ИСКУССТВЕННЫЙ ИНТЕЛЛЕКТ И ЭКСПЕРТНЫЕ СИСТЕМЫ
(Вычислительные системы)
1996 год Выпуск 157
УДК 519.95

СРАВНЕНИЕ ИЕРАРХИЧЕСКИХ СТРУКТУР

Н.Г.Загоруйко, А.Г.Пичуева

Введение

В последне время в области Анализа Данных отмечается рост интереса к анализу так называемых "символьных объектов" [1], с помощью которых описываются разного рода обобщенные характеристики некоторого массива исходных данных. Символьным объектом может быть обнаруженная в этом массиве логическая закономерность типа "Если ..., То ...", направленный граф, отражающий зависимость одних объектов от других и т.п. В частности, результаты иерархической таксономии выявляют структуру множества объектов, которую можно наглядно представить графически в виде иерархического дерева, начальные вершины ("листья") которого отображают все объекты исходного множества, промежуточные вершины ("ветви") описывают все более крупные таксоны, а конечная вершина ("корень") представляет собой объединение всего исходного множества объектов в один таксон. При изучении нескольких различных массивов данных может потребоваться сравнение между собой внутренней структуры этих массивов, что приводит к необходимости измерять степень "близости", "похожести" между иерархическими описаниями этих структур.

В работах [2,3] описаны методы анализа символьных объектов, имеющих форму конъюнкций типа "Если ...,

То ...". Данная работа посвящена введению меры близости или расстояния на множестве символьных объектов типа иерархий.

1. Иерархия

Определим понятие "иерархия". Обозначим через W конечное множество объектов, $W = \{w_1, w_2, \ldots, w_l, \ldots, w_q\}$, а через H — множество непустых частей множества W, называемых таксонами и обозначаемых через h. Теперь воспользуемся определением иерархии, данным в [1].

Иерархией H множества W называется множество подмножеств W таких, что:

1) $\forall\, w \in W \{w\} \in H$ (терминальные вершины ("листья") — одноэлементные множества);

2) $W \in H$ (наибольший таксон ("корень") содержит все элементы W);

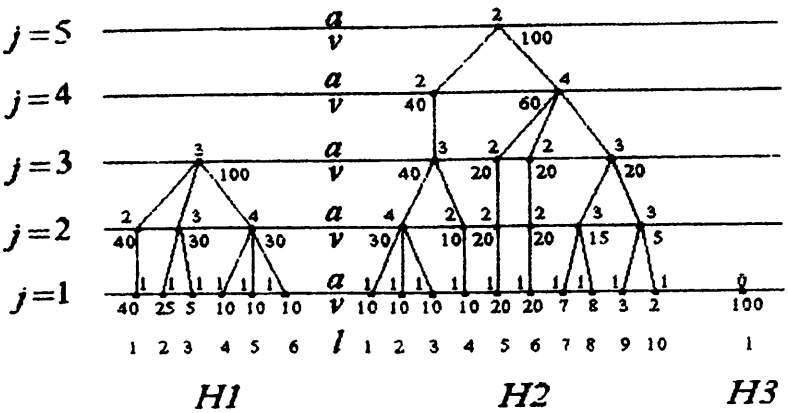
3) для любых вершин $h, h' \in H$ мы имеем либо $h \cap h' = 0$, либо $h \subset h'$, либо $h' \subset h$.

Таким образом, иерархия — это многоуровневая структура, в которой объекты, находящиеся в одном таксоне на некотором (j-м) уровне, остаются в одном таксоне на $j+1$-м и всех других более высоких уровнях. Первому уровню соответствуют терминальные вершины (см.п. 1 в определении иерархии), а последнему, максимальному, уровню (обозначим его m) — наибольший таксон, содержащий все элементы W, этот таксон можно обозначить тем же знаком W (см. п. 2 в определении иерархии). На каждом уровне происходит или не происходит объединение таксонов (см.п. 3 в определении иерархии).

