Technical Report of Knowledge Grid Research Center, KGRC-2007-06, August, 2007
知识网格研究组技术报告 KGRC-2007-06，2007 年 8 月；www.knowledgegrid.net/TR

互联网资源空间模型

诸葛海

物以类聚，人以群分。
互联网资源应以类来管理并提供有效服务。

人类生活空间与分类

人类生活在一个具有资源多样性的空间里。为了有效管理形形色色的资源，人们使用了多种分类方法并发明了各种分类工具。

例如，我们去超市购物时可以观察到商品是分类摆放的。这种超市管理员和顾客所共同熟悉的分类提高了顾客的购物效率也提高了商品的管理效率。走进中药店，你就会看到药品被分门别类地存放在抽屉中，每个抽屉上还贴着标签用来区分其中的药品，而且一个抽屉只放一种药物，标签之间看上去并无一定的顺序关系。人们还常常使用抽屉来管理个人资料。抽屉、书架、档案袋、货架等都是分类管理物品的工具。我们还可以观察到学生被分为班级，幼儿园用分类的方法教孩子们学习基本概念。我们还可以举出许多例子来说明分类是人类有效管理各类资源的方法也是认识现实世界和综合经验的基本方法。

资源空间模型是对文件系统和数据库系统的发展

在计算机世界，文件系统是计算机资源管理的一个里程碑。它可以看作是一个以文件类型为维的一维资源空间。 它是后来操作系统和数据库系统实现的重要基础。

数据库系统是计算机资源管理的另一个里程碑[2]。特别是关系数据模型，它以坚实的数学基础和优美的模型成为集中式数据管理的典范。

互联网的计算环境与三十多年前发明关系数据模型时的计算环境相比已发生了很大变化。原来集中稳定的计算环境发展为分散而动态的计算环境，处理的对象已不再是单纯的数据而是多样异构的资源，应用范围也由原来单纯的数据管理发展为多样的资源管理和智能服务，用户和数据拥有者也呈现由原来的以机构为中心发展为以个人为中心的趋势。这些变化对数据模型的要求已超出了经典数据模型的适用范围。

互联网资源空间模型是一种能够管理互联网资源的语义数据模型。

资源空间模型的基本概念

资源空间模型是一个通过对资源内容进行分类的规范、存贮、管理和定位网络资源的语义数据模型。*n*维空间代表了对一个资源集合的*n*种分类方法。在每一维上给定一个坐标就可定位一个点 —— 一组同类内容的资源。资源空间只关注内容，因而一个点中的资源可以是任何形式。

资源空间可以通过在维上设置约束来实现规范化，从而来增加资源管理的正确性。资源空间的范式就是用来实现这种规范化的。第一范式要求坐标名不重复，第二范式要求各坐标相互独立，第三范式要求各维（轴）互相正交（即，互相细分）。我们还可以根据应用需要定义更严格或更宽松的范式。

资源空间模型的内在特性决定了它非常值得研究，因为它不是一个普通的距离空间，它的维是离散的，而且每个坐标可以是树型结构，每个点是一个资源集合，可以是一个链接也可以是一个资源空间。在某些情况下，我们不需要给出所有维上的坐标就可确定一个点。

图 1 和图 2 分别是两种支持用户操作的资源空间可视化界面。 用户可以旋转、切分、合并资源空间，选择资源空间中所关心的点。只要选定点，所有同类资源都可以一次性获得，无论它们是何种形式。资源空间模型还允许应用系统运用类似 SQL 的语言来操作资源空间。

