

数据空间研究进展

1 引言

2007年7月，我们开始承担863课题“海量数据空间模型、索引与查询技术研究”。本课题旨在研究海量数据空间的理论方法和实现技术，在数据空间模型、组织与分类、演化集成、查询优化等核心技术方取得进展。在此基础上开发具有自主知识产权的数据空间管理原型系统。2007至2008年，我们在数据空间模型及查询技术方面进行了研究，取得了一些成果，并开发了两个数据空间原型系统：计算机科技文献集成系统C-DBLP和个人数据空间管理原型系统OrientSpace。在这些成果的基础上，2009年针对数据空间的演化集成、查询优化等问题进行了研究，基于取得的研究成果完成了相关专利的申请，并对两个数据空间原型系统进行了升级。图1是我们在2008年提出的数据空间管理系统框架。结合这一框架，对2009数据空间项目研究进展做一简单总结。

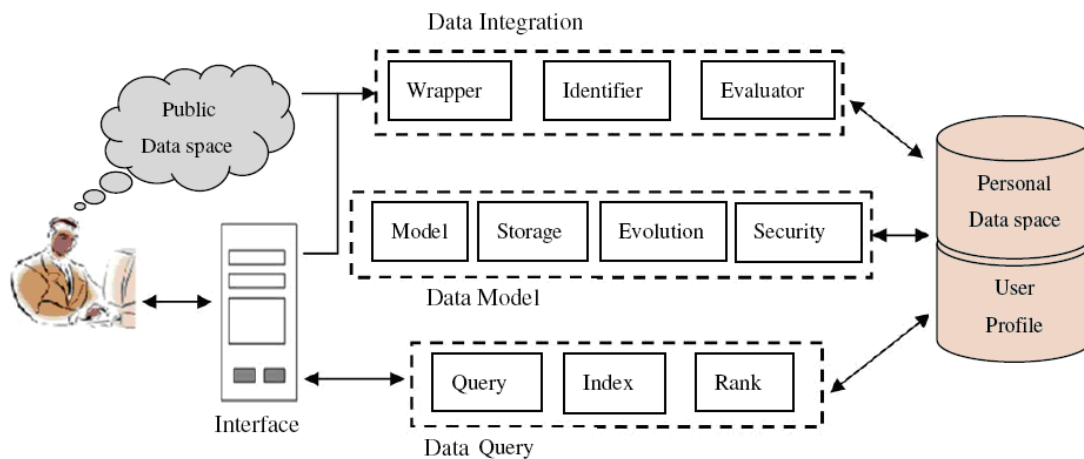


图1 基于主体的数据空间管理系统框架

2 研究进展

由图1可以看出，数据空间模型、数据集成和数据查询是数据空间研究中的三个基本问题。在已有研究成果基础上，进一步针对这三方面的问题进行了研究，取得了一些研究成果。

2.1 数据空间模型

去年我们初步提出了任务空间模型和核心数据空间的思想。2009年在新的研究成果基础上，对这两个概念进行了明确的阐述和形式化，使之更加完整。

任务空间模型

在维基百科中，任务定义为旨在完成特定目标的一系列行为的集合。在信息检索领域，人们很早就对于任务的概念、分类及特征进行了研究。随着信息技术的发展，计算机日益成为一个帮助人们解决问题、完成任务的重要工具，从而为任务管理带来新的特点。由此我们希望在相关工作基础上，从数据管理的角度对任务给出一个形式化的定义。

我们将任务定义为：用户具有明确目标的一系列数据操作的集合。静态的看，任务是一个与特定目标相关的数据集合；动态的看，任务是针对这一数据集合的操作序列。在完成任务目标的过程中，用户一方面会创造出新的数据对象，同时也会参阅许多已存在的数据文件。例如，当用户写一个项目报告时，任务目标可以看作最终的项目报告文件，在书写过程中，可能需要引用其他文档（网页、邮件等）上的一些内容（数字、图片、表格，等）。基于这一观察，我们将任务目标物化为用户生成的数据对象，并将与任务相关的数据集合分为两个子集：目标数据集和参考数据集。目标数据集是指用户在完成任务过程中生成的目标文件集合；参考数据集是指用户生成目标数据过程中参考过的数据文件。

基于目标数据集包含数据对象的数量，我们将任务区分为基本任务和复合任务。基本任务是指目标数据集中仅包含一个数据对象的任务。例如，用户写一封邮件可以看作是一个基本任务。复合任务是指涉及多个目标对象的任务。当用户写一篇论文的时候，往往会生成多个数据文件，这就构成一个复合任务。一个复合任务可以分解为多个基本任务。此外，时间是任务的一个重要属性，一个任务往往有一个生命周期。以上概念构成了任务概念模型。

