数据基因:数据的进化过程管理模型*)

奚建清　郭玉彬　汤德佑
(华南理工大学计算机科学与工程学院　广州 510640)

摘　要　同生物一样,数据也具有进化过程。本文参照生物进化论的观点,给出一种数据进化过程管理模型——数据基因模型。在数据基因模型中,数据进化过程中产生的信息称为数据的遗传信息,它们被保存在数据基因中。通过数据基因的遗传和变异来记录数据的特征及与其它数据的关系。该模型可用于企业信息处理、审计事务处理等对数据来源、去向需要清楚了解的应用中。本文给出了数据基因模型的基本概念、性质,并给出了模型的一个实例。

关键词　数据基因,数据基因序列,数据基因组,直接传播树,家族树

Data Gene: Model of Managing the Evolution of Data

XI Jian-Qing　GUO Yu-Bin　TANG De-You
(College of Computer Science and Engineering, South China University of Technology, Guangzhou 510640)

Abstract　Like life in biology, data is in the journey of evolution. According to life evolution theory, a model to manage the evolving information of data, Data Gene Model, is presented in this thesis. In this model all evolving information of data are named as genetic information of data, and kept in data gene. And data gene will transmit or mutate with the varying of data. So that most information about data like origin, creator, travel experience, relation with other data and so on can be obtained from its data gene. This model can be used in applications that need much information about evolution of data, for example enterprise information modeling, audit system. Several aspects of this model have been defined in this thesis, including the basic concepts such as data gene, data gene sequence and data genome, relationships of data genes and data genomes. And an instance of this model is also given.

Keywords　Data gene, Data gene sequence, Data genome, Direct propagation tree, Family tree

1　引言

生物进化过程是指生物从无到有,从简单到复杂,从低等到高等,一批又一批地"踏上"地球,又"远离"地球走向灭亡,进行着的自然界的"新陈代谢"的过程。相应的数据进化过程指数据在计算机上产生、传播、演化的过程。数据在进化过程中会产生大量关于进化的信息,如一次数据交换中,数据的发布者、发布时间、发布地点等信息都是关于数据进化的信息。数据进化过程中产生的信息对诸如信息监控、审计等应用系统具有重要作用,利用这些信息可追踪数据之间的联系、跟踪数据去向、分析数据的演变。然而目前数据管理研究集中在提高数据本身的利用率,如各种数据库技术、数据仓库技术,而很少对数据的进化过程进行管理。

生物进化过程中表示物种特征及与其它物种关系的信息称为遗传信息。遗传信息被记录在一系列的基因中,随着生命的进化而不断地遗传和变异。参照生物学中基因的概念,我们将数据在进化过程中产生的信息称为数据的遗传信息,并将这些信息保持在一系列的"数据基因"中[1]。数据基因贯穿整个数据的进化历程,随着数据的产生而产生,并随着数据的进化进行遗传和变异。文[1]给出了数据基因的概念及应用领域。在此基础上,本文给出了完整的数据基因模型,包括数据基因、基因序列、数据基因组等基本概念、数据基因、数据基因组之间的关系。分析了数据基因模型的性质,并作为数据基因模型的实例给出了文件数据基因组的具体结构。利用数据基因模型,用户可方便地求解数据进化过程中的动态变化,如数据来源和去向、数据的传播历程、相关数据之间的关联等。

本文第2部分介绍相关工作。第3部分给出数据基因模型的定义、关系、模型性质。第4部分给出数据基因模型的一个实例一文件数据基因组的具体结构形式。最后给出全文总结并指出下一步工作。

2　相关工作

数据在其进化历程中会产生大量的相关信息,对这些信息管理,最常用的数据模型是元数据模型。目前最有影响的元数据模型是 DCMES (Dublin Core Metadata Element Set)[2],它对信息资源的描述由 Title, Identifier, DataType, Subject, Description, Creator, Publisher, Contributor, Dissemination 等元素项组成。它所描述的信息资源是任何可被标识的信息,能够比较全面的涵盖信息资源的静态特征。然而元数据模型只反映了数据的当前状态及特征,没有对数据从产生到消亡的过程及与其它数据的关联进行建模。

