上述 5 个属性组成了 RFID 数据流的起源信息(见表 1)。

表 1 带起源信息的 RFID 监控数据流(Monitor)

RecordNo	TagID	Intension	ReaderNo	TimeStamp
15 340	12	2	2	2011-7-3 14:23:22
15 341	21	13	3	2011-7-3 14:23:23
15 342	102	24	1	2011-7-3 14:23:24

当带有 RFID 标签的物品进入到 RFID 阅读器的有效范围内,RFID 阅读器就会读取标签,生成相应的 RFID 数据集,从这些数据集中可以生成原始的 RFID 数据集(RawDataSet)(见表 2)。

表 2 带 RFID 标签的物品信息(Product)

TagID	ProductName	Category	Price	Unit
12	HP 6320	Computer	7000	Box
21	Rice	Food	100	Bag
102	IBM T63	Computer	13000	Box

3.2 RFID 数据流在世系追踪查询

已有的数据世系研究大多只是针对科学数据库,未考虑以数据流处理为特征的 RFID 数据世系追踪快速响应的需要,这使得现有的世系跟踪方法很难直接应用到蕴含新特性的 RFID 数据管理,需要从创建新世系追踪模式的高度来解决 RFID 数据世系的跟踪问题。文献[13]研发发现,从海量 RFID 数据库中追溯冰山查询结果是影响系统性能的瓶颈问题。

由于转化后的模型仍然是关系模式,且支持世系查询,因此可以直接利用 SQL 对其进行冰山查询。用 SQL 可以改写

为:

```
SELECT *,count(*) FROM AllComputer where
SELECT Monitor. TagID,Product. ProductName,LINEAGE(Moni-
tor,Product)
FROM Monitor,Product
HAVING count(*)≥2
```

关键词 LINEAGE 表示在查询结果中需要利用查询重写技术标注出结果数据元组的世系信息。

查询求逆是在计算时通过对查询或者视图定义进行分析,求逆(执行逆查询)的结果就是视图的世系,由于它是在需要用到数据世系时才进行计算分析,因而又称为“lazy”方式。这种方式主要是在早期把数据世系用于视图维护和更新问题时提出来的。查询求逆计算方法的缺点在于,它不完全适用于复杂查询。前面的研究中都是假设复杂查询满足一定的条件,可以被规范化或者改写,但实际上并不是所有的查询都是如此,即使满足这些条件求出的数据世系有时也不精确。

在很多情况下结果数据并不是从源数据计算得来,而是通过一些复杂转换或者手工劳动而来,因此,数据世系通过逆查询就不可能计算出来。这时,通常采用的是用标注的方法,即直接在数据上注明其来源。将标注用于数据世系就是在标注中记录一些关于数据出处或者产生过程历史的信息。由于这种方式是在一开始就让数据通过标注携带一些数据世系信息,因而又称为“eager”方式。其用标注来存放数据世系信息,在原有数据之外要引入辅助数据-标注,关于标注的组织、管理等有一系列问题有待解决。

3.3 面向 RFID 数据流冰山查询的世系追踪模型

本文设计了一种分布式事件处理系统,以中央服务方式接受查询请求,在多个分布式查询执行引擎上部署查询,查询各个执行引擎上在各自生命周期时间内执行。系统监控查询引擎上的负载,按照分布式负载动态的需求产生新的查询执行引擎。查询计划根据重用规则、查询和网络代价估计对查询进行优化,将收到的查询分布到有效的查询执行引擎,面向 RFID 数据流冰山查询的世系追踪查询的模型如图 1 所示。

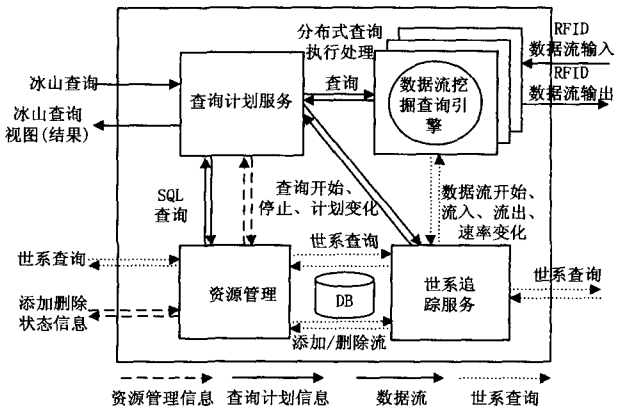


图 1 面向 RFID 数据流冰山查询的世系追踪模型

世系查询信息模型中的基本实体包括数据流、查询、实体的世系信息,数据流由基本流与导出流两种类型组成,基本流源自事件处理系统之外的某一设备、传感器网络或者一个服务,事件处理产生导出流,导出流来自于基本流或其它的导出

流。信息模型由相互关联的实体组成,能够为用户提供充足的信息,用来构建某一异常事件导出历史的跨越时间的结构图。

实体的世系信息分为如下几个部分:在数据流和查询注册器件收集到信息为静态世系,例如元信息是静态世系的一部分;在查询处理期间收集到的信息为动态世系,例如数据流流动速率的变化就是一种动态世系。

资源管理负责维护系统的当前状态,与查询计算进行交互。世系追踪服务对系统中当前资源的变化保持跟踪,同时提供 GUI 监控当前的查询、数据流和计算性资源。用户通过调用世系服务注册输入数据流和冰山查询。当提交新的查询时,系统产生导出流的注册。流/冰山查询注册之后,用户可以向世系数据集添加注释和元信息等附加信息。静态世系数据存储于世系知识库。