在任务定义的基础上，我们进一步提出了任务空间的概念。任务空间定义为一个图，节点表示任务，边表示任务之间的关系，理论上可以定义多种任务关系，定义任务关系的目的是为了提高数据管理的效率。研究表明，内容和时间是人们基于任务查询个人数据时经常参考的两个因素。基于此提出并定义了两种任务关系：内容关系和时间关系。内容关系表示两个任务内容的相似性，可以支持基于内容的任务查询，如“查询与数据空间相关的所有任务”；时间关系表示两个任务的并行性，可以支持基于时间的任务查询，如“查询去年所完成的一些任务”、“查询准备 SIGMOD2010 论文期间所做生成的一些文件”等。任务空间为我们提供了一种用户任务描述数据的方法，基于此模型，用户可以执行基于任务的数据操作。

核心数据空间

2008 年我们提出了根据数据对象和用户的关联程度，构建核心数据空间的思想。核心数据空间由用户曾将访问过的数据对象组成，数据量往往还是比较大，用户很难通过直接浏览快速找到需要的数据对象。研究发现，用户记忆呈现出一定的规律性，这些记忆规律可以用来帮助用户提高访问效率。在信息检索、人机交互、认知行为学等领域，有许多关于用户认知行为规律的研究成果。通过这些成果应用于数据管理，并结合数据空间查询的特点，我们将核心数据空间概念从以下几个方面进行了深化：

- (1) 核心数据空间是一个用户曾经访问过的数据对象构成的数据集合。由于人们对个人数据信息的访问是一种“重访问”，因此我们提出只将用户曾经访问过的数据对

象作为核心数据空间的内容。

(2) 核心数据空间是基于用户记忆规律的多维视图。基于诸葛海研究员提出的资源空间模型，我们提出利用多维空间描述个人数据对象集合。其中每个数据轴对应一个数据属性，每个数据轴的坐标对应数据对象在该数据轴的取值。

(3) 数据轴上坐标的确定基于人的记忆规律。例如，研究表明，“随着时间推移，人们对于数据特征的记忆不断减弱”，基于此我们将最近访问时间属性区分为{今天，本周，本月，本年，一年以前}，等等。

基于以上核心数据空间的定义，我们提出了一个核心数据空间本体。其考虑了数据对象的 10 个共有属性作为数据轴，并给出了各个数据轴上坐标值的确定方法。

任务空间和核心数据空间是我们针对数据空间特点和用户查询需求，提出的两种框架模型，为进一步的研究奠定了基础。

2.2 数据空间演化集成

演化集成是数据空间管理的基本问题。传统的数据库往往是先构建数据模式，然后输入数据。在数据空间中，完全让用户手工地输入、修改数据信息是不现实的，因此必须探索一种自动的数据空间集成与演化策略，使数据空间能够在尽可能少的用户干预的情况下保持数据的高质量，以满足用户查询要求。数据空间演化集成包括很多研究问题，例如，如何从众多数据源中抽取数据；如何进行数据模式的匹配；如何实现“从数据到模式”的演化集成；如何高效地建立初始数据空间；如何通过演化保持数据质量。

Web 数据抽取

数据多样性是数据空间的主要特征之一，如何从众多数据源中抽取信息是数据空间集成面临的首要问题。目前 Web 已经成为最大的数据源，如何从 Web 数据库中高效地集成数据成为重要的研究问题。针对这一问题，我们创造性的提出了一种基于视觉的 Web 数据自动抽取方法 (ViDE)。该方法考虑了网页信息的位置特征、布局特征、外观特征和内容特征，克服了原有方法依赖 HTML 树结构的不足。实验证明了这一方法的有效性。

数据空间自动构建

针对如何高效地构建数据空间的问题，提出了一种自动构建个人数据空间的策略。该策略通过分析操作日志，挖掘用户的兴趣特征，基于该特征计算数据对象与主体的相关度，从而自动识别与主体相关的数据对象，建立初始的数据空间。问题的挑战性在于如何自动挖掘用户兴趣特征，如何形式化的表示用户兴趣特征，以及如何计算数据对象与主体兴趣特征的相关度。我们提出了基于用户行为日志的方法，并从文件类型、文件夹、关键词等几个方面考虑了用户兴趣特征。从而对于任给的一个数据文件，通过计算该文件与用户兴趣特征的相似度，就可以计算出该文件属于数据空间的概率值。

