# §0.2.1.1 A Survey of Multidimensional Modeling Methodologies

## 引用:

Romero, O. , and A. Alberto . "A Survey of Multidimensional Modeling Methodologies." International Journal of Data Warehousing & Mining 5.2(2009):1-23.

---

## 论文获取:

- A Survey of Multidimensional Modeling Methodologies.pdf @google drive

---

## 摘要:

---

## 正文:

**Kimball 等人(1998)** 介绍了我们现在所知的多维建模。此外，他们还介绍了一种推导多维模式的方法。作为第一种方法，它没有提供明确的设计程序，而是提供了详细的提示指南，以识别多维概念并产生多维模式。该方法论的介绍是非常非正式的，它依赖于示例而不是明确的规则。 Kimball 的方法遵循需求驱动的框架来推导出数据仓库的逻辑设计。

Ø 首先，该方法要求我们命名所有可能需要构建的数据集市。数据集市被定义为相关事实的实用集合，但它不一定是排他性的。不考虑数据源，仅建议查看数据源以找出我们可能感兴趣的数据集市。

Ø 下一步旨在列出每个数据集市的所有可能的维度。此时，建议构建一个临时矩阵来捕获我们的多维需求。行代表数据集市，而列代表维度。给定的交叉标记表示数据集市在该处拥有相对应的维度。如果我们查看数据集市共享哪些维度，则该矩阵还用于显示数据集市之间的联系。

Ø 一旦我们选择了数据集市，第三步使用四步法来设计每个事实表：

Ø 首先，我们声明细节粒度。尽管在此过程中可以重新考虑，但建议在开始时由设计团队声明。通常，它将由主要维度确定。

Ø 接下来，我们为特定的事实表选择应该针对所选粒度进行测试的维度。**这一定是一个创造性的步骤**，因为我们需要在不同的模型和不同的文档中寻找维度的片段（即级别和描述符），这最终会导致一项耗时的任务。此时，也建议选择大量的描述符来填充/表达维度。

Ø 最后一步是在声明的粒度的上下文中添加尽可能多的度量。

Cabibbo & Torlone (1998) 提出了最常被引用的设计方法之一。这种方法从 ER 图生成逻辑模式。此外，它可能会产生关系数据库或多维数组方面的多维模式。乍一看，这种方法可能被认为是供应驱动的，因为它对数据源进行了深入分析，但没有给出正式的规则来识别数据源中的多维概念。事实上，多维概念必须由用户手动识别，因此需要从需求中识别。出于这个原因，我们认为它遵循混合框架。总的来说，就像 Kimball 的方法一样，这种方法非常不正式。然而，这些方法奠定了后来被其他方法使用的基础。该方法包括四个步骤。第一步和第二步旨在确定事实和维度并重构 ER 图。这两个步骤可以并行执行并受益于每个步骤检索到的反馈。事实上，作者建议以迭代方式执行它们以改进得到的结果。但是，作者没有给出关于如何识别事实、度量和维度的线索，必须从最终用户的需求中识别它们。一旦它们被识别，每个事实都被表示为一个实体。接下来，我们添加可能在模式中缺失但可以从外部源或与我们的数据源相关联的元数据派生的感兴趣的维度。此时，它还必须通过以下转换来细化每个维度的级别：替换多对多关系，添加新概念以表示新的兴趣级别，为每个级别实体选择一个简单的标识符并删除不相关的概念。最后，第三步和第四步旨在推导出多维模式。为此，作者给出了一些线索来导出将直接映射到多维模式的多维图。

Golfarelli & Rizzi (1998) 提出了该领域的一种参考方法。他们提供了多维设计过程的一般概述，其中包含他们以前的工作，例如 (Golfarelli & Rizzi, 1998a)。这种方法提出了一种形式化和结构化的方法，该方法部分可自动化，由六个明确定义的步骤组成。但是，第四步旨在估算超出本调查范围的数据仓库工作负载：

第一步分析底层信息系统并生成概念模式（即 ER 图）或逻辑模式（即关系模式）。第二步收集和过滤需求。在此步骤中，确定事实很重要。作者给出了一些提示，以从 ER 图（实体或 n 元关系）或关系模式（经常更新的表是很好的候选者）中识别它们。第三步从前面步骤中确定的需求和事实中派生出多维概念模式，该步骤可以如下半自动地执行：

