Три подхода к определению понятия «Количество информации» *

А. Н. Колмогоров

1 Комбинаторный подход

Пусть переменное x способно принимать значения, принадлежащие конечному множеству X, которое состоит из N элементов. Говорят, что «энтропия» переменного равна

$$H(x) = \log_2 N.$$

Указывая определенное значение $x = a$ переменного x, мы «снимаем» эту энтропию, сообщая «инфомацию»

$$I = \log_2 N.$$

Если переменные $x_1, x_2, \ldots, x_k$ способны независимо пробегать множества, которые состоят соответственно из $N_1, N_2, \ldots, N_k$ элементов, то

$$H(x_1, x_2, \ldots, x_k) = H(x_1) + H(x_2) + \ldots + H(x_k). \quad (1)$$

Для передачи количества информации I приходится употреблять

$$I' = \begin{cases} I, & \text{при } I \text{ целом,} \\ [I] + 1, & \text{при } I \text{ дробном} \end{cases}$$

двоичных знаков. Например, число различных «слов», состоящих из k нулей и единиц и одной двойки, равно $2^k(k+1)$.
Поэтому количество информаци в такого рода сообщении равно

$$I = k + \log_2(k+1),$$

т.е. для «кодирования» такого рода слов в чистой двоичной системе требуется[1]

$$I' \approx k + \log_2 k$$

нулей и единиц.

При изложении теории информации обычно не задерживаются надолго на таком комбинаторном подходе к делу. Но мне кажется существенным подчеркнуть его логическую независимость от каких бы то ни было вероятностных допущений. Пусть, например, нас занимает задача кодирования сообщений, записанных в алфавите, состоящем из s букв, причем известно, что частоты

$$p_r = s_r/s \quad (2)$$

появления отдельных букв в сообщении длины n удовлетворяют неравенству

$$\chi = -\sum_{r=1}^{s} p_r \log_2 p_r \leq h. \quad (3)$$

Легко подсчитать, что при больших n двоичный логарифм числа сообщений, подчиненных требованию (3), имеет асимптотическую оценку:

$$H = log_2 N \sim nh.$$

Поэтому при передаче такого рода сообщений достаточно употребить примерно nh двоичных знаков.

Универсальный метод кодирования, который позволит передавать любое достаточно длинное сообщение в алфавите из s букв, употребляя не многим более чем nh двоичных знаков, не обязан быть чрезмерно сложным, в частности, не обязан начинаться с определения частот p_r для всего сообщения. Чтобы понять это, достаточно заметить: разбивая сообщение S на m отрезков $S_1, S_2, \ldots, S_m$, получим неравенство

$$\chi \geq n^{-1}[n_1\chi_1 + n_2\chi_2 + \ldots + n_m\chi_m]. \quad (4)$$

Впрочем, я не хочу входить здесь в детали этой специальной задачи. Мне важно лишь показать,

*Новое в жизни, науке, технике. Сер. «Математика, кибернетика». 1'91, стр. 24-29, ISBN 5-07-001613-X

[1] Всюду далее $f \approx g$ обозначает, что разность $f - g$ ограничена, а $f \sim g$, что отношение $f : g$ стремится к единице.

что математическая проблематика, возникающая на почве чисто комбинаторного подхода к измерению количества информации, не ограничивается тривиальностями.

Вполне естественным является чисто комбинаторный подход к понятию «энтропии речи», если иметь в виду оценку «гибкости» речи - показателя разветвленности возможностей продолжения речи при данном словаре и данных правилах построения фраз. Для двоичного логарифма числа N русских печатных текстов, составленных из слов, включенных в «Словарь русского языка» С. И. Ожегова и подчиненных лишь требованию «грамматической правильности» длины n, выраженной в «числе знаков» (включая «пробелы»), М. Ратнер и Н. Светлова получили оценку

$$h = (log_2 N)/n = 1,9 \pm 0,1.$$

Это значительно больше, чем оценки сверху для «энтропии литературных текстов», получаемые при помощи различных методов «угадывания продолжений». Такое расхождение вполне естественно, так как литературные тексты подчинены не только требованию «грамматической правильности».

