ИНФОРМАЦИОННЫЙ АНАЛИЗ

УДК 001.8:519.237.8]:519.17

Е. В. Трошин

Метод иерархической скалярной кластеризации и его применение к графовым моделям

Разработан метод иерархической скалярной кластеризации. С помощью этого метода происходит размещение объектов, представимых с помощью неориентированных графов, в базах данных и определение уникальности исследуемого объекта, т. е. поиск объекта в базе данных, идентичного данному. Метод позволяет исключить перебор записей, напротив, прямым доступом выходить в требуемую область базы данных.

Существует множество объектов, представимых с помощью неориентированных графов вида (рис. 1).

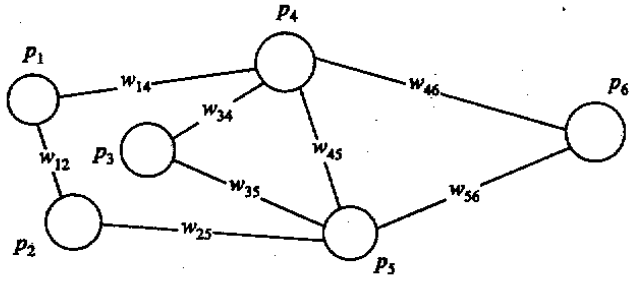

Рис. 1

Вершины данного графа обозначают сущности, из которых состоит моделируемый объект, а дуги — отношения между сущностями в данном объекте. Такой граф является связным. Между парой вершин в общем случае может быть множество связей-дуг или отношений, но при определенном рассмотрении такая сложная связь, состоящая из нескольких элементарных связей, может быть представлена одной дугой, символизирующей сложную связь между данными сущностями. Также в общем случае в графе могут быть петли, но их также можно опустить и представлять сущность, имеющую некую собственную связь, одной вершиной. При этом данная вершина соответствует сущности, имеющей внутреннюю связь, и отличается от аналогичной сущности, не имеющей внутренней связи. Некоторые дуги в графе могут быть ориентированы, но их ориентацию можно соотнести с типом связи (отношения) и на графе показывать в виде неориентированной дуги, изображая сам факт связи между вершинами, полагая, что тип связи, включающий и ориентацию, описан отдельно [1].

Такой граф записывается кортежем:

$$\Gamma =< A, P, W >, \qquad (1)$$

где A — матрица смежности графа, P — матрица-столбец весов вершин (типов сущностей), W — матрица-столбец весов дуг (типов отношений между сущностями).

Для графа на рис. 1 матрица A имеет вид:

$$A = \begin{bmatrix} a_{11} & a_{21} & \dots & a_{i1} & \dots & a_{N1} \\ a_{12} & a_{22} & \dots & a_{i2} & \dots & a_{N2} \\ \dots & \dots & \dots & \dots & \dots & \dots \\ a_{1j} & a_{2j} & \dots & a_{ij} & \dots & a_{Nj} \\ \dots & \dots & \dots & \dots & \dots & \dots \\ a_{1N} & a_{2N} & \dots & a_{iN} & \dots & a_{NN} \end{bmatrix}$$

$$A' = \begin{bmatrix} 0 & 1 & 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 & 1 & 1 \\ 0 & 1 & 1 & 1 & 0 & 1 \\ 0 & 0 & 0 & 1 & 1 & 0 \end{bmatrix}$$

Элементы матрицы A принимают значения 0 и 1, в зависимости от наличия связи между соответствующими элементами. Элементами матриц P и W являются натуральные числа, поставленные в соответствие типам сущностей и связям-отношениям между ними в исследуемом объекте. Способы присваивания натуральных чисел типам сущностей и типам отношений могут быть различными, главное заключается в том, чтобы каждый тип сущности и каждый тип отношения однозначно идентифицировался своим уникальным числом. Два разных типа сущностей, так же как и два разных типа отношений, не могут иметь одинаковые идентификаторы. В базе данных эти идентификаторы могут служить ключами для представленных там записей, соответствующих типам сущностей и связей. В простейшем случае это могут быть порядковые номера записей.