在时态数据库中通过引入"时态"的概念实现了对数据演变的建模[3],描述了结构化数据的演化过程。TXPath 数据模

*)广东省科技攻关计划项目(G03B2040770);广东省自然科学基金项目(B6480598);湖南省自然科学基金(05JJ30122)。奚建清　博士,教授,博士生导师,主要研究领域为数据库与网络计算;郭玉彬　博士生,讲师;汤德佑　博士生,讲师。

型[4~6]在 XPath 基础上增加了"时态"的概念，完成了对半结构化数据演变经历的建模。GEM[7~9]则是基于有向图的时态数据模型，它允许对半结构数据本身及其时态属性进行统一方式的描述。这些数据模型都从时间的角度对数据的历史进行了建模，是对数据的历史信息进行利用的有益尝试。

然而，时态数据模型仅从数据的变更时间上对数据进行建模，对查询如数据与其它数据源之间的关联等动态信息无能为力。此外，文[10,11]给出的数据树模型是一种用于异构数据源集成的通用数据模型，文[12]中给出的是一种根据不同数据源条件和数据源能力进行优化数据查询的数据存储模型。这些数据模型都是对通用数据模型的研究，没有对数据的进化过程进行研究。

Adoc[13]是一种解决协同商业过程集成的应用编程模型。这种模型注重不同商业过程之间的关系，以一种"动态数据上下文(dynamic data context)"的方式表示数据之间的联系。以该文为基础，"smart distance"[14]表达了动态和异构的信息系统之间的关系；ABO[15]则描述了商业实体之间生命周期。然而这些方法侧重于商业过程之间的关系，而没有从数据角度对数据的进化过程进行研究。

相对而言，数据基因模型侧重于对数据进化过程的管理。一份数据的数据基因中保存其创建、加工、使用经历及与其它数据之间的联系。利用数据基因组可以跟踪数据的整个生命历程及血缘关系。

3 数据基因模型

生物学中，生物的基因结构分成基因、基因序列、基因组三个层次。基因组可以随生命体的进化遗传和变异。与之类似，我们将数据的遗传信息记录在数据基因、基因序列、数据基因组三个层次上。其中数据基因是基本的遗传、变异单位，基因序列是数据基因的组织形式，而数据基因组是数据遗传信息的完整表示。它一方面表达数据的特性，另一方面也描述不同数据之间的关联。随着数据从"旧"数据向"新"数据的进化，数据基因组通过遗传变异操作产生新数据的数据基因组。下面先给出数据基因模型的基本概念，然后给出数据基因、数据基因组的关系，并给出数据基因模型的性质。

3.1 基本概念

假设 A 为描述进化信息的属性，V 为属性值，则

定义 1 基因片段(Data Gene Segment)是一个二元组，$f=(A,V)$。数据基因(Data Gene)是 F 的正闭包，即 $D=F^+$，其中 F 为所有基因片段的集合。

一个基因片段是一个"属性/值"对，是数据基因模型中有意义数据的最小单位。数据基因由一到多个基因片段组成，是数据遗传和变异的最小单位。一个数据基因可以是对数据进化过程中一次经历的记录，也可以是数据性质的描述。同生命体基因一样，数据基因也可以分成显性基因和隐性基因两种类型。显性基因是指对数据内容和性质有所影响的数据基因，如记录编辑操作的数据基因。而隐性基因则是对数据内容和性质没有影响的数据基因，如记录转发操作的数据基因。

定义 2 描述数据整体信息的数据基因，称为主导基因，记为 D_a。

主导基因在产生数据时产生，全面地描述数据的静态属性，其中属性包括数据 ID、数据名、数据来源、创建者、数据类型、权限等。其结构可参考 DCMES[2]中对信息资源的定义。

定义 3 基因序列是一个二元组 $s=\langle \$sid, D^+ \rangle$，其中 $\$sid$ 是基因序列的标识符，D 为所有数据基因的集合，D^+ 则是对数据的某种性质或某方面经历进行描述的多个不同数据基因的集合。

基因序列是数据进化过程中记录某方面遗传信息的一条"线"。由于对数据的属性描述可以从多个角度进行描述，因此数据的遗传信息通常保存在多个基因序列中。

定义 4 描述数据整体信息及演化过程的数据基因形成的序列称为数据的主基因序列，记作 S_m。

定义 4 说明一个数据的主导基因属于其主基因序列。主基因序列中其它数据基因则是对数据整体信息进化历史的记录。如对数据类型、数据名的改变作为数据基因存在于主基因序列中。