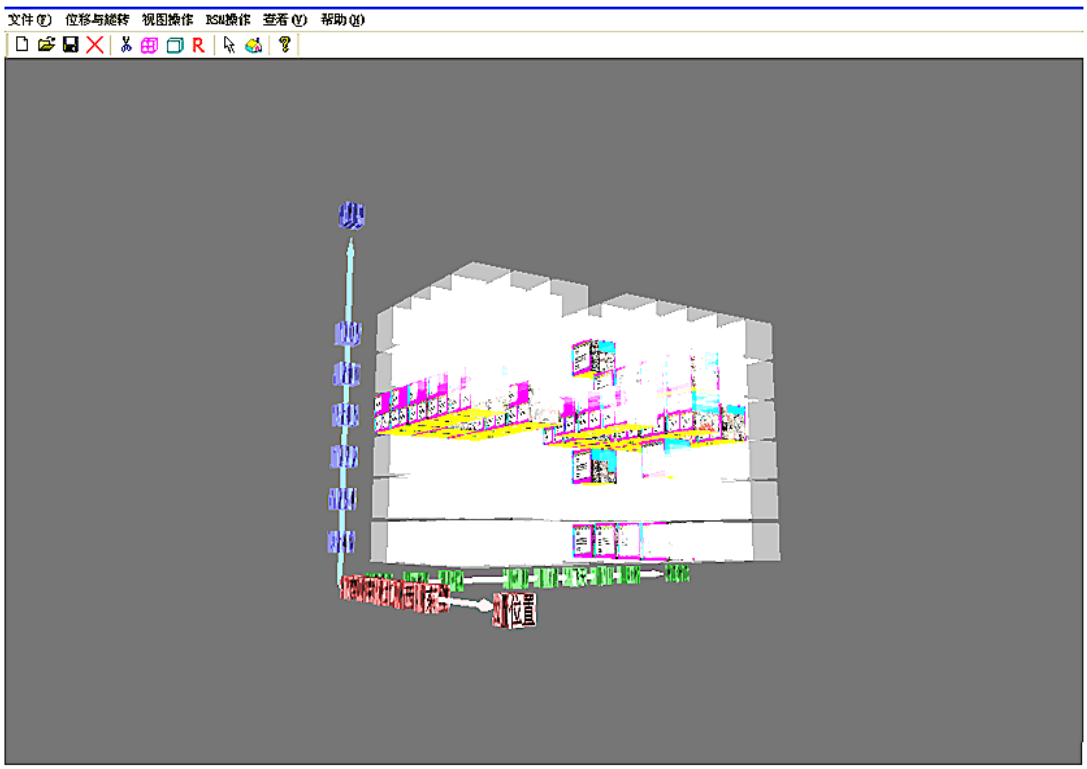

图 1。一个用来管理敦煌文化内容的可视化三维资源空间模型，其中每个小立方体代表资源空间中的一个点，每个点代表属同类内容的各种类型的资源（文字、书画、壁画、彩塑、音乐等）。