Примером использования описанной модели может служить представление в формализованном виде технических устройств. Здесь вершины графа или сущности являются элементами (блоками, субблоками, модулями, приборами) устройства, а дуги или отношения между сущностями являются способами соединения элементов [2].

С помощью описанной модели можно представлять также лингвистические конструкции, социальные, организационные, биологические системы, вообще любые объекты, в которых можно выделить сущности и установить отношения между ними.

Существует задача определения уникальности объекта, представленного моделью, показанной выше. Это может быть задача идентификации или сравнения двух объектов, например, определение новизны изобретения. При этом возникает проблема, которая заключается в том, что приходится обрабатывать достаточно большой массив информации. В этом массиве требуется найти объект, содержащий необходимое множество сущностей и отношений между ними.

Существующие методы основаны либо на прямом переборе объектов в базе данных, либо на способах построения дерева поиска и продвижении по нему, выбирая дальнейший путь в каждой вершине из множества альтернатив, и многократным возвратом в случае неудачи. Последние методы применимы, если база данных тем или иным образом структурирована по типам сущностей и связей [3].

Для решения задачи поиска заданного объекта в базе данных предлагается метод иерархической скалярной кластеризации. Он используется применительно к базе данных, содержащей сложноструктурированные объекты с множеством составляющих их сущностей и связей между ними. Метод позволяет не только обходиться без перебора всех записей в базе данных для поиска похожего объекта, но и абсолютно достоверно утверждать об уникальности данного объекта, если таковая действительно присутствует. Заключается метод в следующем.

Матрица A является некоторым образом графа и, соответственно, структуры объекта, представленного данным графом. Под структурой объекта здесь понимается набор неименованных сущностей и наличие неименованных связей между ними.

Примером такой структуры может являться печатная плата. В ней предусмотрены гнезда для элементов и протравлены дорожки между ними. Гнезда являются вершинами графа, а дорожки — дугами. Потенциально в гнезда могут быть вставлены различные электронные элементы. В этом случае придется говорить о типах вершин, содержание приобретут и связи, соединяющие элементы различных типов и различного функционального назначения. Другим примером структуры являются лингвистические конструкции. Так, предложение естественного языка состоит из подлежащего, сказуемого, дополнения, определения и др. В качестве этих членов предложения могут выступать различные части речи и конкретные слова.

Если структура объекта является уникальной, не похожей на структуры других объектов, то необходимо найти такой показатель для матрицы A, который соответствовал бы матрице A и только матрице A. Тогда этот показатель идентифицировал бы все объекты, имеющие структуру, описываемую матрицей A.

Сформулируем следующую теорему.

Теорема. *Для любого множества A бинарных матриц A_v одинаковой размерности $n \times m$ существует взаимно однозначное отображение множества A в множество S^A скалярных показателей S_v^A, являющихся образами соответствующих матриц A_v.*

Теорема говорит о том, что для любой матрицы A_v заданной размерности, элементами которой являются числа 0 и 1, найдется единственный скалярный показатель S_v^A, отличный от показателей, соответствующих другим аналогичным матрицам. Показатель S_v^A должен однозначно идентифицировать матрицу A_v.

В дальнейшем изложении индекс v опустим, так как будем рассматривать одну единственную матрицу A.