Обозначим точкой каждый таксон иерархии. Тогда отрезки, соединяющие эти точки (или вершины иерархии), передают порядок образования таксонов, который отвечает пп.1–3 определения. На рисунке показана, например, иерархия H_2 с десятью терминальными вершинами $\{w_k\}$, $k = 1, 2, \ldots, l, \ldots, q$, $q = 10$, первого уровня, шестью таксонами на втором уровне, четырьмя таксонами на уровне $j = 3$, двумя таксонами на уровне $j = 4$ и

одним таксоном на верхнем пятом уровне ($j = m_1 = 5$). Иерархия H_1 содержит три уровня с шестью, тремя и одной вершиной на уровнях $j = 1$, 2 и 3 соответственно.

Можно представить себе и вырожденный случай: иерархию, свернутую в одну изолированную вершину на первом уровне (иерархия H_3 на рисунке).



Каждая вершина (jlt) графа H_t может характеризоваться структурным индексом $a(jlt)$, равным числу примыкающих к ней ребер. В нашем примере для всех терминальных вершин иерархий H_1 и H_2 $a(1lt) = 1$, индексы их других вершин принимают значения от 2 до 4: $a(212) = 4$, $a(311) = 3$ и т.д., а индекс $(113) = 0$.

Терминальные вершины могут быть равнозначными, но могут и отличаться по "значимости", по "насыщенности", по "весу". Так может быть, например, когда листья представляют собой таксоны с разным числом входящих в них объектов. Весовой индекс вершины будем обозначать через $v(jlt)$. Значения весовых индексов терминальных вершин, нормированных так, чтобы их сумма на каждом иерархическом уровне равнялась 100, указаны на рисунке в строке v. Весовой индекс нетерминальных

вершин уровня j равен сумме весовых индексов вершин, входящих в эту вершину из предыдущего уровня $j-1$.

2. Расстояние между иерархиями

Как определить расстояние между подобными иерархическими структурами? В работах [4–6] предлагается мера близости между такими графами с поименованными вершинами, списки вершин в которых совпадают или мало различаются. Здесь мы попытаемся предложить решение проблемы измерения расстояний между иерархиями с объектами произвольного состава.

Естественным образом возникает идея оценить расстояния между иерархиями через сложность превращения одной иерархии в другую, добавляя или убирая вершины и связи между ними, где это необходимо, т.е. применяя набор так называемых "редакционных операций". Каждая операция имеет свою стоимость (c). Оптимальному переводу соответствует последовательность элементарных операций с минимальной суммарной стоимостью, которая носит название "редакционного расстояния" [7]. Связанную с ним переменную d — качественную характеристику расстояния или различия во внешнем виде двух иерархических структур — назовем "расстоянием по виду структур". С другой стороны, неплохо было бы учитывать и вес элементов, "собираемых" в таксоны на каждом уровне иерархий. Связанную с этим переменную — количественную характеристику различия по "насыщенности" или "весу" таксонов двух иерархических структур — будем обозначать символом r.

Перейдем к математической постановке задачи нахождения характеристик расстояний d и r.

3. Расстояние по виду структуры

Пусть нам даны две иерархии, H_1 и H_2, с числом уровней m_1 и m_2 соответственно. Будем рассматривать иерархию H_t как (упорядоченное) множество уровней с

$j = 1$ по $j = m_t$, а каждый уровень — как совокупность $h(jt)$ расположенных на нем q вершин (таксонов) $h(jlt)$:

$$h(jt) = \{h(j1t), h(j2t), \ldots, h(jlt), \ldots, h(jqt)\}.$$

В процессе превращения одной иерархии в другую потребуется вершину $h(jl1)$ заменить на вершину $h(jl2)$. Будем считать, что стоимость такой редакционной операции равна $c(jl1, jl2) = |a(jl1) - a(jl2)|$. Для оценки стоимости замен всех q_1 вершин уровня j_1 первой иерархии на все q_2 вершины уровня j_2 второй иерархии будем пользоваться простым алгоритмом "похожих пар". Для этого вначале сделаем число вершин в сравниваемых уровнях одинаковым и равным q, добавив к уровню с меньшим числом вершин f "пустых" вершин, где $f = |q_1 - q_2|$. "Пустой" будем называть вершину с индексом $\{a(jlt) = 0\}$.

