Математика. Компьютер. Образование. Вып. 7. Москва 2000.

ВЕКТОРНЫЕ МЕРЫ СЛОЖНОСТИ, ЭНТРОПИИ, ИНФОРМАЦИИ.

А. В. Коганов

Научно-исследовательский институт Системных исследований Российской Академии Наук (НИИСИ РАН). Россия, 117218, Москва, Нахимовский пр., д.36, корп. 1. При содействии Российского Фонда Фундаментальных Исследований (РФФИ), грант 98-01-00798.
т. (095)332-4818, (095)143-2370, факс (095)719-7651 , mail: koganow@niisi.msk.ru .

1. Постановка задачи.

Проникновение информационных технологий и связанного с ними математического описания объектов во многие области науки и практики выявило недостаточность существующих определений меры сложности, энтропии и информации. Понятие сложности было разработано рядом авторов для конструктивной теории функций, причем характер этих определений не позволяет расширить область их применения [1][2]. Понятие информации, введенное в рамках теории кодирования К. Шенноном, фактически пригодно только для анализа загрузки каналов связи при передаче символьных последовательностей и совершенно не отражает содержание передаваемых сообщений [6]. Формальное применение такой меры информации может приводить к парадоксальным результатам. Например, если получено сообщение, увеличивающее множество возможных (ожидаемых в моделируемой ситуации) альтернатив, мера информации отрицательна (чем больше новых вариантов, тем меньше мера полученной информации), а при изменении множества альтернатив с сохранением их числа информация нулевая независимо от ценности этих сведений. В этой работе автор делает попытку дать достаточно общее комбинаторное определение мер сложности, энтропии и информации сообщения, применимое для описания произвольных конечных структур математических моделей (без использования вероятностных описаний [6]), соответствующее интуитивным представлениям.

Понятие сложности предполагает выбор числовых характеристик в соответствие с требованиями задачи, где используется модель. Одна модель может иметь много неэквивалентных мер сложности. Главная задача введения такой меры - это ранжирование математических моделей по потреблению тех или иных ресурсов при их нормативном прикладном использовании. Несколько подобных числовых характеристик могут образовать вектор сложности, отражающий разные аспекты потребления ресурсов. Скалярную меру сложности можно получить, задавая специально подобранную норму такого вектора с учетом значимости каждого ресурса.

Понятие информации в содержательном аспекте означает *сообщение*, содержащее совокупность сведений об изменении структуры или параметров математической модели. Термин “*Информация*” означает только числовую меру этих изменений. Если для модели введена числовая или векторная мера сложности, то мера информации должна позволять вычислить новое значение

вектора сложности для измененной модели. Поэтому, в общем случае информация также должна быть вектором. Норму вектора можно рассматривать, как общую скалярную меру информации, от которой естественно потребовать монотонности по каждой из компонент вектора.

Кроме понятия сложности и информации, модель характеризуется разнообразием возможного выбора вариантов ее состояния. Эта характеристика традиционно называется комбинаторной энтропией модели. Обычно математическая модель допускает несколько определений понятия "*состояние*", и каждому определению будет соответствовать своя мера энтропии. Содержательно, состояние - это возможный режим или способ использования модели. Следуя Больцману, комбинаторная энтропия определяется, как логарифм числа состояний модели, и именно так энтропийная мера модели зависит от определения состояния.

Такой подход исключает применение мер сложности и информации к объектам, не имеющим математической модели, или к моделям, для которых не определены потребляемые ресурсы. К этой категории относятся произведения искусства или чисто теоретические конструкции. Однако, как только возникает вопрос о восприятии этих объектов человеком, сразу возникают ограничения по ресурсам времени, внимания, чувствительности, памяти и т. д. и тогда мера сложности становится уместной. Однако для ее введения требуется построить математическую модель потребления этих ресурсов.

В данной работе строится модель использования ресурсов памяти и адресного пространства компьютеров для записи широкого класса математических конструкций, и вводится соответствующее понятие информации. Для того же класса моделей вводится мера комбинаторной энтропии нескольких типов. Все меры имеют векторную форму, и между ними имеется интересная связь.