Доказательство. Рассмотрим матрицу B вида

$$B = \begin{vmatrix} 2^0\ 2^1\ 2^2\ \ldots\ 2^e\ \ldots\ 2^{n-1} \\ 2^n \ldots\ldots\ldots\ldots 2^{n-1} \\ 2^{2n} \ldots\ldots\ldots\ldots 2^{3n-1} \\ \ldots\ldots\ldots\ldots\ldots\ldots \\ 2^{(m-1)n} \ldots\ldots\ldots 2^{nm-1} \end{vmatrix}$$

Для n=5, m=4 матрица B принимает вид

$$B' = \begin{vmatrix} 2^0 & 2^1 & 2^2 & 2^3 & 2^4 \\ 2^5 & 2^6 & 2^7 & 2^8 & 2^9 \\ 2^{10} & 2^{11} & 2^{12} & 2^{13} & 2^{14} \\ 2^{15} & 2^{16} & 2^{17} & 2^{18} & 2^{19} \end{vmatrix}$$

Нетрудно увидеть, что элемент b_{ij} матрицы B равен

$$b_{ij} = 2^{n(j-1)+i-1}.$$

Если домножить каждый элемент a_{ij} матрицы A на соответствующий элемент b_{ij} матрицы B, то получим матрицу C, элементы которой есть числа

$$c_{ij} = a_{ij} 2^{n(j-1)+i-1}.$$

Таким образом, матрице A можно поставить в соответствие матрицу C, определенную выше. Сумма всех элементов матрицы C равна

$$a_{11}2^0 + a_{12}2^1 + \ldots + a_{ij}2^{n(j-1)+i-1} + \ldots$$
$$\ldots + a_{nm}2^{nm-1} = S^A. \qquad (2)$$

Но выражение (2) есть ни что иное, как формула перевода числа из двоичной системы счисления в десятичную, т. е. двоичного числа $a_{11}a_{12}\ldots a_{ij}\ldots a_{nm}$ в десятичное (натуральное, если не все a_{ij} равны 0) число S^A.

Каждой матрице A соответствует своя уникальная комбинация из $n \cdot m$ нулей и единиц, а это есть запись определенного двоичного числа. Таким образом, каждой матрице A соответствует число в двоичном коде, которое можно преобразовать по известным правилам в другую систему счисления, например, десятичную. Но данное преобразование всегда однозначно, поэтому каждая матрица A имеет свой уникальный скалярный показатель S^A. Скалярный показатель в виде натурального числа в десятичной системе счисления вычисляется по правилу

$$S^A = \sum_{i,j=1}^{i=n,\,j=m} a_{ij} 2^{(j-1)n+i-1}. \qquad (3)$$

Для нулевой матрицы скалярный показатель равен нулю. В рассматриваемой предметной области нулевая матрица A не имеет смысла за исключением варианта, когда объект состоит из одного единственного элемента, что в принципе возможно, но представляет собой экстремальный случай, и здесь не рассматривается.

Таким образом, любую матрицу A заданной размерности, состоящую из нулей и единиц, всегда

можно отобразить в единственный скалярный показатель, вычисляемый по определенному правилу и соответствующий только этой матрице из множества мощности $2^{nm}-1$, элементы которого являются различными матрицами той же размерности. Что и требовалось доказать.

Для бинарной матрицы смежности можно получить следствие из теоремы.

Следствие. *Для бинарной матрицы смежности A размерности n скалярный показатель равен сумме*

$$S^A = \sum_{j=1}^{n}\sum_{i=j}^{n} a_{ij}2^K, \text{ где } K = (n^2+n-(j-1)\times \times(2n-j))/2+i-2. \quad (4)$$

Матрица смежности обладает той особенностью, что $a_{ij} = a_{ji}$. Элементы, лежащие над главной диагональю, равны элементам, лежащим под главной диагональю. Для сокращения размерности скалярного показателя элементы, лежащие под главной диагональю, можно исключить из двоичного ряда, соответствующего матрице A, в результате чего он примет вид

$$a_{11}a_{21}\dots a_{i1}\dots a_{n1}a_{22}a_{32}\dots a_{n2}a_{33}a_{43}\dots a_{n3}\dots \dots a_{jj}a_{j+1j}\dots a_{nj}a_{j+1j+1}\dots a_{nn}.$$

Данный ряд состоит из $\dfrac{n^2+n}{2}$ членов. Рассмотренному ряду ставится в соответствие ряд $2^0 2^1 2^2 \dots 2^k \dots 2^N$, где $N = \dfrac{n^2+n}{2} - 1$. Необходимо вычислить степень k.