Затем для вершины $h(jl1)$ находится самая похожая на нее вершина $h(jl2)$, т.е. такая, редакционное растояние $c(1)$ до которой минимально. Величина $c(1)$ добавляется в счетчик суммарного расстояния $c(j_1, j_2)$ между данными уровнями и эта пара самых похожих вершин из дальнейшего рассмотрения исключается. Среди оставшихся вершин снова ищется самая похожая пара, величина их редакционного расстояния $c(2)$ добавляется к счетчику $c(j_1, j_2)$, а эта пара так же исключается из дальнейшего анализа. Такая процедура нахождения на каждом (l-м) шагу самой похожей пары, добавления к счетчику $c(j_1, j_2)$ величины $c(l)$ и исключение l-той пары выполняется q раз. В итоге получается величина редакционного расстояния между двумя уровнями: $c(j_1, j_2) = \sum_{l=1}^{q} c(l)$.

Проведя сравнение всех m_1 уровней первой иерархии со всеми m_2 уровнями второй, мы получаем матрицу $C(1,2)$ с номерами строк $1, 2, \ldots, j_1, \ldots, m_1$ и номерами столбцов $1, 2, \ldots, j_2, \ldots, m_2$. На пересечении строки j_1 и столбца j_2 будут стоять величины (частных) редакционных расстояний $c(j_1, j_2)$ между уровнями j_1 и j_2 сравниваемых иерархий (см. табл. 1).

Таблица 1

j_1 \ j_2	1	2	3	4	5
1	(4)	10	8	8	6
2	13	(7)	(3)	3	7
3	11	13	7	(3)	(1)

Редакционным расстоянием d между иерархиями H_1 и H_2 будем называть стоимость не любого, а оптимального перевода уровней иерархии H_1 в соответствующие уровни иерархии H_2. Этот перевод будем искать с помощью метода динамического программирования [7,8]. В результате будет найден путь на матрице $C(1,2)$, соединяющий клеточку (1,1) с клеточкой (m_1, m_2) и проходящий через соседние клеточки либо по горизонтали слева направо, либо по вертикали сверху вниз, либо по диагонали вправо-вниз. На каждом шаге будем прибавлять к счетчику расстояния $d(Q)$ величину $k*c(j_1, j_2)$, взятую из той клеточки, через которую проходит путь. Весовой коэффициент k равен 1 при переходе по диагонали и 2 при переходе по горизонтали или вертикали (схема динамического программирования "2-1-2"). Наша цель состоит в поиске такого пути Q, который набирает минимальную сумму $d(Q)$ стоимостей частных взвешенных редакционных расстояний. Этот путь показан в табл.1 цепочкой выделенных элементов. Он дает величину $d(Q) = 22$.

Для нормировки редакционного расстояния $d(Q)$ в пределы от 0 до 1 нужно $d(Q)$ разделить на "коэффициент нормализации" D, который представляет собой наибольшее редакционное расстояние от иерархии H_1 и H_2 до некоторой предельно отличающейся от них иерархии H_3. В качестве таковой принимается иерархия H_3 в виде одной вершины, находящейся на первом уровне и имеющей индекс $a(113) = 0$.

Матрица частных редакционных расстояний для сочетания всех уровней иерархий H_1 и H_3 приведена в

табл.2а, а для иерархий H_2 и H_3 — в табл.2б.