2. Формализация постановки задачи.

В дальнейшем будет использоваться следующая терминология. "*Состав модели*" - набор структурных элементов модели (множества, их элементы, отображения и отношения), по которым определяется "*сложность*", как векторная мера состава $X=(x_1,...,x_n)$. "*Сообщение*" - описание изменений в модели, данное в терминах ее состава. Фактически, это математическая модель изменений. Тип возможных сообщений входит в описание модели. Относительно этих сообщений вводится векторная мера $Y=(y_1,...,y_m)$ ("информация"), позволяющая вычислить новую сложность измененной модели, используя прежний вектор сложности.

$$(x'_1,...,x'_n)=F(x_1,...,x_n;\ y_1,...,y_m) \qquad (1)$$

"*Состояние*" - комбинация элементов состава модели, определяющая некоторый режим ее работы. "*Энтропия*" - логарифмическая мера S числа N определенных для модели состояний.

$$S=\log N\ ;\ N=|\ STATE\ SET\ |\ ; \qquad (2)$$

Аргументы в пользу такой меры энтропии те же, что в классических работах Больцмана, Шеннона, Колмогорова. Главное характеристическое свойство заключается в том, что энтропия для совокупности двух взаимно независимых моделей равна сумме их энтропий.

Универсальным языком для описания моделей выберем отношение индукции ***I*** на множестве ***T*** элементарных объектов ("*точек*"), которое определяет для каждой точки ***t*** (как для "*центра индукции"*) некоторую систему ***I(t)*** подмножеств ("*индукторов*" этой точки) из того же множества ***T***. Если подмножество ***V*** - индуктор точки ***t*** ,то будем обозначать это ***V~t*** или ***V~t/I*** , если надо указать конкретное отношение индукции. Множество точек, для которых система индукторов не пуста, обозначим ***T/I*** , а дополнение ***T/0*** . В [3][4] показано, что такие отношения позволяют описывать практически любые процессы и группы преобразований. Отношения индукции, удовлетворяющие некоторым дополнительным требованиям (индукторные пространства), обобщая понятия направленного графа и топологии, позволяют моделировать объекты в форме активной среды, где между элементами передаются сигналы.

Далее будет предполагаться, что модель описана в форме конечного отношения индукции, которое будем называть "*И-структурой"*. Для других способов описания моделей ниже приведенные формулы сложности, энтропии, информации не пригодны, однако сам принцип их построения, изложенный в постановочной части статьи, остается в силе.

Для индукторных моделей введем несколько определений множества состояний.

Состояние типа "*простая связь*" определено как одна цепочка

$$(t; V;g),\ t \in T,\ V \in I(t),\ g \in V; \qquad (3)$$

Состояние типа "*потенциальная связь*" определено, как совокупность элементов, выбранных по одному из каждого индуктора И-структуры, совокупность индукторов, выбранных по одному для каждого центра индукции, и одной точки, выбранной из всех центров индукции. Такое состояние для каждого изменения центра индукции уже содержит готовую простую связь.

$$(t; \{V|\ V \in I(z),\ z \in T\}; \{g|\ g \in V,\ V \in I(z),\ z \in T\}),\ t \in T; \qquad (4)$$

Состояние типа "*мультиграф*" образуют три подтипа (tvG), (tWG), (QWG), в каждом из которых потенциальная связь на некоторых уровнях образует не элемент, а подмножество соответствующего вида. Эти состояния описывают для каждого центра или подмножества центров множество простых связей.

$$(t; \{V|\ V \in I(z),\ z \in T\}; \{G|\ G \subset V,\ V \in I(z),\ z \in T\}),\ t \in T; - (tvG).$$

$$(t; \{W|\ W \subset I(z),\ z \in T\}; \{G|\ G \subset V,\ V \in I(z),\ z \in T\}),\ t \in T; - (tWG).$$

$$(Q; \{W|\ W \subset I(z),\ z \in T\}; \{G|\ G \subset V,\ V \in I(z),\ z \in T\}),\ Q \subset T; -(QWG). \qquad (5)$$

Относительно этих пяти типов состояний будут вычисляться варианты энтропии модели.