Ø 构建属性树：根据事实的主键，我们通过函数依赖关系创建树。因此，树的给定节点（即属性）在功能上决定了它的后代。

Ø 修剪和嫁接属性树：必须修剪和嫁接树属性，以消除不必要的细节层次。

Ø 定义维度：必须在属性树中根顶点之间选择维度。

Ø 定义度量：度量是通过应用于树的数字属性、根级别的聚合函数来定义的。

Ø 定义层次结构：属性树显示了层次结构的合理组织。层次结构必须源自每个节点与其后代之间的一对一关系。

最后，最后两个步骤（即第五、六步）导出逻辑（通过将每个事实和维度转换为一个关系表）和物理模式（作者给出了一些有关索引的提示，以在 ROLAP 工具中实现逻辑模式）。

该方法的第四步旨在估计数据仓库的工作量。 作者认为，此过程可用于检查在第三步中生成的概念模式的正确性，因为只有在正确定义度量和层次结构的情况下才能表达查询。但是，作者并没有为此提供更多信息。

Boehnlein & Ulbrich-vom Ende (1999) 提出了一种从 SER（结构化实体关系）图中导出逻辑模式的混合方法。 SER 是 ER 的扩展，用于可视化对象之间的存在依赖关系。出于这个原因，作者认为 SER 是识别多维结构的更好选择。这种方法有三个主要阶段：

Ø 步骤 0：首先，我们必须将 ER 图转换为 SER 图。

Ø 步骤 1：必须从目标中识别业务度量。例如，作者建议寻找业务事件以发现适当的度量。一旦确定了业务度量，它们就会被映射到 SER 图中的一个或多个对象。最终，这些度量将产生事实。

Ø 步骤 2：SER 图的层次结构有助于识别潜在的聚合层次结构。维度和聚合层次结构是通过直接和传递函数依赖关系来标识的。作者认为，**发现维度是一项创造性任务，必须辅以对应用领域的良好了解**。

Ø 步骤 3：最后，推导生成星形或雪花模式，使用其分析维度的主键创建事实表，并相应地对聚合层次结构进行反规范化或规范化。

Mazón 等人 (2007) 提出了一种半自动混合方法，首先从用户需求中获取概念模式，然后通过查询/查看/转换 QVT 关系方法来验证并确保其对数据源的正确性。 他们的方法适用于 i* 框架中表达的关系源和需求。 这种方法从需求分析阶段开始。 他们引入了一个详细的需求驱动阶段，作者认为用户应该根据业务目标在一个高度抽象的级别上陈述他/她的需求，并从信息化业务目标中导出信息需求。 目标和信息需求应通过 i* 框架的适配/改编进行建模。 多维概念模式必须从需求派生出来，并用作者提供的 UML 扩展来表达。

接下来，他们提出了最后一步来检查概念多维模型的正确性。此步骤的目标有两个：它们基于多维范式呈现一组 QVT 关系，以将源自需求的概念模式与数据源的关系模式对齐。因此，输出模式将捕获源的分析潜力，此外，它们将根据 MNF 进行验证。 本文中使用的 MNF 是 (Hüsemann, Lechtenbörger & Vossen, 2000) 中使用的 MNF 的演变，它们具有相同的目标。 沿着可能半自动化的五个 QVT 关系，该文章描述了概念多维模式应该如何与底层关系模式对齐：

Ø 1MNF (a)：概念模式中的函数依赖必须在关系模式中具有相应的函数依赖。

Ø 1MNF (b)：源数据库中包含的维度级别之间的函数依赖必须表示为概念模式中的聚合关系。 因此，它们用源中包含的附加聚合层次结构补充了概念模式。

Ø 1MNF (c)：必须在概念模式中标识可以从其他度量计算的度量。 因此，他们支持派生的措施。

Ø 1MNF (d)：必须以事实的原子级别形成一个键的方式将度量分配给事实。 换句话说，它们要求将度量置于具有正确基(base)的事实中。

Ø 2MNF 和 3MNF：当数据源中的 NULL 结构不能保证完整性时，这些约束要求使用概念的特化。

Song 等人 (2007) 提出了一种自动化的数据驱动的方法，它从 ER 模型中导出逻辑模式。 这些方法提供了一种通过连接拓扑值 (connection topology value, CTV) 从 ER

