ОНТОЛОГИЧЕСКИЕ И РЕЛЯЦИОННЫЕ МОДЕЛИ ДАННЫХ В КООРДИНАТНОМ ПРЕДСТАВЛЕНИИ

В. И. Емельяненков

Белорусский государственный университет
Минск, Беларусь
E-mail: emelvi@bsu.by

В данной работе рассматриваются вопросы координатного представления онтологических и реляционных моделей данных, которые могут быть заданы как иерархические структуры в виде решетки. Обсуждаются механизмы построения двумерного растра и методы отображения на него узлов решетки, которые позволяют использовать координаты точек растра в качестве индексов узлов задаваемых схем моделей данных.

Ключевые слова: иерархические структуры, онтологические модели данных, реляционные модели данных, решетки, индексирование.

В задачах обработки данных распространение получили модели данных, схемы которых представляют некоторую решетку. Под решеткой будем понимать алгебраическую структуру, которая в случае реляционных и онтологических моделей рассматривается как таксономия концептов. Ее суть состоит в следующем [1].

На множестве объектов U и атрибутов V определено отношение $I \subseteq U \times V$, такое, что pIa, где $p \in U$, $a \in V$, тогда и только тогда, когда a есть атрибут объекта p. Тройка $K = (U, V, I)$ называется формальным контекстом.

Формальный контекст может быть представлен в виде бинарной матрицы, строки которой помечены именами объектов, а столбцы – значениями атрибутов.

Тогда пара (P, G), удовлетворяющая условиям: $P \subseteq U$, $G \subseteq V$, $P' = G$, $G' = P$, называется формальным понятием (концептом) контекста $K = (U, V, I)$. Множество объектов P составляет объем понятия, а множество всех атрибутов G, которыми они обладают, – содержание понятия. Таким образом, формальное понятие – это множе-

ство объектов из данной предметной области, каждый из которых обладает всеми атрибутами из некоторого подмножества атрибутов, присущих этим объектам.

Множество формальных понятий (P, G), где $P \subseteq U$, $G \subseteq V$, частично упорядочено отношением R (можно назвать его, например, «менее общий чем или равен»): (P_1, G_1) ≤ ≤ (P_2, G_2), если $P_1 \subseteq P_2$ или $G_2 \subseteq G_1$ (что эквивалентно) – и образует концептуальную решетку контекста K.

Известно, что для формального контекста выделенного фрагмента предметной области множество формальных понятий образует полную решетку концептов [2]. Это позволяет представлять ее в виде многоуровневой иерархической структуры, как это показано на рис. 1. В узлах размещаются концепты с двойными индексами, представляющими, соответственно, номер слоя и номер концепта в слое. На верхнем уровне размещаются первичные понятия (в базах данных таблицы), которые не имеют родителей. Далее вниз по иерархии следуют производные концепты.

В настоящее время при достаточно приемлемых ресурсных ограничениях возможна организация хранения данных, схемы которых содержат сотни тысяч концептов. В то же время известные алгоритмы целенаправленного их отбора и последующей обработки лишь частично покрывают потребности анализа. Поэтому поиск и исследование новых возможностей в этом направлении являются актуальными.

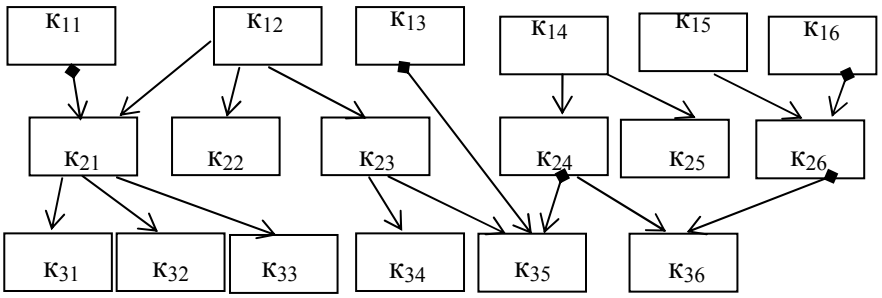

Рис. 1. Пример диаграммы концептуальной решетки

Определенный интерес представляют пространственно-координатные формы реализации моделей данных, которые в определенных условиях предоставляют довольно широкие возможности. Так, например, широкое распространение получили OLAP-технологии, на основе которых развернут богатый спектр аналитических исследований по методологиям Data Mining.

