

数据集成中的不确定性问题

孙 辉

1、引言

数据管理中的不确定问题（uncertainty）其实并不是一个新问题，早在十几年前的论文中就对此有所提及，例如上世纪八十年代末开始出现的概率数据库（probabilistic database）研究^[1,2,3]，这一研究认为元组在数据库中的存在具有不确定性、属性值具有不确定性、查询应答也具有不确定性。但是，一直以来，人们对不确定性问题认识不足，这也决定了人们对待不确定数据管理的态度，很多研究工作虽然遇到了不确定性问题，但往往采取传统的“去除不确定性”方法避开对不确定数据的管理。

近两年来，不确定性问题逐渐引起了人们的广泛关注和兴趣，人们开始承认数据不确定性的本质，VLDB, SIGMOD 等数据库领域重要国际会议上相继出现了这方面的相关论文，VLDB2007 上还专门举办了不确定数据管理的 workshop。

数据管理中的不确定性问题之所以引起人们的普遍重视，成为一个新的研究焦点，主要下面几个方面的原因^[1,4]：

第一，应用的需要。随着计算机网络的飞速发展和信息化的推进，全球的数据量正在以指数的趋势迅猛增长，不断增长的数据对数据管理提出很多新问题，新应用也不断涌现，有些应用需要对不确定数据进行管理。

数据集成（Data Integration）是不确定数据管理最重要的应用^[5]。1996 年，Alon Halevy 等人在 VLDB 国际会议上发表题为《Querying Heterogeneous Information Sources using Source Descriptions》^[6]的论文，这篇文章 2006 年被评为 VLDB 十年最佳论文，文中提出一个数据集成系统——Information Manifold, Information Manifold 和其他相关研究极大地促进了数据集成的发展，并导致了一系列数据集成系统商业产品的诞生。过去十几年的时间内，数据集成一直是数据管理领域的研究热点，相关方面的研究取得了很大进展。目前，数据集成发展到一个新的阶段，不确定性和数据血统（data lineage）问题开始引起研究者的注意^[5]：从本质上说，数据集成系统就是管理多个数据源的数据。来自外部数据源的数据是不确定的，数据处理过程也会产生一些不确定的结果，这是数据集成系统的两个重要特点。如果数据库系统能够为不确定数据以及它们的血统进行建模和处理，那么传统数据库系统与数据集成系统差别也就几乎不存在了。

近几年开始兴起的数据空间（dataspace）研究^[6]对也对不确定数据管理提出很高的要求。数据空间定义为一个实体所拥有的所有数据的集合。数据空间与实体一一对应，数据具有时空特性，其空间特性表现在数据可以来自多个分布的数据源；时间特性表现在数据空间的不断演化。数据空间本质上是数据集成问题，数据集成中不确定性问题数据空间中同样存在。除此之外，数据源不确定性是数据空间自身的主要特征之一，主要包括三个方面：数据源的未知性、数据源位置的不确定性、数据内容的不确定性。由于数据源分布在多台计算机上。这些数据源的物理位置和逻辑位置往往是不确定的，当用户不清楚数据源或者没有提供时，数据空间有责任发现和探测数据源的所在，以此作为提供其他服务的基础。数据源的不确定性给数据空间带来了新的挑战。

第二，数据的需要。当今数据管理系统所要面对的数据已经不再局限于确定性的企业数据，而要处理很多非传统方式产生的数据，这些数据往往是不准确的，具有不确定性的本质。例如，信息抽取系统从文本中自动抽取的数据通常是不准确的；Google Base, Flickr 等系统中收集的大众数据具有不确定性，因为在这些系统中，人们可以随心所欲解释自己的数据；位置信息服务（Location-Based Service）中管理的位置信息是不确定的，因为对象的位置是不断变化的，数据库中保存的位置信息不一定是实时的；传感器（sensor）收集的描述物理世界的的数据也是不准确的，这是由 sensor 网络的本质决定的；日益普及 RFID 电子标签数据也具有不确定性，因为电子标签的识别存在错误率。数据的不确定性本质要求我们对不确定性进行建模和管理。