定义 5 数据基因组是一个三元组：$dg=\langle \$gid, S^+, DG'^* \rangle$，其中 $\$gid$ 是数据基因组的唯一标识符，S 是所有基因序列的集合，S^+ 则是基因序列的正闭包，DG'^* 是其组成数据基因组的闭包。

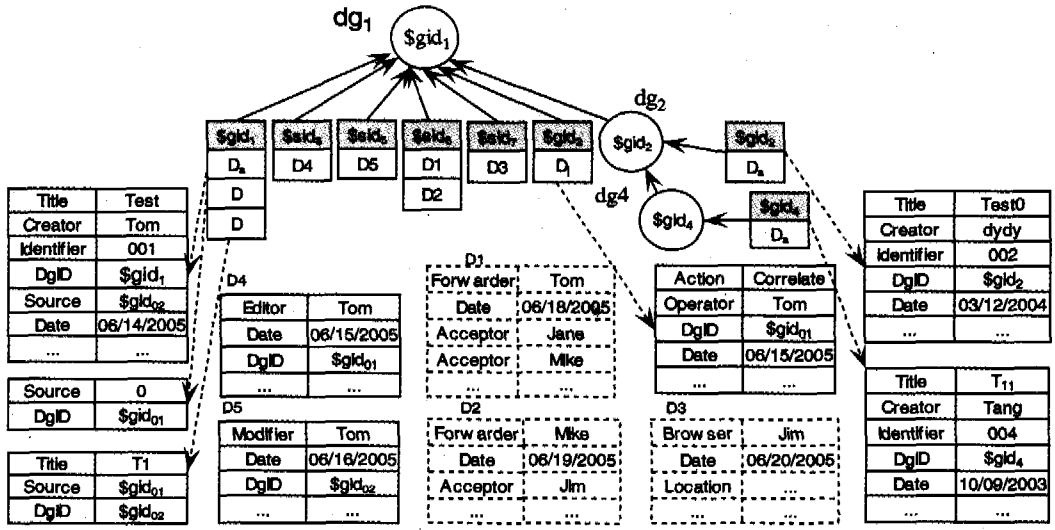
定义 5 中，DG' 称为 dg 的组成数据基因组，是对从其它数据源继承的数据的描述，一般是从来源数据的数据基因组中提取相关部分形成。数据关联了多少数据源，则其数据基因组中就有多少个组成数据基因组。数据基因组中组成数据基因组的个数称为数据基因组的基数，记为 $|dg|$。S^+ 是基因序列的正闭包，数据基因组中至少要包含主基因序列 S_m。对 DG'^* 中的组成数据基因组，在 S^+ 中存在一个与之对应的数据基因序列以记录数据基因组被关联的详细信息。

在此，我们没有限定一个数据基因组中应该包含多少个基因序列，只规定一个基因序列反映数据的某一性质的进化过程。与基因序列相对应，数据基因组是对数据进化过程全方位的描述，是从"面"的角度对数据进化过程的完整描述。

定义 6 不包含组成数据基因组的数据基因组，即 $|dg|=0$ 的数据基因组，称为元数据基因组。

新生成数据的数据基因组就是一个特殊的元数据基因组。它只有一个主基因序列，没有组成数据基因组。在数据的进化过程中，若数据与其他的数据发生关联，则要添加相应的数据基因组作为其组成数据基因组。随着数据的进化，数据基因组会包含多个基因序列和多个组成数据基因组。