Рассмотрим две матрицы:

$$D = \begin{vmatrix} 0 & 1 & 2 & 3 & 4 \\ 5 & 6 & 7 & 8 & 9 \\ 10 & 11 & 12 & 13 & 14 \\ 15 & 16 & 17 & 18 & 19 \\ 20 & 21 & 22 & 23 & 24 \end{vmatrix}$$

$$G = \begin{vmatrix} 0 & 1 & 2 & 3 & 4 \\ & 5 & 6 & 7 & 8 \\ & & 9 & 10 & 11 \\ & & & 12 & 13 \\ & & & & 14 \end{vmatrix}$$

Для матрицы D значение элемента в зависимости от индексов i и j равно $d_{ij} = n(j-1)+i-1$, где n — размерность матрицы, в данном случае — 5. В матрице G показаны только элементы, лежащие не ниже главной диагонали. При переходе с первой строки на вторую значение элемента g_{22} не 6, как в матрице D, а 5, т. е. происходит сдвиг на $j-1$ позицию и соответственно уменьшение значения g_{22} на 1. Разница $d_{22} - g_{22} = 1$.

При переходе на третью строку в матрице G дополнительно происходит сдвиг еще на одну позицию относительно второй строки и на две позиции относительно первой строки. Уменьшение значения g_{33} относительно значения d_{33} равно

$$d_{33} - g_{33} = (j-1) + (j-2) = 2+1 = 3.$$

Нетрудно увидеть, что разница между d_{ij} и g_{ij}, где $i = j$, есть сумма арифметической прогрессии 0, 1, 2, 3, 4, ... Сумма этой прогрессии равна

$$S_j = ((j-1)^2 + j - 1)/2 = (j^2 - j)/2.$$

Следовательно, для того чтобы получить значение элемента матрицы g_{ij}, необходимо значение соответствующего элемента матрицы D уменьшить на величину S_j, т. е.

$$g_{ij} = n(j-1)+i-1-(j^2-j)/2 = (j-1)(2n-j)/2+i-1.$$

Значение g_{ij} есть искомое значение показателя степени k.

Для того чтобы значения скалярных показателей матриц различной размерности не оказались одинаковыми, необходимо поменять местами старшие и младшие разряды в бинарном ряду. Для этого младшему разряду необходимо присвоить степень $(n^2+n)/2-1$, тогда степени остальных разрядов равны

$$K = (n^2+n)/2 - 1 - k = (n^2+n)/2 - 1 - (j-1)\times \times(2n-j)/2+i-1 = (n^2+n-(j-1)(2n-j))/2+i-2.$$

Таким образом, получен показатель степени K из выражения (4).

Для матрицы A, соответствующей модели исследуемого объекта, а также для каждой матрицы смежности AI_k, из находящихся в базе данных моделей других объектов предметной области, можно вычислить скалярный показатель по формуле (4), идентифицирующий данную матрицу и соответствующий ей объект.

Различные объекты могут иметь одинаковые графы и соответственно одинаковые матрицы смежности и их скалярные показатели при различных весах дуг и вершин, т. е. при различных типах сущностей и отношений, которые определяются матрицами P и W соответственно для исследуемого объекта и PI_k, WI_k для известных объектов из базы данных. Для полной идентификации модели объекта необходимо провести идентификацию множества типов элементов и множества типов связей, образующих функциональный смежностный граф объекта.

Идентификацию рассматриваемых множеств возможно провести только в том случае, если эти множества упорядочены. Определить сходство двух смежностных графов можно в том случае, если нумерация вершин и дуг у обоих графов подчиняется одним и тем же правилам. Предлагается начинать нумерацию с вершины, имеющей минимальный вес, при этом дуги окажутся пронумерованы в зависимости от того, какие вершины они соединяют.