Труднее оценить комбинаторную энтропию текстов, подчиненных определенным содержательным ограничениям. Представляло бы, например, интерес оценить энтропию русских текстов, могущих рассматриваться как достаточно точные по содержанию переводы заданного иноязычного текста. Только наличие такой «остаточной энтропии» делает возможным стихотворные переводы, где «затраты энтропии» на следование избранному метру и характеру рифмовки могут быть довольно точно подсчитаны. Можно показать, что классический четырехстопный рифмованный ямб с некоторыми естественными ограничениями на частоту «переносов» и т. п. требует допущения свободы обращения со словесным материалом, характеризуемой «остаточной энтропией» порядка 0,4 (при указанном выше условном способе измерения длины текста по «числу знаков, включая пробелы»). Если учесть, с другой стороны, что стилистические ограничения жанра, вероятно, снижают приведенную выше оценку «полной» энтропии с 1,9 до не более чем 1,1–1,2, то ситуация становится примечательной как в случае перевода, так и в случае оригинального поэтического творчества.

Да простят мне утилитарно настроенные читатели этот пример. В оправдание замечу, что более широкая проблема оценки количеств информации, с которыми имеет дело творческая человеческая деятельность, имеет очень большое значение.

Посмотрим теперь, в какой мере чисто комбинаторный подход позволяет оценить «количество информации», содержащееся в переменном x относительно связанного с ним переменного y. Связь между переменными x и y, пробегающими соответсвенно множества X и Y, заключается в том, что не все пары x, y, принадлежащие прямому произведению $X \times Y$, являются «возможными». По множеству возможных пар U определяются при любом $a \in X$ множества Y_a тех y, для которых $(a, y) \in U$.

x	y			
	1	2	3	4
1	+	+	+	+
2	+	−	+	−
3	−	+	−	−

Естественно определить условную энтропию равенством

$$H(y \mid a) = \log_2 N(Y_a) \tag{5}$$

(где $N(Y_x)$ – число элементов в множестве Y_x), а информацию в x относительно y – формулой

$$I(x : y) = H(y) - H(y \mid x). \tag{6}$$

Например, в случае, изображенном в таблице имеем

$I(x = 1 : y) = 0, \quad I(x = 2 : y) = 1,$
$I(x = 3 : y) = 2.$

Понятно, что $H(y \mid x)$ и $I(x : y)$ являются функциями от x (в то время как y входит в их обозначение в виде «связанного переменного»).

Без труда вводится в чисто комбинаторной концепции представление о «количестве информации, необходимом для указания объекта x при заданных требованиях к точности указания». (См. по этому поводу обширную литературу об «ε-энтропии» множеств в метрических пространствах.)

Очевидно,

$$H(x \mid x) = 0 \quad I(x : x) = H(x). \tag{7}$$

2 Вероятностный подход

Возможности дальнейшего развития теории информации на основе определений (5) и (6) остались в тени ввиду того, что придание переменым x и y характера «случайных переменных», обладающих определенным совместным распределением вероятностей, позволяет получить значительно более богатую систему понятий и соотношений. В параллель к введенным в §1 величинам имеем здесь

$$H_W(x) = -\sum_x p(x)\log_2 p(x), \quad (8)$$

$$H_W(y \mid x) = -\sum_y p(y \mid x)\log_2 p(y \mid x), \quad (9)$$

$$I_W(x : y) = H_W(y) - H_W(y \mid x). \quad (10)$$

По-прежнему $H_W(y \mid x)$ и $I_W(x : y)$ являются функциями от x. Имеют место неравенства

$$H_W(x) \le H(x), \quad H_W(y \mid x) \le H(y \mid x), \quad (11)$$

переходящие в равенства при равномерности соответсвующих распределений (на X и Y_x). Величины $I_W(x : y)$ и $I(x : y)$ не связаны неравенством определенного знака. Как и в §1,

$$H_W(x \mid x) = 0, \quad I_W(x : x) = H_W(x). \quad (12)$$

Но отличие заключается в том, что можно образовать математические ожидания $\mathbf{M}H_W(y \mid x)$, $\mathbf{M}I_W(x : y)$, а величина

$$I_W(x, y) = \mathbf{M}I_W(x : y) = \mathbf{M}I_W(y : x) \quad (13)$$

характеризует «тесноту связи» между x и y симметричным образом.

