

# OrientSpace: 个人数据空间管理原型系统

张相於 李玉坤 寇玉波 赵婧

## 1 简介

随着个人数据量的不断增长以及数据类型的不断丰富，个人数据管理所面临的问题日益凸显。一方面，人们对数据管理的要求越来越高，希望数据管理工具能够提供更多有效的服务，如能力更强大的数据查询、个人任务的管理、facet search 等；而另一方面，现有的数据管理工具不能提供令人满意的服务。以 Windows 资源管理器以及桌面搜索引擎为代表的个人数据管理工具目前仍然只能提供非常有限的功能，如关键字搜索，基于目录结构的数据组织和管理。这种种局限性促使我们去构建一种新的个人数据管理工具，来应对人们所面临的各种挑战。

基于这样的考虑，我们设计实现了个人数据空间管理原型系统——OrientSpace。OrientSpace 是一个以数据空间为理论依托的新一代个人数据管理工具，其目标是以数据空间的思想为基础，为个人用户提供更丰富有效的服务，以解决个人数据管理中存在的问题。OrientSpace 也是我们数据空间研究课题的工作展示平台，在展示我们研究成果的同时，也能够开发过程中不断发现新的研究问题。与现有的个人数据管理系统相比，OrientSpace 以用户为核心，具有以下的特点：

- **充分考虑数据之间的关联。**

在数据项之间构建丰富的关联信息，充分将这些关联信息用于查询、浏览等各项功能，为用户提供更好的服务。

- **提供基于任务的数据组织方式。**

经过我们的观察和实验，任务是人们在日常数据管理中经常用来组织数据的一种重要形式。以任务为依据来组织数据能够提供更为高质量的服务，极大地方便用户的使用。

- **以用户为核心的 Pay-As-You-Go 演化。**

个人数据管理系统的核心，只有充分考虑用户在系统中的作用才能提供高质量的数据管理服务，一个系统的好坏要以它是否能够很好地服务于某个具体用户来评价。因此我们使用了一套以用户为核心的演化机制，通过分析用户行为对系统进行不断地演化，使系统能够随着用户使用的增加而提供越来越好的服务。

图 1 所示为 OrientSpace 的系统架构图。系统包括 8 个模块，有两个前端模块：用户界面模块和用户反馈管理模块，以及六个后端模块：数据导入模块，存储管理模块，关联管理模块，演化管理模块，查询处理模块以及任务管理模块。前端模块负责将服务呈现给用户以及收集用户的反馈并将其返回到系统中；后端模块负责服务的生成和维护。

## 2 主要功能

目前 OrientSpace 系统主要包括以下几个功能：

- **关键字查询**

用户可以向系统提交关键字查询，系统会将关键字查询分别提交到文本索引和 RDF 元信息存储上，返回满足查询要求的数据。在文本索引方面我们没有使用传统的全文索引，而是采取了选择性索引的策略，该策略会在后面进行具体介绍。