图 1 是数据 Test 的数据基因组，dg_1 表示整个数据基因组，它由六个基因序列和一个组成数据基因组构成。其中 $\$gid_1$ 为其主基因序列，是对数据 Test 的整体信息描述。它包含一个主导基因和记录主导基因两次发生变化的数据基因。基因序列 $\$sid_3$ 是对数据编辑信息的历史记录。$\$sid_5$ 详细记录用户对数据整体信息的修改，如用户改变数据的标题、关键词等。$\$sid_6$ 则记录数据转发信息。如 Tom 2005 年 6 月 18 日将信息转发给 Mike 和 Jane。Mike 于 2005 年 6 月 19 日将数据又转发给 Jim。基因序列 $\$sid_7$ 中数据基因记录浏览数据的操作。基因序列 $\$gid_2$ 与组成数据基因组 dg_2 相对应。$\$gid_2$ 中只有一个数据基因，表示 Tom 于 2005 年 6 月 15 日将此数据基因组结合到 $\$gid_{01}$ 中。基因序列 $\$sid_5$ 中只包含主导基因。dg_2 是 Test0 数据基因组的组成数据基因组，这说明 Test0 的部分遗传信息传递到了 Test 中。而 dg_4 又是 dg_2 的组成数据基因组，即 Test0 从 T_{11} 继承了部分内容且又被 Test 引用。



1)对图中 source 属性的取值,我们规定若一份数据有原创成份,则在其 source 属性中添加 0 作为标识。当一份数据从其它数据中复制数据时,将其数据基因组的 $sid 写入 source 属性中。可参看数据基因组融合操作定义。2)图中为表示简单,序列的标识符和数据基因均表示在一个表中,以背景色区分。下同。

图 1 数据的数据基因、数据基因序列和数据基因组

3.2 数据基因、数据基因组之间的关系

设数据基因的集合为 G,基因序列的集合为 S,数据基因组的集合为 DG,属性的全集为 A,$g\in G$,$a\in A$,$s\in S$ 且 $dg\in DG$。下面列出下文中用到的一些符号及其含意。

$ID(s)$:基因序列 s 的唯一标识符 $\$sid$;

$T(s)$:S 中所有数据基因的集合;

$DgID(dg)$是数据基因组 dg 的唯一标识符 $\$gid$;

$I(dg)$:dg 的主导基因;

$S(dg)$:dg 的基因序列的集合;

$S_m(dg)$:dg 的主基因序列

$\Psi(dg)$:dg 的组成数据基因组集合;

$\Pi_a(g)$:数据基因 g 在属性 a 上的投影值。

另外,数据标识我们用 identifier 表示,它存在于数据的主导基因中。一个数据基因组的主基因序列唯一,且主基因序列的标识也唯一,因此可以用主基因序列的标识来标识整个数据基因组,即 $DgID(dg)=ID(S_m(dg))=\$gid$。

定义 7 等价关系(Equivalence,≡):

两数据基因 g_i,g_j 处于等价关系,$g_i\equiv g_j\Leftrightarrow$

$(\forall f\in g_i:f\in g_j)\wedge(\forall f'\in g_j:f'\in g_i)$

两基因序列 s_i,s_j 处于等价关系,$s_i\equiv s_j\Leftrightarrow$

$(\forall g_i\in T(s_i),\exists g_j\in T(s_j):g_i\equiv g_j)\wedge(\forall g_i\in T(s_j),\exists g_i\in T(s_i):g_j\equiv g_i)$

两数据基因组 dg_i,dg_j 处于等价关系,$dg_i\equiv dg_j\Leftrightarrow$

$(DgID(dg_i)=DgID(dg_j)\wedge(\forall s_i\in S(dg_i),\exists s_j\in S(dg_j):s_i\equiv s_j)\wedge(\forall dg_k\in\Psi(dg_i),\exists dg_l\in\Psi(dg_j):\in dg_k\equiv dg_l)\wedge(\forall s_j\in S(dg_j),\exists s_i\in S(dg_i):s_i\equiv s_j)\wedge(\forall dg_k\in\Psi(dg_j),\exists dg_l\in\Psi(dg_i):dg_k\equiv dg_l)$

若两数据基因包含属性完全相同且对应属性取值完全相等,则两数据基因等价。若两基因序列中包含的数据基因对应等价,则两个基因序列等价。若两数据基因组标识相等、基因序列对应等价且所包含组成数据基因组也对应等价,则两数据基因组处于等价关系。可见若两数据基因等价,则对应数据所经历的一次操作相同。若两基因序列等价,则对应数据在某一方面的经历相同。而若两数据基因组等价,其对应的数据内容、经历完全相同。或者是其中一个数据是另一个数据的复本。

定义 8 同源关系(Isogeny,≈)

两数据基因 g_i,g_j 处于同源关系:

$g_i\approx g_j\Leftrightarrow\exists f\in g_i:f\in g_j$

两数据基因序列 s_i,s_j 的同源关系:

$s_i\approx s_j\Leftrightarrow\exists g_i\in T(s_i),\exists g_j\in T(s_j):g_i\approx g_j$

另外,当 s_i,s_j 是主基因序列时我们还要求:$I(s_i)\approx I(s_j)$

两数据基因组 dg_i,dg_j 的同源关系:

$dg_i\approx dg_j\Leftrightarrow(S_m(dg_i)\approx S_m(dg_j))\wedge\prod_{Identifier}(I(dg_i))=\prod_{Identifier}(I(dg_j))$

若两数据基因处于同源关系,我们称其中一个是另一个的同源数据基因。若两基因序列处于同源关系,我们称其中一个是另一个的同源基因序列。若两数据基因组处于同源关系,则我们称其中一个是另一个的同源数据基因组。两个数据基因组处于同源关系,说明两个数据基因组是同一数据的数据基因组,它们可能对应数据的不同版本或者具有不同传播经历的同一版本。可见数据基因、基因序列和数据基因组是的等价关系是都是对应同源关系的特例,且三种结构上的同源关系都是自反、对称且传递的。

对数据基因组的同源关系,我们又把它分覆盖和继承两类。覆盖是指具有相同 $\$gid$ 的两个数据基因组的同源关系。此时两数据基因组对应数据的同一版本,但经历不同。继承则指两个不同 $\$gid$,但具有相同数据 identifier 的数据基因组的同源关系,此时两数据基因组对应数据的不同版本。

定义 9 覆盖(Cover,≻)

数据基因 g_i 覆盖 g_j,$g_i\succ g_j\Leftrightarrow(\forall f'\in g_j:f'\in g_i)$。

基因序列 s_i 覆盖 s_j,$s_i\succ s_j\Leftrightarrow\forall g'\in T(s_j),\exists g''\in T(s_i):g''\succ g'$。

数据基因组 dg_i 覆盖 dg_j,$dg_i\succ dg_j\Leftrightarrow$

$(\forall s'\in S(dg_j),\exists s''\in S(dg_i):s''\succ s')\wedge(\forall dg'\in\Psi(dg_j):dg'\in\Psi(dg_i))\wedge DgID(dg_i)=DgID(dg_j)$

定义 9 中,我们称 g_j 为 g_i 的子数据基因,称 g_i 为 g_j 的覆盖数据基因。同样,称 s_j 为 s_i 的子基因序列,s_i 是 s_j 的覆

盖基因序列。称 dg_j 为 dg_i 的子数据基因组，称 dg_i 为 dg_j 的覆盖数据基因组。数据基因组的覆盖关系是指 dg_j 中数据基因都在 dg_i 中，是 dg_i 中某一个数据基因的子数据基因。可见数据基因、基因序列及数据基因组上的覆盖关系都是自反的、传递的。

定义 10 继承(Derivation，$\rightarrowtail$)

数据基因组 dg_j 继承于数据基因组 dg_i 当且仅当

$dg_i \approx dg_j \wedge DgID(dg_i) \neq DgID(dg_j) \wedge \amalg_{Identifier}(I(dg_i)) = \amalg_{Identifier}(I(dg_j)) \wedge (T(S_m(dg_i)) - \{I(dg_i)\} \subseteq T(S_m(dg_j)) - \{I(dg_j)\} \wedge ((\forall s' \in S(dg_i) - \{S_m(dg_i)\}) \wedge (\forall s'' \in S(dg_j) - \{S_m(dg_j)\}) \wedge (I(s'') = I(s')) : s'' \succ - s'$，记作 $dg_i \rightarrowtail dg_j$。

此时我们称 dg_j 为 dg_i 的后代数据基因组，dg_i 称为 dg_j 的祖先数据基因组。显然，后代数据基因组从其祖先数据基因组中选择的继承部分遗传信息。具体继承的内容我们在另文中给出。数据基因组的继承关系也是可传递的。

定义 11 数据基因组族(Data Genome Family)