Стоит, однако, отметить и возникновение в вероятностной концепции одного парадокса: величина $I(x : y)$ при комбинаторном подходе всегда неотрицательна, как это и естественно при наивном представлении о «количестве информации», величина же $I_W(x : y)$ может быть и отрицательной. Подлинной мерой «количества информации» теперь становится лишь осредненная величина $I_W(x, y)$.

Вероятностный подход естествен в теории передачи по каналам связи «массовой» информации, состоящей из большого числа не связанных или слабо связанных между собой сообщений, подчиненных определенным вероятностным закономерностям. В такого рода вопросах практически безвредно и укоренившееся в прикладных работах смешение вероятностей и частот в пределах одного достаточно длинного временно́го ряда (получающее строгое оправдание при гипотезе достаточно быстрого «перемешивания»). Практически можно считать, например, вопрос об «энтропии» потока поздравительных телеграмм и «пропускной способности» канала связи, требующегося для своевременной и неискаженной передачи, корректно поставленным в его вероятностной трактовке и при обычной замене вероятностей эмпирическими частотами. Если здесь и остается некоторая неудовлетворенность, то она связана с известной расплывчатостью наших концепций, относящихся к связям между математической теорией вероятностей и реальными «случайными явлениями» вообще.

Но какой реальный смысл имеет, например, говорить о «количестве информации», содержащемся в тексте «Войны и мира»? Можно ли включить разумным образом этот роман в совокупность «возможных романов» да еще постулировать наличие в этой совокупности некоторого распределения вероятностей? Или следует считать отдельные сцены «Войны и мира» образующими случайную последовательность с достаточно быстро затухающими на расстоянии нескольких страниц «стохастическими связями»?

По существу, не менее темным является и модное выражение «количество наследственной информации», необходимой, скажем, для воспроизведения особи вида кукушка. Опять в пределах принятой вероятностной концепции возможны два варианта. В первом варианте рассматривается совокупность «возможных видов» с неизвестно откуда берущимся распределением вероятностей на этой совокупности[2]. Во втором варианте характеристические свойства вида считаются набором слабо связанных между собой случайных переменных. В пользу второго варианта можно привести соображения, основанные на реальном механизме мутационной изменчивости. Но соображения эти иллюзорны, если считать, что в результате естественного отбора возникает система согласованных между собой характеристических признаков вида.

[2] Обращение к множеству видов, существующих или существовавших на Земле, даже при чисто комбинаторном подсчете дало бы совершенно неприемлемо малые оценки сверху (что-либо вроде <100 бит!).

3 Алгоритмический подход

По существу, наиболее содержательным является представление о количестве информации «в чем-либо» (x) и «о чем-либо» (y). Не случайно именно оно в вероятностной концепции получило обобщение на случай непрерывных переменных, для которых энтропия бесконечна, но в широком круге случаев

$$I_W(x, y) = \int\int P_{xy}(dxdy) \log_2 \frac{P_{xy}(dxdy)}{P_x(dx)P_y(dy)}$$

конечно. Реальные объекты, подлежащие нашему изучению, очень (неограниченно?) сложны, но связи между двумя реально существующими объектами исчерпываются при более простом схематизированном их описании. Если географическая карта дает нам значительную информацию об участке земной поверхности, то все же микроструктура бумаги и краски, нанесенной на бумагу, никакого отношения не имеет к микроструктуре изображенного участка земной поверхности.

Практически нас интересует чаще всего количество информации в индивидуальном объекте x относительно индивидуального объекта y. Правда, уже заранее ясно, что такая индивидуальная оценка количества информации может иметь разумное содержание лишь в случаях достаточно больших количеств информации. Не имеет, например, смысла спрашивать о количестве информации в последовательности цифр 0 1 1 0 относительно последовательности 1 1 0 0. Но если мы возьмем вполне конкретную таблицу случайных чисел обычного в статистической практике объема и выпишем для каждой ее цифры цифру единиц ее квадрата по схеме

0	1	2	3	4	5	6	7	8	9
0	1	4	9	6	5	6	9	4	1,

то новая таблица будет содержать примерно

$$(\log_2 10 - \frac{8}{10})n$$

информации о первоначальной (n – число цифр в столбцах).