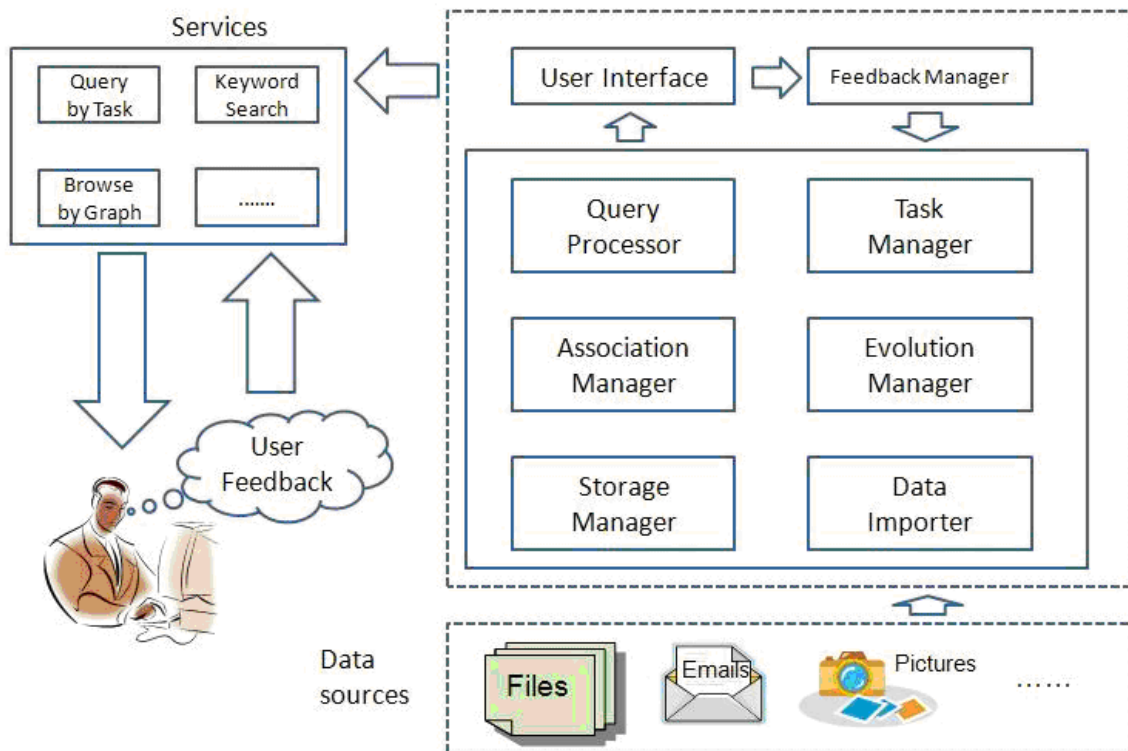


图 1 OrientSpace 系统架构图

## ● 基于内容分析的关联建立

由于具有高度的异质性，个人数据空间中的数据不能像结构化数据那样提供丰富的操作。这其中一个重要的原因就是个人数据空间中缺乏丰富的数据项之间的关联信息，数据之间不能形成有机的整体，从而严重制约着服务的质量。因此数据空间中一个重要的问题就是建立数据项之间有效的关联。在 OrientSpace 中我们使用了一套基于内容分析的关联建立方法，这套方法通过对数据项内容的分析，在数据项之间建立不同类型的关联，这些关联信息将被用来支持包括基于图的数据浏览在内的各种服务。目前我们已经实现了 5 种关联的建立策略，而且可以很容易加入新的关联类型。图 2 显示的是数据空间中丰富的关联信息。

## ● 基于图的数据浏览和查询

用户常常会忘记要寻找的文档的具体内容，甚至有时会连一个合适的搜索关键词都记不起来，这时最好的办法就是从一个相关的文档入手，通过文档之间的关联进行逐步探索，最终找到所需文档。OrientSpace 提供的基于图的数据浏览和查询就是为了给用户这样的便利，用户可以从一个文档入手，通过查看与之相关的文档进而逐步逼近要找的文档。我们目前提供了 5 种关联类型，并将在以后的改进中逐步加入更丰富的关联类型，使这个功能变得更加强大。下图是系统中所有数据构成的一个全局图，其中每个点代表一个数据项（文档），每条边包含一个或多个关联信息，因为两个文档之间可能存在多种关系。用户在使用基于图的数据浏览时可以指定关联的类型和关联的层数，也可以从一个关键字搜索的结果作为起点开始图的浏览。