Таблица 2а

j_3 \ j_1	1	2	3
1	(6)	(9)	(3)

Таблица 2б

j_3 \ j_2	1	2	3	4	5
1	(10)	(16)	(10)	(6)	(2)

Оптимальные пути здесь идут только по горизонтали, так что редакционные расстояния между H_t, $t = 1, 2$, и H_3 будут равны $D_t = \Sigma a(1lt)+2*\Sigma a(jlt)$, где $l = 1, \ldots, q_t$, $j = 2, \ldots, m_t$. В нашем примере расстояние $D_1 = 6+2*(9+3) = 30$, а расстояние $D_2 = 10+2*(16+10+6+2) = 78$. Следовательно, в качестве нормирующего коэффициента выбирается $D = 78$, и редакционное расстояние между иерархиями H_1 и H_2 по виду структуры равно $d = \dfrac{d(Q)}{D} = \dfrac{22}{78} = 0.282$.

4. Расстояние по весовым индексам таксонов

Теперь опишем процесс нахождения другой характеристики расстояния (r) между иерархиями по весовым индексам входящих в их состав таксонов. Здесь также будем применять метод динамического программирования, так как идея состоит в том же самом желании оптимально преобразовать все уровни одной иерархии в соответствующие уровни другой. Для оценки редакционных расстояний между уровнем j_1 первой иерархии и уровнем j_2 второй воспользуемся описанным выше алгоритмом "похожих пар". Если число таксонов в данных уровнях неодинаково, т.е. если $q(1) \neq q(2)$, то устраняем этот

"дефект" путем добавления к уровню с меньшим числом таксонов недостающего числа $f = |q(1) - q(2)|$ таксонов с нулевым весом $v = 0$. После этого находятся самые похожие пары вершин (таксонов) сравниваемых уровней и частные редакционные расстояния между этими вершинами суммируются в накопитель редакционного расстояния между рассматриваемыми уровнями: $c(j_1, j_2) = \sum_{l=1}^{q} |v(jl1) - v(jl2)|$.

Как и в предыдущем случае, формируем матрицу (см. табл.3) редакционных расстояний размером j_1 на j_2 и ищем на ней оптимальный путь Q перевода одной иерархии в другую. Применяем такую же схему динамического программирования "2-1-2" и находим величину редакционного расстояния $r(Q)$ (в нашем примере оптимальный путь показан в табл.3 и $r(Q) = 230$).

Таблица 3

j_1 \ j_2	1	2	3	4	5
1	(50)	(30)	40	70	120
2	100	60	(40)	60	120
3	160	140	120	(80)	(0)

Наибольшая величина расстояния R была бы найдена при сравнении заданных иерархий H_1 и H_2 с наиболее на них непохожей иерархией. В качестве таковой выбираем иерархию H_3, представленную одной вершиной первого уровня с весом $v(113) = 100$.

Таблица 4а

j_3 \ j_2	1	2	3
1	(120)	(120)	(0)

Таблица 46

j_3 \ j_1	1	2	3	4	5
1	(160)	(140)	(120)	(80)	(0)

В нашем примере $R(1,3)=360$ а $R(2,3)=840$ (см. табл. 4а и 4б), так что редакционное расстояние между H_1 и H_2 по насыщенности таксонов равно $r=\frac{r(Q)}{R}=\frac{230}{840}=$ $=0,274$.

Общее редакционное расстояние P между двумя иерархиями примем равной средней величине расстояний d и r: $P=\frac{(d+r)}{2}$ и в нашем случае $P=\frac{0,282+0,274}{2}=$ $=0.278$.

5. Расстояние между множествами иерархий

Перейдем от меры P между двумя иерархиями к мере расстояния S между двумя множествами иерархий.

Если количества иерархий в этих множествах не одинаковы, то к меньшему множеству добавляется недостающее число "пустых" иерархий. "Пустой" будем называть иерархию, состоящую из одной вершины первого уровня с нулевым числом входящих в нее ребер $\{a(11)=0\}$ и с нулевым индексом насыщенности $\{v(11)=0\}$. В результате в каждом множестве будет одинаковое число q иерархий.