图中自动识别事实的新方法。 这种方法的主要思想是事实和维度通常通过多对一关系相关联。 处于关联关系中多侧的概念是事实候选，一侧的概念是维度候选。 此外，它还区分了直接和传递的多对一关系：

Ø 首先，这种方法需要一个预处理将 ER 图转换为二元（即没有三元或多对多关系）ER 图。

Ø 实体的 CTV 是直接和间接多对一关系的拓扑值的复合函数，其中直接关系相对于传递关系具有更高的权重因子。 因此，所有那些 CTV 值高于阈值的实体都被提议为事实。请注意，事实是由他们的 CTV 识别的，因此，可以考虑无事实的事实。

对于每个事实，实体维度通过多对一关系来标识。此外，作者建议使用 Wordnet 和带注释的维度（代表业务流程中常用的维度）来丰富所描述的聚合层次结构。

- MDA, Model Driven Architecture, 模型驱动架构

Hüsemann 等人 (2000) 提出了一种以多维范式 (MNF) 导出多维模式的需求驱动方法。这项工作引入了一组约束，该方法生成的任何多维模式都将满足这些约束。此外，尽管这种方法产生了概念模式，但他们也认为设计过程必须包括四个连续的阶段（需求获取以及概念、逻辑和物理设计），**就像在任何经典的数据库设计过程中一样：**

需求分析和规范：尽管有人认为操作 ER 模式应该提供基本信息来确定多维分析潜力，但没有给出关于如何识别数据源中的多维概念的线索。业务领域专家必须战略性地选择相关的操作数据库属性，并指明将它们用作维度或是度量。

Ø 概念设计：此步骤将半正式的业务需求转换为正式的概念模式。这个过程分为三个连续的阶段：

n 度量的上下文定义：此方法需要为每个度量确定一个基（即从功能上确定度量值的最小维度级别的集合）。此外，共享相同基的度量被分组到相同的事实中，因为它们共享相同的维度上下文。

n 维度层次设计：从这一步确定的每一个原子维度层次，通过函数依赖逐步形成维度层次。描述符和级别根据需求进行区分，并且根据此分类，它们也区分简单和多个（至少包含两个不同的聚合路径）层次结构。此外，在聚合数据时，必须考虑维度的特化以避免结构化 NULL 值。

n 约束可汇总（概括）性的定义：作者认为，某些维度上的某些度量聚合没有意义。因此，他们建议在概念模式的附录中区分有意义的度量聚合与无意义的聚合。

最后，作者认为通过这种方法导出的多维模式是多维范式 (MNF) (Lehner, Albrecht & Wedekind, 1998)，因此它具有完全的多维意义；也就是说，我们可以产生一个没有可汇总性问题的数据立方体（即多维空间）。

Moody & Kortink (2000) 提出了一种从 ER 模式开发多维模式的方法，这是文献中引入的第一个数据驱动方法之一，也是该领域引用最多的论文之一。 尽管这不是第一种适用于 ER 模式的方法，但它们提出了一种结构化和形式化的方法来开发逻辑模式。 他们的方法论分为四个步骤：

Ø 预处理：此步骤开发企业数据模型（如果尚不存在）。

Ø 第一步：这一步将 ER 实体分为三个主要组类：

n 事务实体：这些实体记录有关业务中发生的特定事件（订单、销售等）的详细信息。他们认为，这些是数据仓库中最重要的实体，构成了星型模式中事实表的基础，因为这些是决策者想要分析的事件。 尽管作者没有考虑需求，但他们强调了识别事实与需求的相关性，因为并非所有的事务实体都会引起用户的兴趣。 此外，它们提供了找到此类实体的关键特征：它描述了在某个时间点发生的事件，并且包含可以汇总的度量或数量。

n 组件实体：这些实体通过一对多关系与事务实体直接相关，它们定义每个业务事件的细节或组件。这些实体将产生星型模式中的维度表。

n 分类实体：这些实体通过一对多关系链与组件实体相关联。换句话说，它们在功能上直接或传递地依赖于组件实体。它们将表示多维模式中的维度层次结构。

Ø 第二步：塑造维度层次结构。作者提供了一些形式化的规则来识别它们。具体而言，维度层次结构被定义为通过一对多关系连接在一起的实体序列，所有实体都在同一方向上对齐。