所有同源的数据基因组的集合称为数据基因组族，记作 $\xi(dg_i)$，dg_i 是其中任意数据基因组。

一个数据基因组族中所有数据基因组都是对同一数据相同或不同版本的遗传信息的描述。任意两个数据基因组若处于同一数据基因组族中，则它们处于同源关系，具体讲，它们可能具有覆盖或继承关系，或者都与另一个数据基因组具有覆盖或继承关系。

对不同数据，若数据之间存在关联，其数据基因组之间也同样存在关联，下面给出定义。

定义 12 关联(correlate，$\triangleright$)

若两数据基因组 dg_i，dg_j，满足条件 $dg' \in \rho(dg_i) \wedge dg' \in \Psi(dg_j) \wedge (\exists s \in S(dg_j) \wedge ID(s) = ID(S_m(dg_i)))$($\rho(dg_i)$ 见定义 14)，则称 dg_i 关联到 dg_j 或者 dg_i 是 dg_j 的关联数据基因组，记作 $dg_i \triangleright dg_j$。

由定义 12 条件可知，数据基因组 dg_j 融合了 dg_i 中的部分遗传信息。此时 dg_j 对应的数据中包含了部分 dg_i 对应数据的内容。如图 1 中，dg 有一个组成数据基因组 dg_2，它对应 Test 中部分数据是从 Test0 中拷贝来的。

4 数据基因模型的性质

下面以命题形式给出数据基因组模型的性质：

命题 1 每个数据基因至少含有一个数据基因片断。

定义 1 中，数据基因是基因片断的正闭包，通常每个数据基因至少含有一个表示操作的属性。但是仅有这一个属性，数据基因没有任何意义，因此通常数据基因含有多个基因片段。如记录操作的数据基因一般具有下列属性：operator，timestamp，place，action 等。对主导基因，则至少包括下列属性：identifier，source，timestamp，owner 等。

命题 2 每个数据基因组有且仅有一个主导基因。

主导基因包含数据的整体信息，其结构和作用与元数据类似。每个组成数据基因组均有自己的主导基因，但是属于整个数据基因组的主导基因只有一个。这个主导基因一般存放在主基因序列中。

命题 3 每个数据基因组可以有多个基因序列，但有且仅有一个主基因序列。

主基因序列记录数据的整体信息和演变历程。由于主基因序列在数据基因组中是唯一的，我们也用主基因序列标识符来标识数据基因组。相对而言，其它基因序列包含的则是数据在某一方面的信息。我们没有严格规定每个数据基因组中必然包含多少基因序列，使用数据基因组时可依据不同应用来确定具体基因序列的划分。

命题 4 若将数据基因组看作根结点、基因序列看作叶子节点，组成数据基因组看作分支节点，则它是一棵树。

对新建的原创数据，其数据基因组是最简单的树，只含有一个根节点和一个叶子节点，其中的叶子节点是主基因序列。

为描述同一数据的数据基因组之间的关系，我们给出如下定义。

定义 13 直接覆盖(The directed cover data genome)

若数据基因组 dg' 满足 $(dg' \succ - dg) \wedge (\neg \exists dg'' : dg'' \succ - dg' \wedge dg'' \succ - dg)$ 则，它是数据基因组 dg 的直接覆盖，记作 $dg' \geqslant dg$。

定义 14 覆盖数据基因组集(Cover data genome set)

所有 dg 的覆盖数据基因组的集合称为其覆盖数据基因组集，记作 $\rho(dg)$。

$\rho(dg)$ 中的数据基因组与 dg 描述的是同一版本数据的遗传信息。且 dg 是其中最小的数据基因组，即 $\rho(dg)$ 任意数据基因组都是它的覆盖数据基因组。它们之间具有如下关系。

命题 5 直接传播树(Direct propagation tree，DPT)

dg 与其所有覆盖数据基因组集 $\rho(dg)$ 中的数据基因组按直接覆盖关系形成一棵以 dg 为根的有向树，我们称之为 dg 的直接传播树，记作 $T = \langle V, E \rangle$

其中 $V = \rho(dg)$，$E = \{\langle dg_j, dg_k \rangle | (dg_j, dg_k \in \rho(dg)) \wedge (dg_k \geqslant dg_j)\}$。