В соответсвии с только что сказанным предлагаемое далее определение величины $I_A(x:y)$ будет сохранять некоторую неопределенность. Разные равноценные варианты этого определения будут приводить к значениям, эквивалентным лишь в смысле $I_{A_1} \approx I_{A_2}$, т.е.

$$|I_{A_1} - I_{A_2}| \leq C_{A_1A_2},$$

где константа $C_{A_1A_2}$ зависит от положенных в основу двух вариантов определения универсальных методов программирования A_1 и A_2.

Будем рассматривать «нумерованную область объектов», т.е. счетное множество $X = \{x\}$, каждому элементу которого поставлена в соответствие в качестве «номера» $n(x)$ конечная последовательность нулей и единиц, начинающаяся с единицы. Обозначим через $l(x)$ длину последовательности $n(x)$. Будем предполагать, что

1) соответствие между X и множеством D двоичных последовательностей описанного вида взаимно однозначно;

2) $D \subset X$, функция $n(x)$ на D общерекурсивна [1], причем для $x \in D$

$$l(n(x)) \leq l(x) + C,$$

где C – некоторая константа;

3) вместе с x и y в X входит упорядоченная пара (x, y), номер этой пары есть общерекурсивная функция номеров x и y и

$$l(x, y) \leq C_x + l(y),$$

где C_x зависит только от x.

Не все эти требования существенны, но они облегчают изложение. Конечный результат построения инвариантен по отношению к переходу к новой нумерации $n'(x)$, обладающей теми же свойствами и выражающейся общерекурсивно через старую, и по отношению к включению системы X в более обширную систему X' (в предположении, что номера n' в расширенной системе для элементов первоначальной системы общерекурсивно выражаются через первоначальные номера n). При всех этих преобразованиях новые «сложности» и количества информации остаются эквивалентными первоначальным в смысле $\approx$.

«Относительной сложностью» объекта y при заданном x будем считать минимальную длину $l(p)$ программы p получения y из x. Сформулированное так определение зависит от «метода программирования». Метод программирования есть не что иное, как функция $\varphi(p, x) = y$, ставящая в соответсвие программе p и объекту x объект y.

В соответсвии с универсально признанными в современной математической логике взглядами следует считать функцию φ частично рекурсивной. Для любой такой функции полагаем

$$K_\varphi(y \mid x) = \begin{cases} \min\limits_{\varphi(p,x)=y} l(p), & \\ \infty, & \text{если нет такого } p, \\ & \text{что } \varphi(p,x) = y. \end{cases}$$

При этом функция $v = \varphi(u)$ от $u \in X$ со значениями $v \in X$ называется частично рекурсивной, если она порождается частично рекурсивной функцией преобразования номеров

$$n(v) = \Psi[n(u)].$$

Для понимания определения важно заметить, что частично рекурсивные функции, вообще говоря, не являются всюду определенными. Не существует регулярного процесса для выяснения того, приведет применение программы p к объекту x к какому-либо результату или нет. Поэтому функция $K_\varphi(y \mid x)$ не обязана быть эффективно вычислимой (общерекурсивной) даже в случае, когда она заведомо конечна при любых x и y.

Основная теорема. *Существует такая частично рекурсивная функция $A(p, x)$, что для любой другой частично рекурсивной функции $\varphi(p, x)$ выполнено неравенство*

$$K_A(y \mid x) \leq K_\varphi(y \mid x) + C_\varphi,$$

где константа C_φ не зависит от x и y.

Доказательство опирается на существование универсальной частично рекурсивной функции $\Phi(n, u)$, обладающей тем свойством, что, фиксируя надлежащий номер n, можно получить по формуле $\varphi(u) = \Phi(n, u)$ любую другую частично рекурсивную функцию. Нужная нам функция $A(p, x)$ определяется формулой[3]

$$A((n, q), x) = \Phi(n, (q, x)).$$

В самом деле, если

$$y = \varphi(p, x) = \Phi(n, (p, x)),$$

то

$$A((n, p), x) = y,$$
$$l(n, p) \leq l(p) + C_n.$$

Функции $A(p, x)$, удовлетворяющие требованиям основной теоремы, назовем (как и определяемые ими методы программирования) *асимптотически оптимальными.* Очевидно, что для них при любых x и y «сложность» $K_A(y \mid x)$ конечна. Для двух таких функций A_1 и A_2

$$|K_{A_1}(y \mid x) - K_{A_2}(y \mid x)| \leq C_{A_1 A_2},$$

где $C_{A_1 A_2}$ не зависит от x и y, т. е. $K_{A_1}(y \mid x) \approx K_{A_2}(y \mid x)$.