В работе [3] был рассмотрен координатный подход по отношению к иерархическим моделям. В данной работе делается попытка использования координатных представлений рассматриваемого типа моделей данных в достаточно общей форме. В основу кладется разделение решетки на две компоненты: древовидную иерархию и набор удаленных дуг. После этого формируется представление решетки, так, как это изображено на рис. 2.

В данном случае в качестве примера берется та же решетка, что представлена на рис. 1. Из нее после удаления дуг, начало которых помечено зачерненным ромбиком, получается набор изолированных деревьев, где вершинами являются концепты верхнего уровня $к_{11}$, $к_{12}$, $к_{13}$, $к_{14}$, $к_{15}$ и $к_{16}$. Этот набор деревьев представлен в верхней части рис. 2. При этом концепты показаны как узлы, помеченные зачерненными кружочками с указанием их индексов, которые они имели на рис. 1. Соответственно, в

нижней части рисунка представлены те узлы, между которыми удалены соединяющие их дуги. Это пары узлов ($к_{11}$, $к_{21}$), ($к_{13}$, $к_{35}$), ($к_{24}$, $к_{35}$) и ($к_{16}$, $к_{26}$).

Особенностью данного представления деревьев является то, что их узлы имеют координатное представление, идея которого состоит в следующем. На оси Y строится равномерная шкала целочисленных значений r = 1, 2, …, а затем на ней строится растр, состоящий из горизонтальных линий $Y_r = r$ равномерной шкалы и двух семейств гипербол вида

$$Y^0_i = i/x \quad \text{и} \quad Y^1_j = j/(1 - x), \tag{1}$$

где значения координаты x заданы на интервале от нуля до единицы $x \in (0, 1)$, индексы i = 1, 2, ... , …, «выбирают» необходимые линии семейства Y^0_i, а индексы j = 1, 2, ..., соответственно, образуют номер линии семейства Y^1_j.

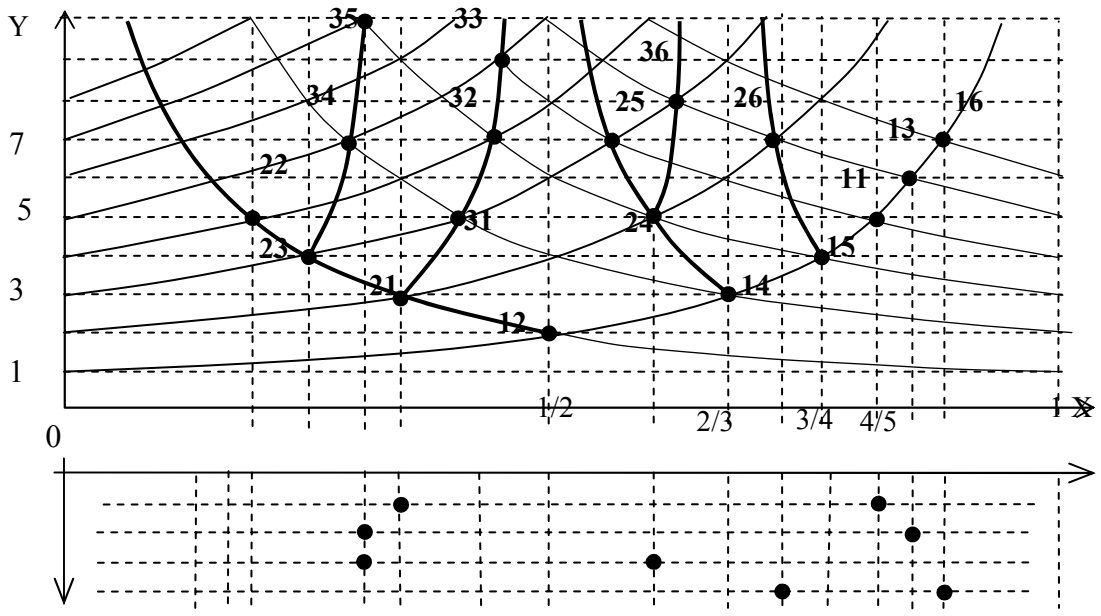

Рис. 2. Отображение концептов на узлы растра

Пересечение гипербол Y^0_i и Y^1_j между собой происходит в точках, расположенных на линиях равномерной шкалы $Y = r$, которые образуют множество $\{Y_{i/r}\}$ узлов растра (где это не оговорено особо, координаты задаем по Y^0_i). Несложно заметить, что координаты x_q точек $Y_{i/r}$ являются элементами множества рациональных чисел из интервала (0, 1), и имеет место равенство $x_q = i/r$, где i – номер кривой Y^0_i, а r – номер линии равномерной шкалы. Поэтому индекс в обозначении $Y_{i/r}$ можно рассматривать как запись значения координаты $x_q = i/r$ в виде структуры i/r.