这一命题成立的依据是数据的传播路径是一棵有向树，在此我们不给出严格证明。dg 的直接传播树表示了数据从 dg 所表示状态开始的传播历程。集合中每一个结点是其传播历程中的一个中间站。由此我们还可以得出，对某一个版本的数据，设其最初版本数据的数据基因组为 dg_{IN}，则任意满足 $ID(S_m(dg_i)) = ID(S_m(dg_{IN}))$ 的数据基因组 dg_i 都是其覆盖数据基因组。且以 dg_{IN} 为根的直接传播树描述这一版本数据传播的全历程，其它所有 $dg \in \rho(dg_{IN})$ 的直接传播树是这棵直接传播树的子树。

可见，任意两个数据基因组，若二者有覆盖关系，或者二者与同一个数据基因组有覆盖关系，则它们是同一版本数据的数据基因组。所有同一版本数据的数据基因组可看作一棵树，若将这样的树看作一个结点，则在数据基因组族中我们可得到另一棵有向树。

定义 15 直接继承数据基因组(Direct derived gata genome)若数据基因组 dg' 满足条件

$(dg \rightarrowtail dg') \wedge (\neg \exists dg'' : dg \rightarrowtail dg'' \wedge dg'' \rightarrowtail dg')$

则称 dg' 是 dg 的直接继承数据基因组，记作 $dg \vdash\!\!\rightarrowtail dg'$。

命题 6 数据基因组的家族树(Family tree)

对每一个数据基因组族 $\xi(dg)$，令 $DG_i (i = 1 \cdots n)$ 为 $\xi(dg)$ 的子集，且 $\forall dg', dg'' \in DG_i : \rho(dg') \cap \rho(dg'') \neq \phi$，则 $\xi(dg)$ 中 $DG_i (i = 1 \cdots n)$ 按直接继承关系形成一棵有向树，我们称之为 dg 的家族树，记作 $T = \langle V, E \rangle$ 其中

$V = \{DG_i | DG_i \subseteq \xi(dg)\}$，$E = \{\langle DG_j, DG_k \rangle | (DG_j, DG_k \in V, \forall dg' \in DG_j, dg'' \in DG_k) dg' \vdash\!\!\rightarrowtail dg''\}$

这一命题的正确性由数据传播历程的树型结构保证，在此不给出严格证明。家族树中根结点是数据第一版本的覆盖

数据基因组集，根结点的孩子结点是其继承数据基因组的覆盖数据基因组集。如此类推，叶子结点则是“最新版本”数据的数据基因组集。家族树也可看作一个版本树，不同结点表示数据的不同版本的基因信息。而每个结点内部又是一棵关于此版本数据传播历程的直接传播树(DPT)。

5 DGM的一个参考实现

上面几节我们给出了一个数据基因模型的概念模型。本节给出数据基因模型的一个实例——文件的数据基因模型。此模型中我们定义了6种基因序列，见表1。基因序列中所使用的部分属性见表2，其它属性可参看文[16]。图1也可看作是一个文件的数据基因模型。限于篇幅，基因片段结构没有详细列出。

表1 基因序列的类型

Type	描述	属性集	图1中对应序列
S_m	主基因序列，描述数据的整体信息	{DgID, Identifier, Title, Creator, Date, Description, Format, Keywords, Language, Location, Notes, Organization, Publisher, Type, Source, Rights}	$\$gid_1$
S_e	数据的编辑历史	{Editor, Date, Location, Validity, DgID}	$\$sid_3$
S_f	数据的传播历程	{Forwarder, Acceptor, Date, Location, Rights}	$\$sid_6$
S_b	数据的浏览信息	{Browser, Date, Location, Validity}	$\$sid_7$
S_c	与其它数据关联的详细信息	{DgID, Operator, Action, Date, Location}	$\$gid_2$
S_g	主导基因的修改历程	{Modifier, Date, Location, DgID}	$\$sid_5$

表2 表1中部分属性描述

属性	描述	是否显性基因?
DgID	数据基因组标识	N
Source	数据来源(当前数据从何处继承生成)($gid)	N
Validity	操作有效期	N
Acceptor	文件的acceptor	Y
Browser	文件浏览者	Y
Forwarder	数据发布者	N
Editor	数据编辑者	N
Modifier	数据属性编辑者	N
Operator	数据操作者	N
Action	操作者采取何种动作	N