3. Оценка сложности модели и информации сообщения.

Сложность модели будем оценивать относительно ресурса памяти, необходимой для ее записи. Вообще говоря, природа памяти несущественна, но классифицируя ее на память кодов и память адресов, мы неявно предполагаем компьютерную систему хранения данных. Это достаточно актуально для современного способа математического моделирования.

Коды точек требуют длины кода $S(T)=\log(|T|)$бит. Адресация индуктора для данного центра индукции требует $L(I(t))=|I(t)|\log(|I(t)|)$ бит с длиной одного адреса $S(I(t))$. Адресация точек индуктора требует $L(V)$ бит с длиной одного адреса $S(V)$. Для общности обозначений будем считать $S(0)$ символом, удовлетворяющим условию $2^{S(0)}=0;\quad L(0)=0$. Тогда общий объем адресной памяти:

$$Amem(T,I)=L(T/I)+sum\{L(I(t))+sum\{L(V)|V\ in\ I(t)\}|t\ in\ T\}; \qquad (6)$$

Необходимая длина адресного регистра:

$$Reg(T,I)=S(T/I)+max\{S(I(t))|t\ in\ T\}+max\{S(V)|V\ in\ I(t),t\ in\ T\} \qquad (7)$$

С учетом запоминания кода точки по каждому адресу в индукторе общая память составляет:

$$Mem(T,I)=L(T/I)+sum\{L(I(t))+ sum\{|V|(S(V)+S(T))|V\ in\ I(t)\}|t\ in\ T\}=$$
$$=Amem(T,I)+S(T)sum\{|V|\ |\ V\ in\ I(t)\},\ t\ in\ T\}; \qquad (8)$$

В качестве вектора сложности можно принять

$$Comp(T,I)=(S(T),S(T/I),\ Amem,Reg,Mem) \qquad (9)$$

Для введения вектора информации потребуются две величины:

$$Z(T,I)= sum\{L(I(t))+sum\{L(V)|V\ in\ I(t)\}|t\ in\ T\};$$
$$R(T,I)=sum\{\ |V|\ \ |\ V\ in\ I(t),\ t\ in\ T\ \}. \qquad (10)$$

Сообщение об изменении И-структуры содержит три множества: множество M точек, отбрасываемых вместе со своими индукторами, множество W новых точек, и новые индукторы этих точек, образующие отношение индукции B на новом множестве точек. Точка считается отброшенной, если изменился хотя бы один ее индуктор. Индуктор считается удаленным, если удален его центр. По множеству M однозначно определяется множество удаленных индукторов A. Тогда в качестве меры информации сообщения можно принять следующий десятимерный вектор:

$$Inf(M;\ W,B)=(\ S(M/I),\ S(M/0),\ Reg(T\text{-}M,\ I\text{-}A),Z(A),R(A);$$
$$S(W/B),S(W/0),\ Reg(W,B),\ Z(B),R(B)\ ). \qquad (11)$$

По этим данным можно вычислить новое значение ввектора сложности.

$$|T_{new}| = 2^{S(T)} - 2^{S(M/I)} - 2^{S(M/0)} + 2^{S(W/B)} + 2^{S(W/0)};$$
$$|T_{new}/I| = 2^{S(T/I)} - 2^{S(M/I)} + 2^{S(W/B)};$$
$$Amem(new) = Amem(T,I) - L(T/I) + L(T_{new}/I) - Z(A) + Z(B);$$
$$Reg(new) = max\{Reg(T-M, I-A); Reg(W,B)\};$$
$$R(I) = (Mem(T,I) - Amem(T,I))/S(t);$$
$$Mem(new) = Amem(new) - R(I)S(T) + (R(I) - R(A) + R(B))S(T_{new}); \quad (12)$$

4. Векторная энтропия модели.