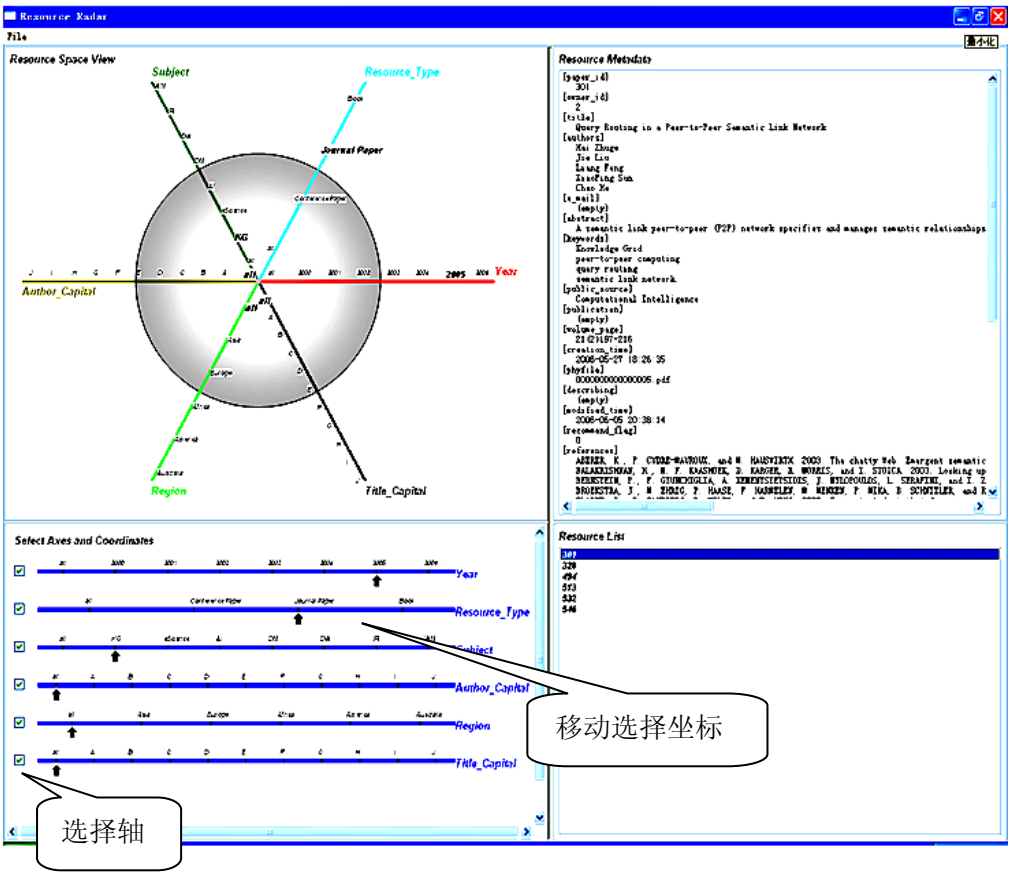

图 2。一个 *n* 维资源空间可视化界面。

与资源空间的规范管理相应的是语义链网络，它的任意一个结点可以连接到任意一个语义相关的结点。体现了互联网应用的自主性。

语义链网络是对超链网络的自然扩展 [10]，它在超链上附加一个语义因子来反映语义关系。在一组链接规则的支持下，语义链网络支持关系推理。相比之下，超链网不具备关系推理能力。

规范化和自主性是实现互联网语义数据模型的两个重要需求。有机结合资源空间和语义链网络可构成兼有规范化和自主性的语义层为高层智能服务提供一种共享基础[5]。

资源空间模型的主要内容

资源空间模型包括以下主要内容：

1. **资源空间模型方法学**。它指导资源空间模型的学习、设计和研究。包括资源空间的基本定义和特征、操作的定义、范式理论、完整性理论、查询语言和开发方法。这套理论和模型从形式上看是与关系数据库理论并行的。两者模型的不同决定了他们的范式理论的不同，进而决定了他们的完整性理论的不同和开发方法的不同。
2. **资源空间模型和语义链网络的集成及其关系理论**。资源空间模型和语义链网络虽然是独立发展出来的两个模型，但二者之间具有内在联系，在某种条件

下可以互相转换。有机结合两种模型形成一个支持资源分类管理和关系查询的语义数据模型。在这种模型下，一个资源既属于一个点又可与属于其它点或其它空间的资源建立语义关系。资源不仅可以按内容来定位，而且可以找到相关的资源。

3. **资源空间模型查询操作的完备性和必要性理论**。资源空间模型需要一套资源操作语言来进行资源的查询、更新和管理。该理论首先提供一个理论基础，能够用来判断所提出来的任何一个资源操作子语言的选择能力的完备性，并且判别结果要独立于任何这个子语言所嵌入的主语言。例如，以下问题首先必须回答：所定义的操作是否足够，即是完备的？是否必需？其次，对于各种子语言，哪种子语言的表达能力更强？通过研究得出结论：操作 Union, Difference, Intersection, Extended Cartesian Product, Selection, Join, Disjoin, Merge 和 Split 是资源空间上的一组完备的操作集合；操作 Union, Difference, Extended Cartesian Product, Selection 和 Disjoin 是资源空间上的一组完备且必需的操作集合。
4. **资源空间模型查询操作的代数和演算理论**。它从资源空间代数和资源空间演算的视角来探究资源空间模型的查询能力和表达能力。代数由一个操作数集合以及定义在这些操作数上的一组满足封闭性的操作构成。演算则是定义在数据模型上的一阶谓词逻辑，它可以描述用户需要的查询结果。演算可以用来描述用户的需求，而代数则是用来计算查询的结果。资源空间模型的代数包括了资源空间模型的 5 个完备的基本操作。我们证明了资源空间模型的代数和演算具有相同的查询能力和表达能力，即给定代数中的一个操作可以用演算将该操作的结果表达出来，并且给定一个用演算表达的查询结果，可以使用代数中一系列的操作将该结果计算出来。该理论还说明资源空间模型至少具有关系模型的表达能力。
5. **资源空间搜索的复杂性理论**。它揭示搜索效率和坐标分布的关系以及搜索效率与维的关系。在资源空间中查找一个点（基于关键字的比较），从查找效率的角度来看，空间的维数是越高越好，还是越低越好，或者是其它的情况呢？每个轴上坐标个数的分布是越平均越好，还是越不平均越好？查找的复杂度和每个轴上坐标个数的分布有没有关系？通过考查了查找复杂度和每个轴上坐标分布的关系得到结论：从查找复杂度的角度来看，每个轴上坐标的分布越平均越好。通过研究查找复杂度和空间的维数的变化的关系得出结论：资源空间的维数不是越低越好，也不是越高越好，存在一个唯一的临界维数。从查找复杂度的角度来看，具有临界维数的资源空间是最优的。我们还得出这个临界维数的取值大约是 lnN（N 是资源空间中的点的总个数）。得到的理论结果有助于对资源空间查找复杂度的计算和理解，同时也能够用于资源空间结构的设计与分析。
6. **资源空间的物理存贮机制**。资源空间的多维离散特性不同于关系数据库的一维索引和多维索引。它的独特性需要特定的存贮机制来保障查询效率。传统多维空间的存贮要求维上坐标满足线性序，并用欧氏距离度量资源的邻近度。存贮资源时把邻近的资源放在磁盘上相邻的位置，从而实现高效的资源定位。但资源空间模型中维上坐标代表分类，通常没有线性序，较为常见的是层次语义关系。为保证资源插入、删除和查找等操作的效率，我们同样希望分类语义越相似的资源在磁盘上的存储位置越紧邻。因此，我们定义同轴概念间的语义距离为它们在该轴概念树上的最短路径长度，并通过各种组合方式定