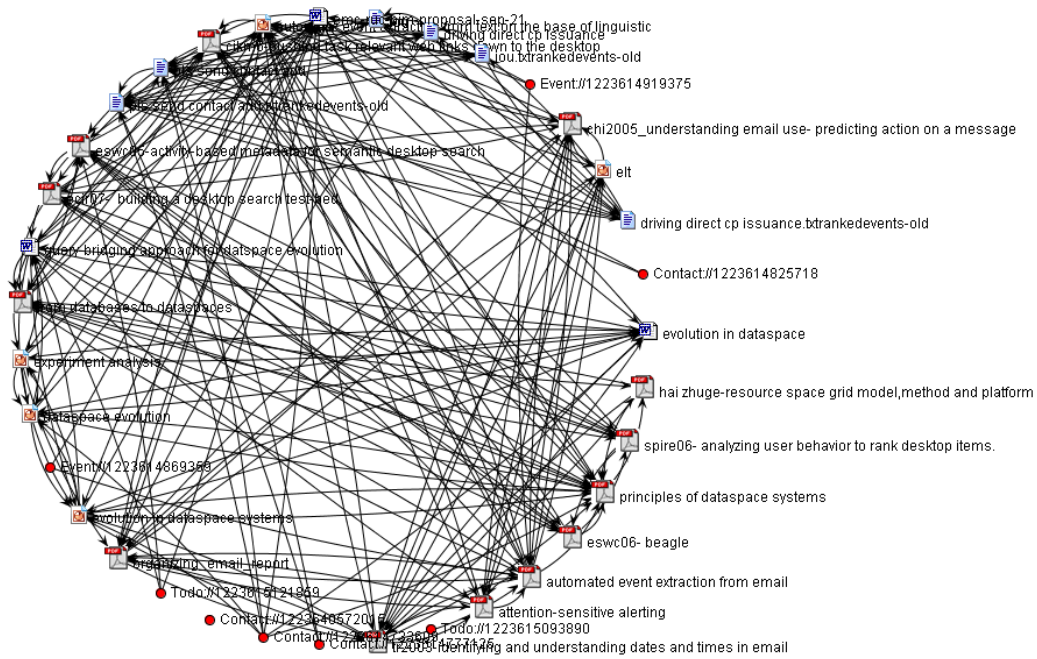


图 2 OrientSpace 中的数据以及数据间的关联

下面的例子展示了从关键字搜索出发进行基于图的数据查询：用户想要查找一篇关于邮件处理方面的论文，但是却无法记起论文的题目，只记得和曾经写过的一个有关 PIM 的 proposal 有关，尝试了一些关键字搜索也不能找到这篇论文。这是我们在日常信息处理中经常会遇到的一个问题，在 OrientSpace 中，这个问题可以很好地被解决。首先，用户用 PIM 和 proposal 为关键字进行关键字查询，会得到一个结果列表，如下图所示。

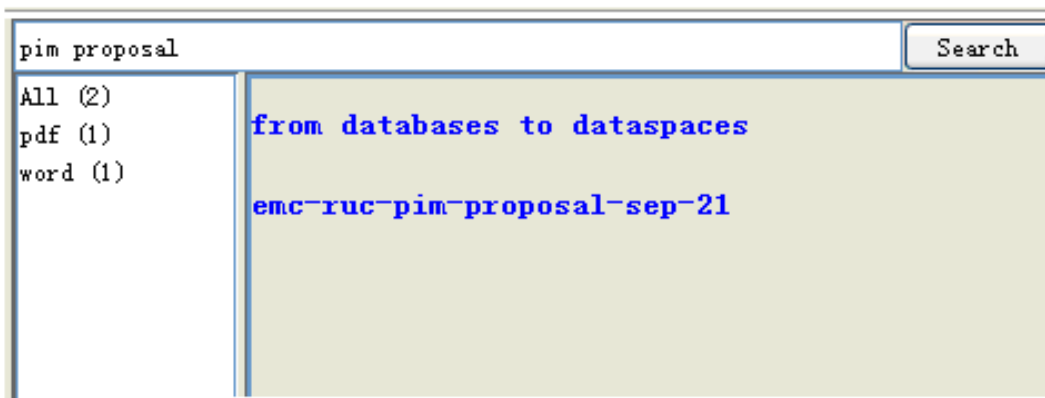


图 3 关键字搜索

结果列表中名为 `emc-ruc-pim-proposal-sep-21` 的文件就是用户能够回忆起来的 `proposal`，随后在这条记录的右键菜单中选择展示关联图，就得到了与该文件相关联的所有其他文件。图 4 显示了上述文件的关联图。在图 4 所示的关联图中，用户可能会通过各种关联找到自己要找的数据项，如果在这个关联图中找不到，还可以从图中的某一个其他节点将图继续展开，直到找到最后需要的数据项。这种从关键字查询出发，结合图结构的数据查询方式充分利用了用户的思维习惯，提供了利用关联信息进行查询的方法，能够在无法一次性准确定位数据时实现高效的查询。

## ● 模式的自由创建和管理

在 OrientSpace 中我们允许用户根据自己的意愿自由建立、修改和删除模式，并可以在各个模式

下自由地添加实例。这里体现的是数据空间所提倡的“从数据到模式”的思想：用户一开始可能无法得知某种数据的具体模式，因而只能建立一个初步的模式，但是已经可以向这个不完整的模式里面添加数据；随着用户对模式的不断了解，系统中的模式可能会被不断修改，逐渐趋近最终完整的模式。由于我们在底层采用 RDF 作为存储支持，OrientSpace 允许在模式变化的整个过程中随时修改数据。这种方式更适合目前形势下的数据管理技术发展趋势。