Для состояния типа простая связь энтропия определяется, как логарифм числа простых связей (скаляр). Для потенциальных связей естественно разделить энтропию центров индукции и энтропию индукторов. Аналогично выглядят и двухкомпонентные энтропии мультиграфа.

$$S_{sim}(T,I) = log(R(I)); \quad (13)$$
$$S_{pot}(T,I) = (S(T), sum\{S(I(t)) + sum\{S(V(t)) | V \text{ in } I(t)\} | t \text{ in } T/I\}); \quad (14)$$
$$S_{tVG}(T,I) = (S(T), sum\{S(I(t)) + sum\{|V(t)| \; | V \text{ in } I(t)\} | t \text{ in } T/I\}) =$$
$$= (S(T), sum\{S(I(t))| \; t \text{ in } T/I\} + R(I)); \quad (15)$$
$$S_{tWG}(T,I) = (S(T), sum\{|I(t)| + sum\{|V(t)| \; | V \text{ in } I(t)\} | t \text{ in } T/I\}) =$$
$$= (S(T), sum\{|I(t)| \; | \; t \text{ in } T/I\} + R(I)); \quad (16)$$
$$S_{QWG}(T,I) = (|T|, sum\{|I(t)| + sum\{|V(t)| \; | V \text{ in } I(t)\} | t \text{ in } T/I\}) =$$
$$= (|T|, sum\{|I(t)| \; | \; t \text{ in } T/I\} + R(I)); \quad (17)$$

Можно отметить интересную аналогичность, но не тождественность, выражений для сложности и энтропии. Для каждого типа энтропии также можно ввести свой вектор информации сообщения. Заметим, что векторы информации, введенные для разных векторных оценок сложности или энтропии, моно объединять в расширенный вектор информации, пригодный для всех этих оценок. Например, для потенциальных связей вектор информации и оператор преобразования сложности имеют вид

$$Inf_{pot}(M;W;B) = (S(M), P(A); S(W), P(B)) \;\; where$$
$$P(I) = sum\{S(I(t)) + sum\{S(V) | V \text{ in } I(t)\} | t \text{ in } T\}; \quad (18)$$
$$S(T_{new}) = log(2^{S(T)} - 2^{S(M)} + 2^{S(W)}); \;\; P(I_{new}) = P(I) - P(A) + P(B);$$

5. Последовательности сообщений.

В рассмотренном выше подходе класс возможных сообщений должен быть строго определен вместе с типом математической модели. Вследствие этого, некоторая последовательность сообщений может не образовывать нового сообщения. Это происходит, если очередное сообщение оказывается не соответствующим той модели, в которую преобразовали исходную модель предыдущие сообщения. Например, это сообщение может потребовать устранить уже несуществующий в модели фрагмент, или добавить фрагмент, для согласования с которым требуются удаленные ранее детали. Поэтому сообщения обладают лишь частичной способностью к образованию допустимых цепочек, и это требование принципиально для любой

формализации содержательной информации. Договоренности типа «игнорировать бессмысленные сообщения в цепи» являются чисто формальным расширением теории, затрудняющим ее интерпретацию. В тех случаях, когда сообщения образуют смысловую цепочку, по ним можно сформулировать одно эквивалентное сообщение, сразу преобразующее исходную модель к окончательному виду.

Рассмотрим это на примере индукторных структур. Сообщения в этом случае имеют вид, указанный в аргументе информации формулы (11): ***(M; W,B).*** Пусть имеются два последовательных совместимых сообщения ***(M_1; W_1,B_1), (M_2; W_2,B_2)***. Причем каждая ***M***-компонента позволяет определить по модели свою индукцию ***A.*** Тогда эквивалентное их последовательному применению единое сообщение имеет следующий вид (все операции над множествами).

$$W= W_1 -M_2 +W_2 ;\ \ B= B_1 -A_2+ B_2 ;\ M= M_1 + TM_2 ; \qquad (19)$$

6. Интерпретация в статистической физике.

Модель взаимодействия в некоторой совокупности частиц можно построить в форме отношения индукции. В этом случае энтропия модели приобретает смысл статистической энтропии взаимодействия, и, варьируя модели, можно анализировать влияние изменения структуры коллектива на энтропию.