数据空间演化

数据空间本质上可以看作一个语义链网络，其中每个节点是一个数据对象，两个数据对象之间的边称之为语义链。数据演化就是指系统能够自动地更新这个语义链网络，使之真实准确的反应数据空间的状态。例如，当用户访问了新的数据对象的时候，数据空间能够自动将该数据对象加入到该语义链网络中，并自动与相关的数据对象建立语义关联。数据空间中的语义链可以根据用户的需要建立很多种，例如，具有参考关系的文件之间可以建立语义链；曾经合作过的两个人可以建立语义链；同属于一个任务的两个数据对象可以建立语义链，等等。我们提出了基于用户行为的数据空间演化策略：基于用户操作，及时补充新的数据对象和语义关联，及时更新现有的语义链信息，从而使数据空间保持高数据质量。例如，当用户访问一个新的数据对象的时候，该数据对象会自动增加到核心数据空间中，会自动与相关的数据对象建立关联。

2.3 数据空间查询

传统的数据库查询往往是通过特定的查询语言实现，例如关系数据库中的 SQL 语言、XML 数据查询语言等。而数据空间面对的往往是没有数据管理背景的普通用户，因此需要为用户提供简洁的查询接口。在不同的场景下，用户往往需要不同的查询方式。例如，当用户编辑一个文档的时候，可能需要查询“该文档所引用的文件有哪些？”；当用户写一个总结报告的时候，可能需要查询“我最近半年完成了哪些任务？”；当用户想要查询一个文件的时候，可能只记得一些模糊的信息（如最近半年访问过、类型是 JPG 或 VSD，被存放在 D 盘等）。因此需要针对这些特定的查询场景设计与之相应的查询策略。针对不同应用需求，我们提出了基于核心数据空间的导航式查询策略、基于任务的数据查询方法、以及基于上下文的数据查询方法。

基于核心数据空间的导航式查询

该方法采用导航式查询，支持用户根据记忆线索，逐步定位到所查找的数据对象。提出了一种基于用户行为规律的简易的查询逻辑，从而支持用户表达比较复杂的查询语义。实现了一个个人核心数据空间本体，并提出了基于该本体建立多维查询接口的方法。

基于用户操作上下文查询个人数据空间

用户书写一个文档的时候，往往需要调用其他文档的一些数据信息，如数字、图片、表格、参考文献等。当用户重新编辑该文档的时候，经常需要访问该文档曾经引用的一些文档。对于这一用户需求，目前并没有好的解决方法。基于此我们定义了一种数据关系：上下文关联关系（Context-based Relation），并提出了一种自动识别这种数据关系的方法，进一步提出了基于这一关系查询个人数据空间的策略（C-Query）。

基于子图匹配的数据空间搜索技术

传统数据库中数据关联基于表，而数据空间的数据关联是元组一级的，因此要复杂的多。数据空间本质上可以看作一个复杂的语义网络，也可以抽象为一个复杂的图结构。研究证明，一个图中的子图匹配查询是 NP 完全问题。针对这一问题，我们提出了一个双层索引的思想，对大图上每个点的两种特征进行索引：点的身份信息 and 位置信息。在进行查询处理时，首先通过第一层索引（点的身份信息）找到符合要求的点，然后再通过第二层索引（点的位置信息）进一步排除位置上不满足要求的点。实验证明我们提出方法是有效的。

2.4 数据空间管理原型系统

基于在数据空间模型、集成与查询方面的研究成果，我们对原有的两个数据空间原型系统进行了升级：

在 OrientSpace 系统中，增加了以下功能：

- (1) 对于核心数据空间查询的支持
- (2) 基于任务的数据查询功能
- (3) 基于用户日志自动构建初始数据空间
- (4) 基于 RDF 的数据存储系统
- (5) 个人数据文件的自动标识
- (6) 数据之间关联的挖掘等。

在 C-DBLP 系统中，增加了如下功能：

- (1) 同名区分功能；
- (2) 文献 BibTex 信息展示功能；
- (3) 作者的相关图片的自动收集与展示；
- (4) 合作关系的可视化展示，利用直观友好的图形化界面为用户提供更丰富的信息；
- (5) 论文数量趋势变化的图形化显示，帮助用户更直观地了解作者论文发表情况。

3 结论

在 2008 年研究成果基础上，进一步开展了深入细致的研究工作，特别加强了在数据空间模型、演化集成、查询优化方面的研究以及在原型系统研发方面的工作，取得了一些研究成果，并将取得的研究成果应用于已经开发的数据空间管理原型系统，验证了我们所提出的方法的可用性。这些成果为我们下一步的工作打下了坚实的基础。