义空间中分类点间的语义距离。该距离反映了资源空间分类语义的邻近度，语义相近的资源可以被放到磁盘的相邻位置。若干语义相似的空间分类点可以用更为抽象的空间分类点表示，并作为它们的索引项存放在磁盘中。同样，对抽象后的分类点作进一步抽象并建立相应的索引项，直至形成一棵树。该自底向上生成索引树的方法类似于传统多维空间中的索引树，但优化准则更为复杂，要充分考查分类点在各维上层次语义的邻近度。由于概念的字符串值长短不一、概念间的层次语义关系可能较为复杂，而资源操作却要频繁地判断概念间层次语义关系和计算它们的语义距离，因此需要对概念进行编码。资源空间每个轴对应一棵概念树，所有轴构成一个概念森林。通过将森林转换到二叉树，对生成的二叉树的边进行编码，并将根到每个概念的路径的编码串作为该概念的编码。该编码方案不仅完全保留了概念间层次语义关系，而且支持高效的语义计算。针对编码可能过长的问题，我们还设计了无损压缩编码。

7. **基于 P2P 的分散式资源空间**。这是一种使资源空间兼有规范化和自主性的方法，目前包括结构化 P2P 资源空间和非结构化 P2P 资源空间两种解决方案。非结构化对等网络允许资源随机存放在自组织节点上，节点间的链接是任意的，它具有简单和可用性、低维护代价以及健壮性好等优点。一个 n 维的资源空间可以映射到一个分类树。对等网络上的节点根据自身兴趣和资源类型分成不同的社区，这些社区相应于分类树的叶子节点。每个节点以分层结构化的形式维护邻居信息，而分层的数目取决于该节点所在分类树的深度。当一个节点发送查询时，首先决定选择哪层邻居节点来传递查询。当查询抵达邻居节点所在的社区时，一个基于 Gossip 的算法就发布该查询信息。收到查询信息并且能回答该查询的节点直接把结果反馈给初始节点。非结构化对等资源空间同时拥有资源空间模型和非结构化对等网络的优点，可提高对等网络的性能，是资源空间模型的一种分布式应用方法。在结构化的 P2P 解决方案中，资源空间被划分成若干单元，每个单元代表了资源空间中的一个多维矩形。P2P 网络中的每个节点均负责一个单元，并且存储着位于该单元内的资源信息。每个节点都维持一个邻居列表。邻居所负责的单元在空间上是相邻的。节点利用它所存储的邻居信息来路由消息。路由算法采用贪心的策略，即当前节点总是把消息发送给在资源空间中距离目标最近的邻居。路由的时间复杂度与资源空间的维度和节点个数相关，为：$(n/4)(m^{1/n})$，其中 n 是资源空间的维度，m 是节点个数。结构化 P2P 中的节点以这种自组织的形式形成一个介于底层 P2P 网络和上层资源空间模型之间的覆盖层。该层中的所有节点都是平等的，不存在超级节点。这样的设置提高了整个网络的可扩展性。结构化 P2P 资源空间解决方案支持带有树结构坐标的资源空间。通过增加长链提高了整个系统的路由效率。
8. **概率资源空间模型**。它支持用户或应用系统以不确定性的方式存贮和管理资源，是一种更为普遍的资源空间模型。在最初的资源空间模型中，一个资源要么属于一个类，要么不属于一个类。但在很多应用中，人们往往不能准确地判断一个给定的资源是否属于一个类。为此，任意一个资源都在每一维上被赋予一个概率隶属函数，从而将一个资源空间映射到概率资源空间。资源空间模型的范式理论，完整性约束理论和操作也在概率资源空间模型中得到了更一般化的定义和解释。