若应用程序支持DMG模型，当我们创建文件时，会同时生成一个包含主基因序列和主导基因的数据基因组。若对数据进行某种操作，数据基因组也会相应进行遗传、变异操作。如从数据源向数据拷贝内容，则数据基因组会进行融合操作。而数据的发布、浏览等操作，对应数据基因组会执行加尾操作。若对数据属性等整体信息进行修改，数据基因组会执行变异操作。融合、加尾、变异等操作是数据基因组遗传、变异的具体实现，详细定义另文给出。

总结与展望 本文给出了一种对数据进化过程中产生的信息进行记录、管理和应用的数据模型——数据基因模型。其中包括数据基因、数据基因组等定义，数据基因、数据基因组之间的关系，并以给出文件的数据基因模型的参考实现。但对这种模型的研究还仅仅处于初始阶段，还要进行不断的完善。

我们近期的工作包括：首先数据基因模型的操作与代数操作定义我们将另文给出。以此为基础建立数据基因查询语言是我们近期要做的工作。第二，给出数据基因模型的存储模型，结合实验建立数据基因模型的抽取、存储与分析工具。第三，结合实验对数据基因模型进行优化。第四，根据数据的基因信息建立信息的可信度、可靠性模型，分析数据真伪，并根据继承信息恢复对数据的破坏，以及解决信息加工过程中的权限问题，加强信息的可控度等问题也是我们下一步要做的工作。

另外，考虑对多变环境下工作流、service等动态概念的描述，在较新的更复杂的数据环境下，对信息、功能、service的集成与发布。以数据基因模型为基础，结合P2P网络，建立语义更丰富的信息集成与发布平台，也是今后的工作重点之一。

参考文献

1 奚建清，汤德佑，郭玉彬. 数据基因：数据的遗传信息载体. 计算机工程(已录用)
2 http://dublincore.org/
3 Tansel A, Clifford J, Gadia S, et al. Temporal Databases: Theory, Design, and Implementation. Benjamin/Cummings, 1994
4 http://www.w3.org/TR/xpath20/
5 Amagasa T, YoShikawa M, Uemura S. Realizing Temporal XML Repositories Using Temporal Relational Databases. In: CODAS, 2001: 63～68
6 Amagasa T, YoShikawa M, Uemura S. A data model for temporal XML Documents. In: Proc of 11th International Confrence on Database and Expert Systems Applications(DEXA 2000), 2000. 334～344
7 Damiani E, Oliboni B, Quintarelli E, et al. Modeling Semistructured Data by Using Graph-based Constraints. In: OTM Workshops Proceedings, LNCS, 2003. 20～21
8 Damiani E, Oliboni B, Quintarelli E, et al. Temporal Aspects of Semistructured Data. In: SEBD, 2001. 215～222
9 Combi C, Oliboni B, Quintarelli E. A Graph-Based Data Model to Represent Transaction Time in Semistructured Data. In: DEXA, 2004. 559～568
10 王宁，等. 数据树——一种用于异构数据源集成的公共数据模型. 计算机研究与发展，1998，35(7)：610～615
11 王宁，王能斌. 异构数据源集成系统查询分解和优化的实现. 软件学报，2000(2)
12 陈彤兵，等. 分布式自治数据源的联合查询. 计算机研究与发展，2004(4)
13 Kumaran S, Nandi P, Heath T, et al. Raja Das: ADoc-Oriented Programming. In: SAINT, 2003. 334～343
14 Ye Yiming, Nandi P, Kumaran S. Smart Distance for Information Systems: The Concept. IEEE Computational Intelligence Bulletin, 2003, 2(1): 25～30
15 Nandi P, Kumaran S. Adaptive Business Objects - A new Component Model for Business Integration. ICEIS, 2005(3): 179～188
16 Päivärinta T, Tyrväinen P, Ylimäki T. Defining Organizational Document Metadata: A case beyond standards. Proc of the 5th European Conf on Information Systems, Gdansk, 2002
17 Agrawal R, Imielinski T, Swami A. Mining Association Rules Between Sets of Items in Large Databases, In: Proc ACM SIGMOD, 1993. 207～216
18 Chen M S, Park J S, Yu P S. Efficient Data Mining for Path Traversal Patterns. IEEE Trans on Knowledge and Data Eng, 10(2): 209～221