На это множество $\{Y_{i/r}\}$ и отображаются узлы иерархической структуры. Рассмотрим более подробно механизм реализации данного подхода.

Вначале отобразим концепты верхнего уровня на некоторое подмножество точек растра, координаты которых образуют сходящуюся последовательность на всем интервале (0, 1). В данном случае мы можем взять непосредственно одну из кривых семейств Y^0_i или Y^1_j.

Координаты x_q кривой Y^1_1 семейства Y^1_j образуют последовательность чисел {0, 1/2, 2/3, 3/4, 4/5, …}, сходящуюся к значению x = 1. Выберем ее для размещения концептов первого уровня в отмеченных точках, как это показано на рис. 2. Узел с координатой x = 0 будем рассматривать как виртуальный корень. Последующие узлы рас-

тра $Y_{i/r}$ переопределим как $K^{mn}{}_{i/r}$, оставив нижний индекс i/r и введя верхний индекс mn для концепта, где m – номер слоя, а n – номер концепта в слое. Тогда на выбранной кривой $Y^1{}_1$ будет получен следующий набор таких узлов: $K^{12}{}_{1/2}$, $K^{14}{}_{2/3}$, $K^{15}{}_{3/4}$, $K^{11}{}_{4/5}$, $K^{13}{}_{5/6}$, $K^{16}{}_{6/7}$, соответственно с координатами 1/2, 2/3, 3/4; 4/5, 5/6, 6/7.

Аналогичным образом можно построить ветви следующего (второго) уровня иерархии. Здесь уже концепты $K^{12}{}_{1/2}$, $K^{14}{}_{2/3}$, $K^{15}{}_{3/4}$, $K^{11}{}_{4/5}$, $K^{13}{}_{5/6}$, $K^{16}{}_{6/7}$ будут корневыми для своих деревьев. При этом следует иметь в виду, что сходящиеся последовательности на сей раз формируются не на всем интервале (0, 1), а каждая в своем. Дело в том, что концепты первого уровня делят его на области следующим образом. Первую область образует интервал (0, 1/2), вторую область образует интервал (1/2, 2/3) и т. д. За каждым концептом закрепляется определенная область общего интервала (0, 1), так, чтобы пространства образованных ими деревьев не пересекались. А именно дерево концепта $K^{12}{}_{1/2}$ будет размещаться в области (0, 1/2), дерево концепта $K^{14}{}_{2/3}$ будет расположено на интервале (1/2, 2/3) и т. д.

Поясним вначале порядок построения концептов второго слоя на примере узла $K^{12}{}_{1/2}$. В данном случае есть кривая $Y^0{}_1$ семейства $Y^0{}_i$, которая проходит через узел $K^{12}{}_{1/2}$ и содержит узлы растра, координаты которых образуют сходящуюся последовательность чисел {1/2, 1/4, 1/5, 1/6, …} на интервале (0, 1/2). На ней мы и расположим узлы второго уровня $K^{21}{}_{1/3}$, $K^{23}{}_{1/4}$, $K^{22}{}_{1/5}$, которые подчинены узлу $K^{12}{}_{1/2}$.

Теперь перейдем к узлу $K^{14}{}_{2/3}$. Для него нет кривой семейства $Y^0{}_i$ или $Y^1{}_j$, проходящей через рассматриваемый узел так, чтобы непосредственно на ней в интервале (1/2, 2/3) получить сходящуюся последовательность точек растра. Однако из имеющихся точек растра всегда можно выбрать такие, координаты которых будут образовывать нужную последовательность, сходящуюся в данном случае к значению $x = 1/2$. Как показано на рис. 2, на них отображаются концепты $K^{24}{}_{3/5}$, $K^{25}{}_{4/7}$.

Аналогичным образом получим все оставшиеся концепты $K^{26}{}_{5/7}$, $K^{31}{}_{2/5}$, $K^{32}{}_{3/7}$, $K^{33}{}_{4/9}$, $K^{34}{}_{2/7}$, $K^{35}{}_{3/10}$, $K^{36}{}_{5/8}$.

Итак, получено конструктивное отображение концептов решетки на множество точек построенного растра. При этом все они проецируются на ось X, образуя одномерный эквивалент структуры набора деревьев.