在执行期间,查询处理引擎将查询的更新发送到世系服务,当用户感兴趣的事件发生时,查询计划与查询执行引擎更新世系服务。用户感兴趣的事件包括:新查询的开始、由于流集减少或处理节点失败时查询执行计划的改变、某个已注册的查询过期、流速率变化、流/查询近似度与精度的改变等。

结束语 总之,目前缺少科学合理的流数据世系追踪模型,实现逆向跟踪 RFID 数据流的冰山查询存在较大难度。本文以 RFID 数据流的冰山查询中热门元素的数据世系追踪为研究对象,提出了面向 RFID 冰山查询的世系追踪模型。世系可以一直跟踪数据的读取到数据显示的整个过程。通过世系可以标注每一条原始数据流的来源和时间来准确定位异常发生的时间和位置,下一步将针对 RFID 数据流中的异常事件进行原子事件的世系追踪的检测与查询研究。

参 考 文 献

- [1] Wang F, Liu P. Temporal management of RFID data[C]//Proceeding of the 31th International Conference on Very Large Data Base(VLDB05). 2005;1128-1139
- [2] Derakhshan R, Orlowska M E, Xue Li. RFID Data Management: Challenges and Opportunities[C]//Proceeding of IEEE International Conference on RFID 2007. March 2007;175-182
- [3] Benjelloun O, Sarma A, Halevy A, et al. Uldbs: Databases with uncertainty and lineage[C]//Proceeding of the 32th International Conference on Very Large Data Base(VLDB06). 2006
- [4] Fang M, Shivakumar N, Molina H, et al. Computing Iceberg Queries Efficiently[C]//VLDB'1998. 1998;299-310
- [4] Woodruff A, Stonebraker M. Supporting Fine-grained Data Lineage in a Database Visualization Environment[C]//Proc. of the Int'l Conf on Data Engineering(IEEE ICDE). 1997;91-102
- [5] Charikar M, Chen K, Farach-Colton M. Finding Frequent Items in Data Streams[C]//Widmayer P, ed. Proceedings of the 29th International Colloquium on Automata, Languages and Programming. Malaga, Spain; Springer-Verlag, 2002;693-703
- [6] Gibbons P B, Matias Y. New Sampling-based Summary Statistics for Improving Approximate Query Answers[J]. SIGMOD Record(USA), 1998,27(2);331-342
- [7] 金澈清, 钱卫宁, 周傲英. 流数据分析与管理综述[J]. 软件学报, 2004,15(8):1172-1181
- [8] 聂国梁. 流数据统计算法研究[D]. 武汉: 华中科技大学, 2006: 46-48
- [9] 骆吉洲, 李建中, 赵锴. 大型压缩数据仓库上的 Iceberg Cube 算法[J]. 软件学报, 2006,17(8):1743-1752
- [10] 潘定. 持续时态数据挖掘及其实现机制[M]. 北京: 经济科学出版社, 2008;156
- [11] 高明, 金澈清, 王晓玲, 等. 数据世系管理技术研究综述[J]. 计算机学报, 2010,33(3):373-389
- [12] Blount M, Davis J, Misra A, et al. A time-and-value centric provenance model and architecture for medical event streams[C]//Proceeding of the 1st International Workshop on Systems and Networking Support for Healthcare and Assisted Living Environments. San Juan, Puerto Rico, June 11, 2007;95-100
- [13] Wang Yong-li, Qian Jiang-bo, Ma Ran. RFIDSLT: A Data Lineage Tracing Method for Complex Query over RFID Streams[C]//ICEBE'09. Oct. 2009,22;233-240
- [5] Hinneburg A, Keim D. An Efficient Approach to Clustering Large Multimedia Databases with Noise[C]//Proceedings of the 4th ACM SIGKDD. New York, 1998;58-65
- [6] Tung A K H, Hou J, Han J. Spatial Clustering in the Presence of Obstacles[C]//Proceedings of the 17th ICDE. Heidelberg, 2001;359-367
- [7] Han J, Kamber M, Tung A K H. Spatial Clustering Methods in Data Mining: A Survey[C]//Geographic Data Mining and Knowledge Discovery. 2001
- [8] Kohonen T. Self-Organizing Maps[M]. Springer Series in Information Sciences, 2001;30
- [9] Cao Yong-qiang, Wu Jian-hong. Dynamics of Projective Adaptive Resonance Theory Model; The Foundation of PART Algorithm[J]. IEEE Transactions on Neural Network, 2004,15(2):245-260
- [10] Brown D, Huntley C. A Practical Application of Simulated Annealing to Clustering[R]. University of Virginia, 1991
- [11] Tung P C, Zaki Y. CARPENTER: finding closed patterns in long biological data sets[C]//Proceedings of ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. 2003
- [12] Ferhatosmanoglu H, Tuncel E, Agrawal D, et al. High dimensional nearest neighbor searching[J]. Information Systems, 2006,31;512-540
- [13] Zaiane O R, Lee C H. Clustering Spatial Data When Facing Physical Constraints[C]//Proc. of the IEEE International Conference On Data Mining. Maebashi City, Japan, 2002;737-740
- [14] Li Yi-fan, Han Jia-wei. Clustering Moving Objects[C]//ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. 200;492-110

(上接第 145 页)