Наконец, $K_A(y) = K_A(y \mid 1)$ можно считать просто «сложностью объекта y» и определить «количество информации в x относительно y» формулой

$$I_A(x : y) = K_A(y) - K_A(y \mid x).$$

Легко доказать[4], что величина эта всегда в существенном положительна:

$$I_A(x : y) \gtrsim 0,$$

что понимается в том смысле, что $I_A(x : y)$ не меньше некоторой отрицательной константы C, зависящей лишь от условностей избранного метода программирования. Как уже говорилось, вся теория рассчитана на применение к большим количествам информации, по сравнению с которым $|C|$ будет пренебрежимо мал.

Наконец, $K_A(x \mid x) \approx 0$, $I_A(x : x) \approx K_A(x)$.

Конечно, можно избегнуть неопределенностей, связанных с константами C_φ и т. д., остановившись на определенных областях объектов X, их нумерации и функции A, но сомнительно, чтобы это можно было сделать без явного произвола. Следует, однако, думать, что различные представляющиеся здесь «разумные» варианты будут приводить к оценкам «сложностей», расходящимся на сотни, а не на десятки тысяч бит. Поэтому такие величины, как «сложность» текста романа «Война и мир», можно считать определенными с практической однозначностью. Эксперименты по угадыванию продолжений литературных текстов позволяют оценить сверху условную сложность при

[3] $\Phi(n, u)$ определена только в случае $n \in D$, $A(p, x)$ только в случае, когда p имеет вид (n, q), $n \in D$.

[4] Выбирая в виде функции сравнения $\varphi(p, x) = A(p, 1)$, получим $K_A(y \mid x) \leq K_\varphi(y \mid x) + C_\varphi = K_A(y) + C_\varphi$.

заданном запасе «априорной информации» (о языке, стиле, содержании текста), которой располагает угадывающий. В опытах, проводившихся на кафедре теории вероятностей Московского государственного университета, такие оценки сверху колебались между 0,9 и 1,4. Оценки порядка 0,9–1,1, получившиеся у Н. Г. Рычковой, вызвали у менее удачливых угадчиков разговоры о ее телепатической связи с авторами текстов.

Я думаю, что для «количества наследственной информации» предполагаемый подход дает в принципе правильное определение самого понятия, как бы ни была трудна фактическая оценка этого количества.

4 Заключительные замечания

Изложенная в §3 концепция обладает одним существенным недостатком: она не учитывает «трудности» переработки программы p и объекта x в объект y. Введя надлежащие определения, можно доказать точно формулируемые математические предложения, которые законно интерпретировать как указание на существование таких случаев, когда объект, допускающий очень простую программу, т. е. обладающий очень малой сложностью $K(x)$, может быть восстановлен по коротким программам лишь в результате вычислений совершенно нереальной длительности. В другом месте я предполагаю изучить зависимость необходимой сложности программы $K^t(x)$ от допустимой трудности t ее переработки в объект x. Сложность $K(x)$, которая была определена в §3, появится при этом в качестве минимума $K^t(x)$ при снятии ограничений на величину t.

За пределами этой заметки остается и применение построений §3 к новому обоснованию теории вероятностей. Грубо говоря, здесь дело идет о следующем. Если конечное множество M из очень большого числа элементов N допускает определение при помощи программы длины, пренебрежимо малой по сравнению с $\log_2 N$, то почти все элементы M имеют сложность $K(x)$, близкую к $\log_2 N$. Элементы $x \in M$ этой сложности и рассматриваются как «случайные» элементы множества M. Не вполне завершенное изложение этой идеи можно найти в статье [2].

Список литературы

[1] *Успенский В. А.* Лекции о вычислимых функциях. - М.: Физматгиз, 1960.

[2] *Kolmogorov A. N.* On tables of random numbers // Sanknya. A, 1963. - Vol.25. - 4. - P. 369-376.