第三，数据库技术的发展推动了不确定数据管理的研究。对于不确定数据的建模和处理要比管理确定性数据复杂很多，实现起来非常困难，因此，以前的工作在碰到不确定性问题的时候，通常采取回避的态度，试图通过将不确定问题确定化的方法将问题简化。随着数据库查询处理新技术的出现，管理不确定数据才逐

步成为可能。

2、数据集成中的不确定性问题

2.1 不确定性问题

数据集成系统中的不确定性问题可以分为三个层次（见图 1），即数据本身的不确定性、模式匹配的不确定性和查询处理的不确定性^[4]。

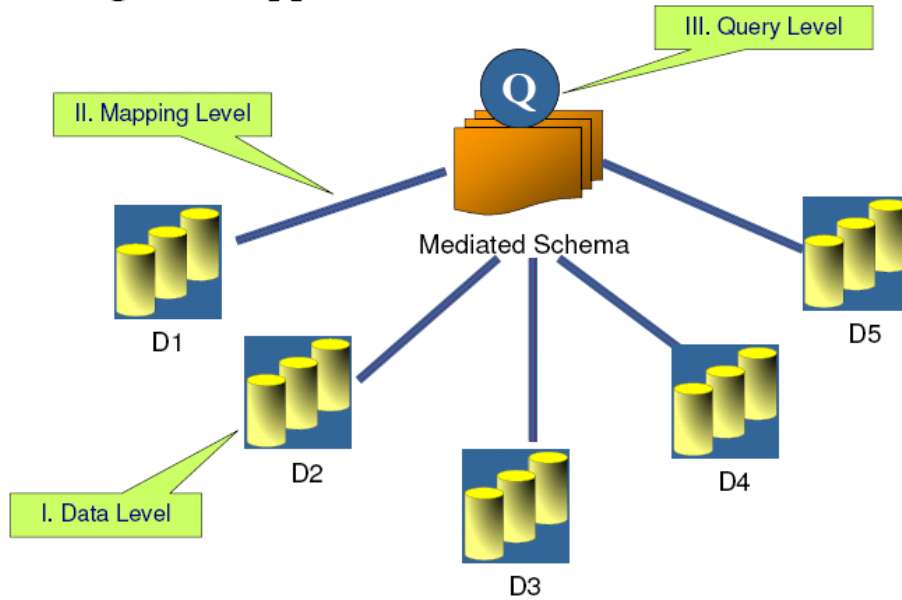


图 1 数据集成中的不确定性问题^[4]

首先，数据本身是不确定的。数据集成系统处理的数据多种多样，有些数据本身就具有不确定性，例如通过信息抽取（Information extraction）系统技术以自动的方式从文本或者半结构化的数据源中抽取的数据，由于抽取技术所限，这些数据通常是不准确的；还有一些数据是从在线数据源中抽取的，数据集成系统很难保证所抽取数据的可靠性和实时性。

其次，模式匹配（schema mappings，也称为语义映射，即 schema matching）是不确定的。数据集成系统一般都是基于中介模式（mediated schema）的，即先建立中介模式与数据源之间的语义映射（schema matching）关系，并通过这种语义映射将用户提交到中介模式上的查询转换为具体数据源上的查询。遗憾的是，中介模式与数据源之间的语义映射关系往往是不准确的，有些应用甚至不可能得到准确的语义映射关系，例如生物信息领域，由于人们对该领域的认识有限，根本就无法确定正确的语义映射。另外，模式匹配也可能是依赖于具体数据的，即数据源与中介模式中介的模式匹配方式本身就是不确定的，需要根据具体数据的特点来确定。