Если известно расстояние между любыми двумя иерархиями, принадлежащими различным множествам, то можно "перевести" все иерархии одного набора в соответствующие иерархии другого набора. Оптимальному переводу будет соответствовать вариант, при котором сумма требующихся для этого редакционных затрат минимальна.

В том случае, когда множества иерархий не структурированы, все входящие в состав этих множеств иерархии могут рассматриваться независимо друг от друга и

минимум суммарных редакционных затрат будет достигаться при использовании описанного выше алгоритма "похожих пар". В счетчик общего расстояния S' между множествами суммируются частные редакционные расстояния $R(ll)$ между похожими парами иерархий.

Нетрудно убедиться, что при сравнении двух иерархий, одна из которых (H) имеет m уровней с $q(j)$ вершинами на j-х уровнях и индексами в вершинах $a(j,l)$ и $v(j,l)$, $j=1,\ldots,m$, $l=1,\ldots,q(j)$, а вторая (H') является пустой, всегда будут иметь место равенства $r=\Sigma a(1,l)+$ $+2\Sigma a(j,l)$, $j=2,\ldots,m$, $l=1,\ldots,q(j)$; $d=100+2*100(m-1)=$ $=100*(2m-1)$.

Обратим внимание на то, что величины $R(ll)$ являются нормированными, так что дополнительная нормировка величины S' состоит в делении ее на число q иерархий в каждом из сравниваемых множеств.

В итоге получается величина редакционного расстояния между двумя неструктурированными множествами иерархий $S=\dfrac{S'}{q}=\dfrac{1}{q}*\sum\limits_{l=1}^{q}R(ll)$.

Если множества иерархий организованы в иерархические структуры, то расстояние между такими структурами определяется описанным выше методом оценки расстояния между иерархиями, так как содержание объектов, образующих иерархию, нами не учитываются.

Если же нужно учесть и характер объектов, то речь в данном случае должна идти о расстоянии между метаиерархиями, на которое влияет не только различие в архитектуре, но также и различия в структурах иерархий, входящих в сравниваемые вершины метаиерархий. Архитектурные различия можно учесть, используя индексы $a(jlt)$, а для оценки различия по составу вершин потребуется вместо весового индекса $v(jlt)$ использовать некоторую величину, в которой отражались бы различия между наборами иерархий в сравниваемых вершинах метаиерархий.

Литература

1. DIDAY E. Symbolic Data Analysis. — INRIA — Roquencourt, Paris, 1995. — P. 1–136.

2. ЗАГОРУЙКО Н.Г. Анализ данных и анализ знаний //Анализ последовательностей и таблиц данных. — Новосибирск, 1994. — Вып. 150: Вычислительные системы. — С. 3–17.

3. ЗАГОРУЙКО Н.Г. Редактирование Баз Знаний //Настоящий сборник. — С.3 11.

4. BOGART K.P. Preference structure. I //J.Math.Sociol. 1973. — Vol.3. — P. 49–67.

5. BOGART K.P. Preference structure.II //SIAM J.appl.math. — 1975. — Vol.29, №2. — P.254–262.

6. РАППОПОРТ А.М., ШНЕЙДЕРМАН М.В. Анализ экспертных суждений заданных в виде структур //Прикладной многомерный статистический анализ. — М.: Наука, 1978. — С. 150–164.

7. НЕМЫТИКОВА Л.А. Методы сравнения символьных последовательностей //Методы обработки символьных последовательностей и сигналов. — Новосибирск, 1989. — Вып. 132: Вычислительные системы. — С.1–34.

8. ВЕЛИЧКО В.М., ЗАГОРУЙКО Н.Г. Автоматическое распознавание ограниченного набора устных команд. — Вычислительные системы. Вып.36. — Новосибирск, 1969. — С. 101–110.

9. ВЕНТЦЕЛЬ Е.С. Элементы динамического программирования. — М.: Наука, 1964.

Поступила в редакцию
12 ноября 1996 года