Моделью коллектива из ***N*** частиц, свободно вступающих в парные взаимодействия, является И-отношение ***T, |T|=N, t~T for all t in T***. Тогда энтропия парного взаимодействия равна $S_{sim}=2\log N$, а энтропия коллективного взаимодействия $S_{tVG}=N^2$. Рассматриваются вторые компоненты вектора энтропии.

Если коллектив разделен на несколько равных частей, между которыми нет взаимодействий, то И-отношение имеет вид

$$T=T_1+...+T_m;\ |T|=N;\ |T_i|=N/m;\ t\sim T_i,\ t \text{ in } T_i,\ i=1,...,m. \qquad (20)$$

$$(S_{sim})_2=2\log(N)-\log(m);\quad (S_{tVG})_2=N^2/m\ . \qquad (21)$$

Таким образом, разбиение коллектива частиц на слабо взаимодействующие фрагменты значительно снижает оба вида энтропии. Это объясняет тенденцию в биологической эволюции к фрагментации организмов и сообществ (клеточные структуры, особи, стаи). В социуме по той же причине возникает разделение общества по территориям, специальностям и т. п. На большую роль структурирования в процессах, изменяющих энтропию, указано в [5][6].Рост энтропии характерен только в тех процессах, где разрушаются границы взаимодействия. Построение и поддержание границ требует затрат энергии, вследствие чего в изоэнергетических процессах энтропия не убывает. Однако для открытых процессов характерно двунаправленное изменение энтропии. Например, при разветвлении русла реки энтропия в сечении потока падает, а при слиянии рукавов в одно русло - возрастает.

7. Случай описательных моделей.

Для случая баз данных, когда модель представляет собой простое описание, т. е. набор элементов-текстов с самоиндукцией (связь только на себя), сложность имеет вид

$$Comp(T,I)=(S(T),\ S(T),L(T),\ S(T),\ 2L(T)\ ) \qquad (22)$$

и не зависит от размера составляющих текстов. Это связано с предположением оптимального кодирования элементов модели, сделанном при оценке общей памяти модели. Для реальных баз данных это предположение, как правило, не выполнено. В этом случае мера сложности не отражает семантику, связывающую сложность с длиной текстов. Ее можно интерпретировать, как сложность адресной структуры базы. Изменение состава базы измеряется информацией

$$Inf(M,W,B)=(S(M),S(0),S(T\text{-}M),0,0,S(W),S(0),S(W),0,0). \qquad (23)$$

Заметим, что при $|T|>2$ вектор остается положительным, даже если число элементов базы данных не уменьшилось, в отличие от обычной меры комбинаторной информации, как негоэнтропии. В то же время, при $|T|=2$, $|M|=1$, $|W|=1$ информация обнуляется, кроме символических координат $S(0)$. Это тоже связано с гипотезой оптимального кодирования и не отражает семантики. Вероятно, случай текстового описания объекта не следует рассматривать, как математическую модель. Он требует другого определения векторов сложности и информации. Например, если за элемент текстовой модели принять один символ, связанный с предыдущим (первый — с собой), то можно потребовать $|T|>2$, и указанные семантические несоответствия практически исчезнут. При этом в качестве T фигурирует множество символов, а модель представляет собой их последовательность. Формулы (22)(23) сохранятся.

Литература.

1.(1968) Algebraic Theory of Machines, Languages and Semigroups (A. Arbib, ed), Academic Press, New York and London.

2. М. И. Мальцев. Алгоритмы и рекурсивные функции. “Наука”, М.,1986г.,367с.

3. А. В. Коганов. Индукторные пространства как средство моделирования. В сб. Вопросы кибернетики. Алгебра. Гипергеометрия. Вероятность. Моделирование; ред. В. Б. Бетелин. РАН, М., с 119-181

4. A. V. Koganov. Processes and Automorphisms on Inductor Spaces. Russian Jornal of Mathematic Phusyk, vol 4, nom 3, 1996, s 315-339

5. А. М. Хазен. Введение информации в аксиоматическую базу механики. М., 1998г. 168с.

6. А. М. Яглом, И. М. Яглом. Вероятность и информация. М., 1957, 160с.