В качестве веса вершины или дуги выступает, как указывалось выше, числовой идентификатор данного признака в виде натурального числа, присвоенный ему в соответствующем словаре (базе данных) и являющийся для него ключом. Множество сущностей в многомерном графе (рис. 1) представлено в виде матрицы-столбца (5). Пример матрицы-столбца типов сущностей изображен в (6). Здесь числа являются ключами соответствующих сущностей.

$$P = \begin{vmatrix} p_1 \\ \dots \\ p_i \\ \dots \\ p_n \end{vmatrix} \quad (5) \qquad P' = \begin{vmatrix} 0034 \\ 0345 \\ 1567 \\ 2912 \end{vmatrix} \quad (6)$$

$$P'' = \left| \begin{array}{c} 01000100000 \\ 10011010100 \\ 11111000110 \\ 00000111101 \end{array} \right| \qquad (7)$$

Если представить каждый идентификатор в виде двоичного числа, то получится матрица (7), состоящая из нулей и единиц. Выше доказано, что каждой матрице, состоящей из нулей и единиц, размерности $n \times m$ соответствует уникальный скалярный показатель вида (4), значит и для матрицы-столбца P, элементы которой упорядочены по возрастанию весов, существует уникальный скалярный показатель. Обозначим его через S^P. Вычисляется этот показатель следующим образом.

Матрица-столбец P преобразуется в матрицу-столбец P'', элементами которой являются веса элементов в двоичном коде. При этом количество разрядов для представления каждого показателя элемента составляет x, а число разрядов показателя S^P соответственно nx. Число x получается из выражения $M \le \sum_{y=1}^{x} 2^{y-1}$, где M есть общее число типов элементов в данной предметной области. Каждый разряд есть определенная степень числа 2, отсюда первый разряд i-го элемента имеет степень числа 2, равную $x(i-1)$. Каждый элемент матрицы-столбца в двоичном коде в зависимости от положения на разрядной шкале величиной nx равен

$$p'_i = p_i \cdot 2^{x(i-1)}.$$

Следовательно, скалярный показатель S^P равен

$$S^P = \sum_{i=1}^{n} p_i \cdot 2^{x(i-1)}. \qquad (8)$$

Аналогично скалярный показатель для матрицы-столбца весов дуг S^W равен

$$S^W = \sum_{b=1}^{l} W_b \cdot 2^{f(b-1)}, \qquad (9)$$

где b — это индекс связи, присвоенный связи w_{ij}, а l — это количество связей в данном графе, f — число двоичных разрядов для обозначения связей получается из выражения $K \le \sum_{z=1}^{f} 2^{z-1}$, где K — число известных связей в предметной подобласти.

Для одного и того же объекта можно построить множество гомоморфных графов [1], задав при этом различную нумерацию вершин. В этом случае матрицы смежности для одного и того же объекта могут оказаться различными, а, следовательно, окажутся различными и скалярные показатели. Для того чтобы этого не произошло, первой назначается вершина, имеющая минимальный вес, далее нумерация осуществляется в порядке возрастания весов, т. е. упорядочивание происходит так, как упорядочиваются элементы матрицы-столбца P. Упорядочивание может осуществляться и по возрастанию весов при назначении первой вершины с максимальным весом.

В результате получен кортеж

$$S = < S^A, S^P, S^W >, \qquad (10)$$

элементы которого вычисляются по формулам (4), (8) и (9), однозначно идентифицирующий любой граф, а, следовательно, и любой объект, представимый с помощью графовой модели. Для каждого известного из данной предметной области и помещенного в базу данных объекта I_k вычислен аналогичный комплексный скалярный показатель

$$SI_k = < SI_k^A, SI_k^P, SI_k^W >, \qquad (11)$$

который также находится в базе данных вместе с самим объектом I_k.

Задача идентификации данного объекта сводится к задаче сравнения показателя S с показателями SI_k в базе данных предметной области. При помещении объекта в базу данных рассмотренный выше комплексный показатель SI_k, рассчитывается и помещается в базу данных. Соответствующая область БД содержит показатели SI_k и идентификаторы объектов, соответствующих этим показателям.