В рассмотренной системе также решается задача идентификации принадлежности концепта соответствующей ветви дерева по его координате $x^*{}_q = i^*/r^*$. По построению координата любого узла лежит внутри определенного интервала, границами которого являются ближайшие слева и справа к i^*/r^* узлы концептов предшествующего уровня. Один из них будет родительским i^p/r^p, а второй сопряженным с ним i^s/r^s. При этом координатой родительского узла будет та из двух, у которой значение r больше ($r^p > r^s$). Далее в свою очередь координата этого родительского узла попадает в интервал, одна из границ которого определяет последующий родительский узел. И так можно продолжать вплоть до узла концепта первого уровня.

Также возможен и вариант определения родительского узла путем прямого вычисления его координаты по заданным значениям $x^*{}_q = i^*/r^*$. Конкретно техника дела здесь такова. Координата любого концепта входит в последовательность, сходящуюся к той границе содержащего ее интервала, которая является координатой i^s/r^s узла, непосредственно сопряженного с родительским узлом. При этом обязательно хотя бы на одной из ближайшей к r^* горизонтали $r^* + 1$ или $r^* + 1$ будут пересекаться кривые (1), одна из которых проходит через узел с координатой i^*/r^*. Рассмотрим в качестве примера концепт $K^{31}{}_{2/5}$. Он находится на пересечении прямой $r = 5$ и кривых $Y^0{}_2$ и $Y^1{}_3$.

Соответственно, на линии $r = 6$ гипербола ${Y^0}_2$ пересекается с Y_{14}, а кривая ${Y^1}_3$ пересекается с ${Y^0}_3$. Аналогично на линии $r = 4$ имеет место узел растра при пересечении ${Y^0}_2$ и ${Y^1}_2$. По крайней мере, один из трех узлов растра будет иметь нужную координату $i^s/r^s = 1/2$. Эта координата будет иметь значение $r^s = 2$, наименьшее из всех тех, которые были получены. Далее определяется родительский концепт ${K^{21}}_{1/3}$.

Таким образом, построено взаимно однозначное отображение узлов дерева на множество рациональных координат параметрической оси, позволяющее задать линейное перечисление элементов иерархической структуры. С этой структурой в той же координатной системе совмещается и множество удаленных дуг, которые можно интерпретировать как гиперссылки.

На практике это позволяет достаточно эффективно проектировать схемы прикладного анализа данных. Рассмотрим некоторые примеры.

Нетрудно заметить, что растровые структуры, содержащие деревья, хотя и формируются на интервалах разной длительности, на самом деле являются однотипными. Это позволяет строить многомерное пространство, задавая параметрические оси для каждого из деревьев в своем собственном измерении. При этом все деревья будут вложены в полностью идентичные растровые структуры. Как частный случай данного представления можно рассматривать такие древовидные информационные конструкции, как OLAP-кубы, сводные таблицы Excel и т.п. В общем случае древовидная иерархия расширяется наличием дуг-гиперссылок, и тогда в многомерной координатной форме может быть представлена схема более общей модели данных.

Одной из распространенных технологий Semantic Web является язык представления онтологий RDF (Resource Description Framework), которая базируется на построении графов, узлами и дугами которых являются триплетные отношения «субъект – предикат – объект». В пределах замкнутых предметных областей такие графы образуют решетки, на которых решаются задачи поиска и отбора данных по заданным наборам признаков. Конкретно решение таких задач сводится к установлению факта существования в общей структуре заданного подграфа, который совпадает с графом-шаблоном запроса.

Известно, что из-за большого объема процедур перебора время выполнения запросов оказывается достаточно высоким. В рассматриваемой координатной системе задача выделения нужного подграфа и сопоставление его с шаблоном запроса может решаться с помощью рассмотренных механизмов достаточно эффективным образом.

ЛИТЕРАТУРА

1. *Пронина, В. А.* Использование отношений между атрибутами для построения онтологии предметной области / В. А. Пронина, Л. Б. Шипилина //Журн. Пробл. управл. 2009. № 1. С. 27–32.

2. *Выхованец, В. С.* Понятийный анализ и контекстная технология программирования / В. С. Выхованец, В. Я. Иосенкин //Журн. Пробл. управл. 2005. № 4. С. 2–11.

3. *Емельяненков, В. И.* Линейное представление ссылок на элементы иерархических структур в задачах многомерного анализа данных / В. И. Емельяненков // БГУ. Информационные системы и технологии (IST’2002): сб. тр. Междунар. науч. конф. Минск. 5–8 ноября 2002 г. С. 223–227.