第三，查询的不确定性。数据集成的很多应用中查询通常都是以关键字的方式提交的，这种查询方式不同于传统的结构化查询，其本身存在着不确定因素：一是关键字表达的查询内容不确定，用户很难通过关键字清楚的表达自己的真实意图，系统通常将关键词查询转化为一些可能的结构化查询，提交到具体的数据源，这一转化过程是不确定的；第二，查询结果也是不确定的，关键字查询返回的结果可能很多，究竟哪些结果是才是用户真正想要的，系统需要对查询结果给出不确定程度的评价。

2.2 系统框架

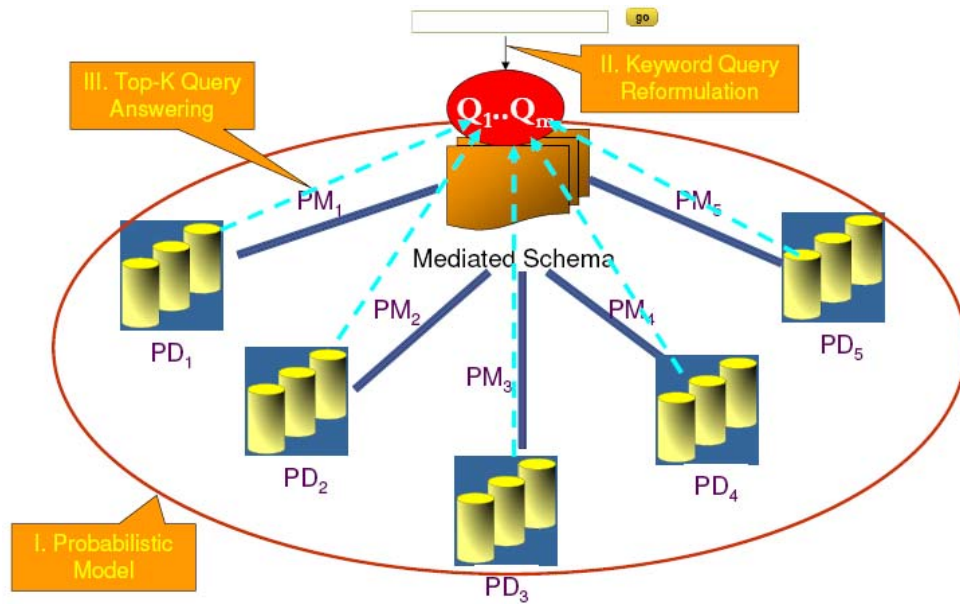


图2 解决数据集成中不确定性问题的系统框架

针对数据集成系统中存在的三个层次的不确定性问题，Dong^[6]提出了一个能够处理不确定性问题的数据集成系统的框架。

首先，与传统数据集成系统不同，新的系统框架是基于概率数据模型（probabilistic data model）的，这里所谓的概率数据模型包含两层意思，一是新系统中处理的数据是有概率的，即每个元组都附有一个概率值；二是中介模式与数据源之间语义映射关系也是附有概率值的。这些概率值将会用于对查询结果进行排名。

在查询转化方面，传统数据集成系统中只需要将中介模式上查询转化为具体数据源的上查询，而在新的系统框架下，基于关键字的查询首先要转化为一系列可能的结构查询，称为关键字查询重写（keyword query reformulation）。

在查询应答方面，这也不同于传统数据集成系统向用户返回所有的查询结果，新的系统框架下，通常只要将前 k 个查询结果返回给用户（Top-k query answering），如何对这个 k 个查询结果进行合理有效的排名也是系统需要关注的问题。

3、研究现状以及相关工作

从介绍数据集成研究的综述文章中看出：目前，数据集成系统中不确定性问题的相关研究还并不太多。2005 年的综述文章首次提到数据集成中模式匹配的不准确问题，但是并没有明确将这一问题提升为不确定性问题；2006 年的综述文章首次明确提出了不确定性问题和数据血统问题是数据集成系统面临的挑战之一。