Ø 第三步：事务实体将产生事实，而维度层次结构将产生它们的分析视角。作者引入了两种不同的运算来生成逻辑模式：

n 折叠层次结构：层次结构中的较高级别可以折叠为较低级别。它是数据仓库中使用的一种非规范化形式，用于提高查询性能。

n 聚合：这可以应用于事务实体以创建包含汇总数据的新实体。为此，一部分属性被选中进行聚合，另一部分属性被选择作为聚合的参照。

对应于这些运算，该方法引入了五种不同的维度设计选项。根据结果模式的非规范化级别和数据的粒度，他们引入规则来派生平面模式、梯级模式、星型模式、雪花模式或星团模式。他们还引入了星座模式的概念，该概念被定义为一组具有分层链接的事实表的星型模式。

Bonifati 等人(2001) 提出了一种由三个基本步骤组成的混合半自动方法：需求驱动阶段、供应驱动阶段和集成的第三阶段。最后一步旨在整合和协调两种范式，并生成最能反映用户需求的可行解决方案。这种方法生成了一个逻辑多维模式，它是第一个引入正式混合方法的方法，该方法具有协调两种范式的集成步骤。此外，该方法已在实际案例研究中得到应用和验证：

Ø 在这种方法中，我们首先**通过访谈收集最终用户需求**，并通过目标/问题/指标范式表达用户期望(Goal/Question/Metrics, GQM)。 GQM 由一组表单和指导方针组成，分为如下四个阶段：一种模糊的方法用于抽象术语制定目标，访谈中使用形式和详细指南确定目标，通过折叠具有相似性的目标整合目标并精简目标的数量，最后，对每个目标进行更深入的分析和详细描述。接下来，作者提出了一个非正式的指南，以从需求中导出逻辑多维模式。作者给出了一些线索和提示，以从过程中使用的表单和表格中识别事实、维度和度量。

Ø 第二步旨在从描述操作型数据源源的 ER 图中执行数据驱动的方法。这一步与前一步并行执行，可以自动执行，并对数据源进行详尽的分析。从 ER 图中，创建了一组将产生星型模式的图，如下所示：

n 他们根据潜在的事实实体具有的附加属性的数量来**标记**它们。每个识别出的事实都被视为图的中心节点。

n 维度通过中心节点的多对一和一对一关系来标识。

**最早期的方法**试图为多维建模提供上下文，提供有关如何设计多维数据仓库的提示和非正式规则。换句话说，他们提出了支持多维设计的最早的指南。**后来**，当关于多维建模的主要概念被建立时，新的正式和强大的方法被开发出来。这些新方法专注于流程的形式化和自动化。自动化是整个数据仓库生命周期中的一个重要特性，多维设计也不例外。事实上，最早期的方法是逐步指导，但随着时间的推移，已经提出了许多半自动和自动方法。这些演变也限制了所使用的输入类型，并且考虑的是**逻辑模式**而不是概念模式。如今，最后引入的方法具有高度的自动化。此外，我们可以说这种趋势也推动了范式的变化。一开始，大多数方法都是由需求驱动的，或者在混合方法的情况下，它们更重视需求而不是数据源。然而，随着时间的推移，数据源变得相关。这是有道理的，因为**自动化与关注数据源而不是需求紧密相关**。因此，最新的方法（高度自动化）大多遵循供应驱动的框架。**然而，设计多维数据仓库的理想方法必须是混合方法，这一点已经被很好地假设了。**在这方面，一些工作以某种方式自动化了他们的需求驱动阶段。