数据空间、数据网格和资源空间模型

谷歌（Google）和 UC Berkeley 等单位的研究人员从应用（个人数据管理、科学数据管理和互联网结构查询等）的角度提出数据空间的概念，最近在数据库领域引起重视[3]。这从一个侧面印证了我们开展资源空间模型研究的选题是正确和及时的。数据空间的目的是围绕实体来管理数据，各实体的数据共存，它还强调数据的多样性、数据源不确定性和数据独立性等。目前还处于概念形成阶段。另外一项相关工作是数据网格[1]，它也是从应用（如科学数据管理）的角度提出了大规模数据存贮和元数据管理的方法。但与关系数据库相比，数据空间和数据网格明显缺乏理论和模型基础。

资源空间模型关注的是：面对一个应用领域、组织或个人的资源，如何建立合适的多维分类体系，并用规范化的分类体系来管理资源，无论资源呈何种形式，存放在何处。目前资源空间模型已具备完整的理论、模型和方法。

资源空间模型的研发

互联网资源空间是我们生活的现实资源空间的一部分。它是一种非常有潜力的有效管理各类网络资源的模型。它的目的不是代替数据库和文件系统，而是提供一种新的模型，在数据库和文件系统不擅长的某些应用中发挥独特的作用。它与语义链网络[5，7]、数据库模型[2]和语义互联网的研究成果（如互联网本体语言 OWL)的结合可望为未来互联环境提供一个强大的语义平台[8]。

资源空间模型的雏形于 2002 年被提出用来管理网络知识资源 [6]。2003-2004 年提出了其主要理论和模型 [5， 9]。 2007 年系统地发展了其理论、模型和方法[4]。

中国科学院计算技术研究所知识网格研究组正在开发资源空间模型系统，完善其理论，并在 e-Culture 和 e-Science 领域开展应用。

以下同志参加了相关研究和开发工作：姚二林、星芸鹏、李向、冯亮、何超、韩旭、时鹏、刘进、唐明董、王震、顾志宇、杨鲲、张均胜等。

研究受到国家重大基础研究计划（2003CB317000）的资助。

参考文献

[1] A. Chervenak, I.Foster, C.Kesselman, C.Salisbury and S.Tuecke, The Data Grid: Towards an Architecture for the Distributed Management and Analysis of Large Scientific Datasets, Journal of Network and Computer Applications, 23(3)(2000)187-200.

[2] E.F. Codd, A Relational Model of Data for Large Shared Data Banks. Communications of the ACM, 13 (6) (1970)377-387.

[3] A. Halevy, M. Franklin and D. Maier, Principles of Dataspace Systems, Proceedings of the twenty-fifth ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems, 2006, pp.1-9.

[4] H. Zhuge, The Web Resource Space Model, Springer, 2007.

[5] H. Zhuge, The Knowledge Grid, World Scientific Publishing Co., Singapore, 2004.
[6] H. Zhuge, A Knowledge Grid Model and Platform for Global Knowledge Sharing, Expert Systems with Applications, 22 (4) (2002) 313-320.
[7] H. Zhuge and X.Li, Peer-to-Peer in Metric Space and Semantic Space, IEEE Transactions on Knowledge and Data Engineering, 6(19) (2007) 759-771.
[8] H. Zhuge, P. Shi, Y. Xing and C. He, Transformation from OWL Description to Resource Space Model, Keynote at 1st Asian Semantic Web Conference, Beijing, China, Sept. 3-7, 2006, LNCS 4185, pp.4-23.
[9] H. Zhuge, Resource Space Grid: Model, Method and Platform, Concurrency and Computation: Practice and Experience, 16 (14) (2004) 1385-1413.
[10] H.Zhuge, Autonomous Semantic Link Networking Model for the Knowledge Grid, Concurrency and Computation: Practice and Experience, 7(19)(2007)1065-1085.

诸葛海（http://www.knowledgegrid.net/~h.zhgue）：中国科学院计算技术研究所研究员，973 语义网格项目首席科学家。