正如前文提到的，2007 年 Xin Dong 等人发表在 VLDB 上的文章《Data Integration with Uncertainty》首次明确定义了数据集成系统中存在的三个层次的不确定性问题，并且针对中介模式与数据源之间模式匹配的不确定性问题，提出了 by-table 和 by-tuple 两种基于概率的模式匹配（Probabilistic schema mappings），给出在这种模式匹配方式下的查询处理方法以及相应的时间复杂度。

其他的一些工作中虽然也提到了数据集成中不确定性问题，但大多数的作者还是仅从数据集成过程中的具体任务考虑。Gal^[7]提到了基于概率的语义映射，文中用半自动方法得到的前 k 个语义映射，并利用这前 k 个映射的结果提高最终语义映射的准确率。这一方法虽然考虑语义映射中的不确定性，但最终还是去除了语义映射的不确定性，选择了一个最优的语义映射，并没有在整个数据集成系统中管理语义映射的不确定性，并在此基础上产生查询应答。

一提到不确定性，人们很自然的能够想到概率，概率是描述不确定性的一种手段，图 2 描述的数据集成系统框架就是基于概率数据模型的，我们可以参考概率数据库（probabilistic database）领域的研究思路来处理数据集成中的不确定性问题。概率数据库中的数据本身是带有概率的，查询结果也是具有概率的。目前，

概率数据库的研究^[1,2]主要集中查询处理上,包括支持概率数据的查询语言、各种类型的查询处理、查询结果的排序、关键字查询、Top-k查询应答等。概率数据库相关研究中的关键词查询技术、Top-k查询应答技术可以用于解决数据集成系统中查询不确定性问题。

4、研究展望

目前,数据集成中不确定问题的研究还处于起步阶段,有很多研究问题亟待解决,具体包括:

- 1、**数据集成中不确定性问题的定义与形式化**。除了[4]中提到三个方面不确定性,数据集成的过程中是否还存在其他的不确定性?例如查询结果抽取、合并的过程中有没有不确定性?
- 2、**数据集成系统中不确定数据的建模**。目前,我们通常通用概率去描述不确定数据,面对一个具体的数据,描述其不确定程度的概率值如何得到?
- 3、**数据不确定性与数据血统的关系**。数据的不确定性往往与数据来源密切相关,考察数据血统能否对处理数据不确定性问题有所帮助?
- 4、**不确定数据的查询优化**。相对确定数据的查询,不确定数据的查询处理要复杂很多,如何优化不确定数据的查询处理是实现实用系统的关键。
- 5、**数据集成系统各个组成部分的不确定性对查询结果的影响**。数据集成系统在数据处理的每一个过程都可能出现不准确的结果,前一模块不准确的结果对后一模块的结果有何影响,对最终的查询结果有何影响?
- 6、**数据空间中不确定问题的研究**。数据空间本质是一个数据集成问题,但是它自身的特点是否引入特殊的不确定性,如何解决?

5、结束语

数据的不确定本质和数据集成过程中的不确定因素决定了数据集成系统具有不确定性。我们必须承认数据集成中的不确定性问题,保留并处理不确定的中间结果,而不是通过种种确定化的手段去除不确定性,只有这样真正解决数据集成中的不确定问题。

参考文献

- [1] N. N. Dalvi and D. Suciu, Management of Probabilistic Data Foundations and Challenges, In PODS, 2007.
- [2] D. Suciu and N. N. Dalvi. Foundations of probabilistic answers to queries. In SIGMOD, 2005.
- [3] O. Benjelloun, A. D. Sarma, A. Y. Halevy, and J. Widom. ULDBs: Databases with uncertainty and lineage. In VLDB, 2006.
- [4] X. Dong, A. Halvey and C. Yu. Data Integration with Uncertainty, In VLDB 2007.
- [5] http://alanhalevy.blogspot.com/2007_01_01_archive.html.
- [6] 数据空间:一种新的数据管理技术。孟小峰,李玉坤,张相於。计算机通讯,2007.8
- [7] A. Gal. Managing uncertainty in schema matching with Top-K schema mappings. Journal on Data Semantics, 6, 2006.