通过比较分析，我们还可以了解到**多维模型的考虑方式的演变**。早期的方法用于生成逻辑多维模式，但随着时间的推移，它们中的大多数会生成概念模式。这种情况的一个原因可能是 Kimball 在逻辑级别引入了多维建模作为特定的关系实现。**随着时间的推移，有人认为有必要在独立于平台的级别生成模式**，实际上，多维设计应该像关系数据库领域一样跨越三个抽象级别（概念、逻辑和物理）。

关于**数据源类型**，大多数早期方法选择了描述数据源的概念实体关系图。 ER 图是表示操作数据库（填充数据仓库的最常见的一种数据源）的最广泛的方式，但是自动化这个过程的必要性以及向数据仓库设计者提供最新概念模式的必要性使得许多方法**适用于关系模式而不是概念模式**。几乎每种方法都会考虑 ER 图或关系模式来描述数据源。最近，随着语义 Web 领域获得的相关性，一些其他工作已经提出了自动化 XML 模式或 OWL 本体的过程。

关于**需求**，他们的表示方法则有很大不同。一开始，提出了诸如表单、表格、工作表或矩阵之类的临时表示，但最近，**许多方法建议使用诸如 UML 图或 i* 之类的框架来形式化需求表示**。此外，一些工作还提出通过 SQL 或 MDX 查询将需求抽象级别降低到逻辑级别，这开辟了自动化的新可能性。

最后，我们还可以发现验证生成的多维模式的趋势以及提供支持该方法的工具的重要性。

关于**如何识别事实数据**，大多数方法都遵循了一些特定的趋势。从数据源来看，数值概念很可能适用于度量的角色，而包含数值属性的概念或那些一旦实施后具有高表基数的概念很可能适用于事实的角色。早期方法主要是需求驱动的，但后来，他们中的大多数使用这些启发式方法来**识别数据驱动阶段中的事实概念**。**但是，这些启发式方法不识别事实或度量，而是识别可能扮演该角色的概念。**因此，必须考虑需求，并且在过去几年中，**需求再次获得关注以识别这些概念**。此外，随着时间的推移，事实之间的关系也变得相关，因为它们在考虑多维代数时开辟了新的分析视角。

最后，尽管 Kimball **从一开始就引入了无事实事实的概念**，但传统上它一直被忽视。最近，一些方法论再次考虑了它们。原因之一可能是难以自动识别没有度量的事实。

维度概念传统上是通过函数依赖来识别的。从一开始，就提出了一些方法来自动识别聚合层次结构。事实上，许多方法使用需求来识别事实数据，然后他们分析数据源寻找函数依赖来识别维度数据，也许正因如此，**使用需求来识别维度概念与使用需求识别事实数据相比，似乎没有被非常重视。**

**关于维度概念的另一个明显趋势是**，一般来说，**方法的自动化程度越高，它就越以事实为中心。**关于维度概念之间的关系，维度间关系（如事实之间的关系）在考虑多维代数时开辟了新的分析视角。然而，在这种情况下，传统上它们比事实之间的关系更容易被忽视。相反，维度内的关系从一开始就获得了相关性。大多数方法都同意区分维度、级别和描述符与分析目的相关。

在本文中，我们深入了解了多维设计方法。本文调查了根据三个因素选出的 17 个工作：引用次数高的参考论文，具有新颖性贡献的论文，如果是同一作者的论文，我们将收录他们作品的最新版本。由于我们仍然缺乏标准的多维术语，并且在方法论中使用的术语来描述多维概念可能会有所不同，因此我们引入了通用的多维符号以避免误解并促进将所调查的方法映射到通用框架以比较每种方法。我们还引入了一套标准来为讨论和检测趋势奠定基础，例如共同特征或沿途假设的演变。这些标准是在对本文所调查方法的增量分析中定义的。对于每种方法，我们捕获了映射到不同标准的主要特征。如果方法论引入了新标准，则对其余工作进行分析，以了解他们对该标准的假设。因此，提出的标准是在多维设计方法分析过程中的迭代过程中定义的。我们将这些标准总结为三个主要类别：一般方面、维度数据和事实数据。一般方面是指关于方法中做出的一般假设的那些标准，维度和事实数据标准是指如何识别维度数据和事实数据并将其映射到多维概念上。我们提供了一个全面的框架，以更好地了解该领域的现状及其演变。