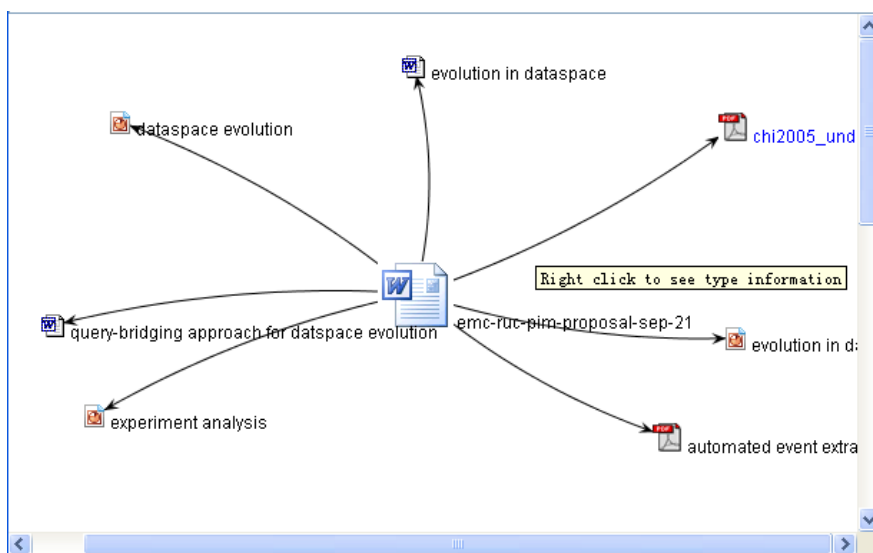


图 4 数据项的关联图

### ● 资源管理器

为了方便用户的操作和使用，我们还提供了传统的资源管理器视图，允许用户用熟悉的方式管理他们的数据。但是我们的资源管理器和传统的并不完全相同，我们还提供更多具有数据空间特色的功能。

## 3 系统特色

如上所述，OrientSpace 与其他数据管理工具相比，具有十分鲜明的特色，具体如下：

### ● 选择性的内容索引

在对文本内容进行索引时，OrientSpace 没有采用传统的全文索引的方式，而是用选择性索引的方式进行索引。采取这种方式基于以下原因：一是全文索引占用空间过大。其结果一方面造成空间的浪费；另一方面建立和维护全文索引的时间开销会很大，从而降低系统性能。二是个人数据管理与 Web 搜索是不同的。在网络搜索中，用户不知道自己所要寻找的东西是否存在，任何符合用户输入条件的都可能是而让用户满意的结果。这就要求搜索引擎必须索引文档中尽可能多的内容。而在个人桌面应用中，用户大多数情况下是知道自己要找的资源是什么的，未知的常常只是这些资源的存放位置。那么用户用来查询这些资源的关键字一般来说也是这些资源中比较“重要”的词汇，因此将这些比较重要的关键词汇索引起来就可以回答用户绝大多数的查询。

## ● 用户对数据模式的自由控制

如上面所介绍，OrientSpace 允许用户灵活自由地创建和修改数据模式。用户可以在使用过程中不断修改创建的模式，使其与现实世界中的数据不断接近，充分体现了数据空间“先有数据，后有模式”的特点，而且也适用于个人数据空间中无结构数据数量大的特点。

## ● 基于内容的关联建立

一些其他的数据管理工具中也提供了数据间的关联信息，但是它们多是通过分析一些简单的结构信息来得到关联信息，如文件存放的目录结构等。在 OrientSpace 中，关联信息是通过分析数据项内容的分析建立起来的，这样的好处是能够建立更加丰富的关联信息，因为元信息能够提供的信息是非常有限的，只有通过内容的分析才能建立足够丰富的关联信息。

## ● 基于图的数据组织方式

个人数据管理所呈现出的很多问题都来源于目前文件系统数据组织结构的限制。单一的树形组织结构已经不能满足用户越来越丰富的应用需求，而数据间的关联信息正在起到越来越重要的作用。在这种背景下，用图的结构来组织数据显得更加合理，也更能够满足用户的各种需求。在 OrientSpace 中，我们通过分析文档的内容和结构信息来生成关联信息，进而利用关联信息构造系统中数据的图结构。

## ● Pay-As-You-Go 式的系统演化

作为数据空间系统的重要特性之一，系统演化是衡量一个数据空间系统能力的重要标准。我们在系统中采用一种基于分析用户行为的方法在个人数据空间的资源之间建立关联信息，并借助用户反馈对关联信息进行不断更新和筛选，以此推动系统不断演化。该方法充分考虑了用户的个性化信息，使得系统不断向着“更好地服务这个用户”的方向发展，而不是泛泛地“变好”，因为个人数据空间具有很强的个人特征，只有更好地服务于某个具体的用户才是有意义的。

## 4 总结

OrientSpace 是一个以数据空间为理论基础的个人数据管理工具，目的是用数据空间的思想解决个人数据管理中日益凸显的各种问题和挑战。该系统特色鲜明，能够完成一些其他工具无法完成的任务。目前系统已完成第一阶段的开发，在第二阶段的开发中，我们将对更多的数据类型进行支持，并加入更多功能。