Дальнейшее базируется на основе кластерного анализа, представляющего собой статистический метод выделения во множестве элементов групп (кластеров) схожих между собой элементов на основе количественных или качественных измерений, выполняемых одновременно над несколькими переменными [4]. В решаемой задаче такими измерениями переменных являются вычисления скалярных показателей.

В рассматриваемой области создаются кластеры, соответствующие скалярному показателю SI_k^A, B k-м кластере находятся объекты, имеющие одинаковые значения SI^A, т. е. объекты, представленные идентичными графами смежности и различающиеся весами (типами) вершин (типами сущностей) и дуг (связей между сущностями внутри объекта).

При нахождении в базе SI^A, равного S^A, в кластере, соответствующем найденному SI^A, осуществляется поиск SI^P, равного S^P, на более низком иерархическом уровне. При положительном результате поиска происходит переход на еще более низкий иерархический уровень для проверки равенства значений S^W и SI^W. Структура области иерархической кластеризации показана на рис. 2.

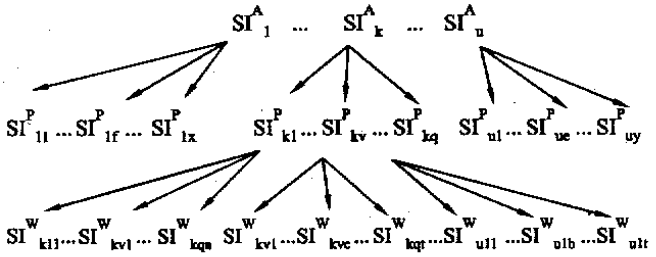

Рис. 2

Кластеры упорядочены по увеличению значений скалярных показателей. При помещении в базу данных нового объекта не происходит изменение ее логической структуры, упорядоченность элементов сохраняется без изменения. Так как кортежи SI уникальны, по определению, для каждого будущего нового объекта "зарезервировано" принадлежащее ему место.

В результате получен метод иерархической скалярной кластеризации множества объектов, представленных структурно-функциональными моделями, позволяющий идентифицировать структурно-функциональные модели кортежем скалярных показателей.

Преимуществом метода иерархической скалярной кластеризации является то, что при его использовании не надо перебирать записи в базе данных для поиска требуемого объекта, и исключается опасность комбинаторного взрыва. Данный метод позволяет войти непосредственно в требуемый кластер по его порядковому номеру, являющемуся значением соответствующего скалярного показателя. При отсутствии требуемого кластера он создается, и это свидетельствует, что искомый объект в базе данных отсутствует, а, следовательно, исследуемый объект является уникальным.

При практическом использовании размерность скалярных показателей может оказаться достаточно большой. В этом случае вместо каждого из трех частных показателей можно создавать иерархические структуры, элементы которых будут иметь меньшую размерность. При этом при поиске требуемого объекта в базе данных число шагов увеличивается, однако, суть метода и его детерминизм не изменяются. Число шагов для обнаружения искомого объекта всегда известно и остается постоянным для данной предметной области и используемого типа базы данных.

Метод иерархической скалярной кластеризации может быть использован в патентных исследованиях при создании новой техники, а также при каталогизации различных объектов в соответствующих предметных областях.

СПИСОК ЛИТЕРАТУРЫ

1. Оре О. Теория графов.— М.: Главная редакция физико-математической литературы, 1980.

2. Арешев Т. А. К вопросу о представлении технических решений в формализованном виде // Патентные проблемы вычислительной техники и кибернетики.— Л.: ЛНИВЦ АН СССР, 1985.

3. Разработка тематических АИПС на основе структурно-функционального анализа: Метод. рекомендации.— М.: ВНИИПИ, 1991.

4. Толковый словарь по вычислительным системам.— М.: Машиностроение, 1989.

Материал поступил в редакцию 13.06.91.